**Reference**

Lefer, M.-A. (2020). Parallel Corpora. In M. Paquot & S. Th. Gries (eds), *Practical Handbook of Corpus Linguistics*. Springer, 257-282.

# 12 Parallel Corpora

**Marie-Aude Lefer**[1]

**Abstract** This chapter gives an overview of parallel corpora, i.e. corpora containing source texts in a given language, aligned with their translations in another language. More specifically, it focuses on directional corpora, i.e. parallel corpora where the source and target languages are clearly identified. These types of corpora are widely used in contrastive linguistics and translation studies. The chapter first outlines the key features of parallel corpora (they typically contain written texts translated by expert translators working into their native language) and describes the main methods of parallel corpus analysis, including the combined use of parallel and comparable corpora. It then examines the major challenges that are linked with the design and analysis of parallel corpora, such as text availability, metadata collection, bitext alignment, and multilingual linguistic annotation, on the one hand, and data scarcity, interpretation of the results and infelicitous translations, on the other. Finally, the chapter shows how these challenges can be overcome, most notably by compiling balanced, richly-documented parallel corpora and by cross-fertilizing insights from cross-linguistic research and natural language processing.

## 12.1 Introduction

This chapter gives an overview of parallel corpora, which are widely used in corpus-based cross-linguistic research (here understood as an umbrella term for contrastive linguistics and translation studies) and natural language processing. Parallel corpora (also called *translation corpora*) contain source texts in a given language (the source language, henceforth SL), aligned with their translations in another language (the target language, henceforth TL). It is important to point out from the outset that the term *parallel corpus* is to some extent ambiguous, because it is sometimes used to refer to comparable original texts in two or more languages, especially texts that belong to comparable genres or text types and deal with similar topics (e.g. Italian and German newspaper articles about migration or English and Portuguese medical research articles). Here, the term will only be used to refer to collections of source texts and their translations.

The compilation of parallel corpora started in the 1990s. Progress has been rather slow, compared with monolingual corpus collection initiatives, but in recent years we have witnessed a boom in the collection of parallel corpora, which are increasingly larger and multilingual. Parallel corpora are highly valuable resources to investigate cross-linguistic contrasts (differences between linguistic systems) and translation-related phenomena, such as translation properties (features of translated language). They can

[1] Marie-Aude Lefer
Université catholique de Louvain
Louvain-la-Neuve, Belgium
e-mail: marie-aude.lefer@uclouvain.be

also be used for a wide range of applications, such as bilingual lexicography, foreign language teaching, translator training, terminology extraction, computer-aided translation, machine translation and other natural language processing tasks (e.g. word sense disambiguation and cross-lingual information retrieval).

This chapter is mainly concerned with the design and analysis of parallel corpora in the two fields of corpus-based contrastive linguistics and corpus-based translation studies. Contrastive linguistics (or *contrastive analysis*) is a linguistic discipline that is concerned with the systematic comparison of two or more languages, so as to describe their similarities and differences. Corpus-based contrastive linguistics was first pioneered by Stig Johansson in the 1990s and has been thriving ever since. Corpus-based translation studies is one of the leading paradigms in Descriptive Translation Studies (Toury, 2012). This field also emerged in the 1990s, under the impetus of Mona Baker, and relies on corpus linguistic tools and methods to elucidate translated text (in particular, the linguistic features that set translated language apart from other forms of language production) (cf. Kruger et al., 2011; De Sutter et al., 2017). Contrastive linguistics and translation studies, which both make intensive use of parallel corpora, are quite close, as demonstrated by edited volumes such as Granger et al. (2003) and the biennial *Using Corpora in Contrastive and Translation Studies* conference series (e.g. Xiao, 2010).

## 12.2 Fundamentals

### *12.2.1 Types of Parallel Corpora*

Parallel corpora can be of many different types. They can be bilingual (one SL and one TL), such as the *English-Norwegian Parallel Corpus* (ENPC; Johansson, 2007), or multilingual (more than one SL and/or TL), such as the *Oslo Multilingual Corpus*, which is fully trilingual (English, Norwegian and German), with some texts available in Dutch, French and Portuguese as well (ibid., 18-19). Other multilingual parallel corpora include the Slavic parallel corpus ParaSol (Wadenfelts, 2011) and InterCorp (Čermák & Rosen, 2012). A further distinction is made between monodirectional corpora, when only one translation direction is represented ($SL_x > TL_Y$, e.g. English > Chinese), and bidirectional (or *reciprocal*) corpora, when both translation directions are included ($SL_x > TL_Y$ and $SL_Y > TL_X$, e.g. English > Chinese and Chinese > English). The ENPC, for example, is bidirectional (from English to Norwegian, and vice versa). Most parallel corpora contain published translations (with some exceptions, e.g. when translations are specifically commissioned for a particular corpus compilation project). In most cases, only one translation of each source text is included. However, there are also *multiple translation corpora*, which include several translations of the same source text in a given TL. Such corpora make it possible to compare the translation solutions used by various translators rendering the same source text.

### *12.2.2 Main Characteristics of Parallel Corpora*

The majority of parallel corpora used in contrastive linguistics and translation studies are characterized by two key features. First, the source and target languages are clearly identified. In other words, the translation direction is known (from $Language_X$ to $Language_Y$ or from $Language_Y$ to $Language_X$). In cross-linguistic research, it is of paramount importance to know, for instance, whether a given text was translated from Spanish into German or vice versa, because corpus studies have shown that translation direction influences translation choices, and hence the linguistic make-up of translated text (e.g. Dupont & Zufferey, 2017). Second, only *direct* translation is included, i.e. no pivot (intermediary, mediating) language is used between the source and target languages. In texts produced by the European Union (EU), for example, English has been systematically used as a pivot language since the early 2000s. In practical terms, this means that a text originally written in, say, Slovenian or Dutch is first translated

into English. The English version is then translated into the other official languages of the EU. In other words, English acts as a pivot language and most target texts originating from EU institutions are in fact translations of translations (see Assis Rosa et al., 2017 on the issue of indirect translation). Parallel corpora that display these two features (known as *translation direction* and *translation directness*) will be referred to as *directional parallel corpora* in this chapter (a term borrowed from Cartoni & Meyer, 2012). Parallel corpora whose translation direction is unknown and/or where a pivot language has been used will be called *non-directional*. Examples of the latter type include the Europarl corpus (Koehn, 2005), the Eur-Lex corpus (Baisa et al., 2016), and the *United Nations Parallel Corpus* (Ziemsky et al., 2016). It is important to bear in mind, however, that the distinction between directional and non-directional parallel corpora is not always clear-cut. In some parallel corpora, both types of parallel texts are included. For example, the *Dutch Parallel Corpus* (DPC; Macken et al., 2011), which is largely directional, contains some indirect, EU translations.

Directional parallel corpora typically (i) contain written texts (ii) translated by expert translators (iii) working into their native language (L1), and (iv) cover a rather limited number of text types or genres. Each of these typical features will be discussed in turn:

(i) Directional parallel corpora mainly cover written translation (e.g. the ENPC), to the detriment of other translation modalities, such as interpreting and audiovisual translation. In recent years, however, efforts have been made to include other forms of translation. A case in point is the compilation of several parallel corpora of simultaneous interpreting (see Russo et al., 2018 for an overview of corpus-based interpreting studies). In these corpora, the main source of data (speeches and their interpreted versions) is the European Parliament (Bernardini et al., 2018). An example of one such European Parliament interpreting corpus is the fully trilingual English-Italian-Spanish *European Parliament Interpreting Corpus* (Russo et al., 2006). Recent developments also include the compilation of *intermodal* parallel corpora, i.e. corpora representing several translation modalities (e.g. written translation and simultaneous interpreting), such as the *European Parliament Translation and Interpreting Corpus* (EPTIC; Ferraresi & Bernardini, forthcoming). EPTIC features two main components: (i) simultaneous interpreting: transcripts of speeches delivered at the European Parliament plenary sittings and transcripts of the simultaneous interpretations of these speeches, and (ii) written translation: the verbatim reports of the plenary sittings, as officially published on the European Parliament website, alongside the official translations of these verbatim reports (the Europarl corpus is also based on this written material, see *Representative corpora* below). Parallel corpora of sign interpreting (e.g. Meurant et al., 2016) and audiovisual translation modalities (such as subtitling, dubbing, and film audio description; cf. Baños et al., 2013) have also been collected recently. Some of these parallel corpora are multimodal, in the sense that they contain different modes, such as language, image, sound and music (e.g. Jimenez Hurtado & Soler Gallego, 2013; Chap. 16).

(ii) In general, parallel corpora include target text translated (or assumed to have been translated) by professional and/or expert translators (it must be stressed, however, that limited metadata on translators' status have been collected to date; see Sect. 12.2.4.2). In some cases, the translators' status is rather unclear (e.g. in translated news items, found in several parallel corpora, from *Le Monde Diplomatique*, a French monthly newspaper with more than thirty international editions, in 20+ languages[2]). Other translators' profiles are also represented, albeit less frequently, such as non-professional, volunteer translators, as in the TED Talks WIT³ corpus (*Web Inventory of*

[2] https://www.monde-diplomatique.fr/diplo/int/

*Transcribed and Translated Talks*; Cettolo et al., 2012). Aside from professional and volunteer translators, some parallel corpora, called *learner translation corpora* (LTC), contain translations produced by foreign language learners or trainee translators, i.e. novices (see also Chap. 13). The first LTC emerged in the early 2000s (Uzar, 2002; Bowker & Bennison, 2003) and have been followed by several similar initiatives, such as the MeLLANGE corpus (Castagnoli et al., 2011), the English-Catalan UPF LTC (Espunya, 2014), the *Russian Learner Translator Corpus* (Kutuzov & Kunilovskaya, 2014), and the *Multilingual Student Translation* corpus[3]. The vast majority of directional parallel corpora contain translations produced by human translators (in some cases, with the help of computer-aided translation tools). Recently, however, translation scholars have started to include machine-translated texts alongside human-translated texts, with a view to uncovering the linguistic traits that differentiate machine translation from human translation (computer-aided or otherwise) (e.g. Lapshinova-Koltunski, 2017).

(iii) Directional parallel corpora tend to be restricted to L1 translation (i.e. when the translation is carried out into the translator's native language), except in the case of some LTC, which contain L2 (inverse, reverse) translation as well, or corpora representing language pairs for which L2 translation is common practice (e.g. Finnish to English) (see Beeby Lonsdale, 2009 on directionality practices).

(iv) Most directional, balanced parallel corpora used in contrastive linguistics and translation studies are restricted to a couple of genres or text types, mainly fictional prose (e.g. the English-Portuguese COMPARA; Frankenberg-Garcia & Santos, 2003; the "core" part of InterCorp), news (news items and opinion articles published in newspapers and magazines) and/or non-fiction, such as popular science texts (e.g. the ENPC; the English-French *Poitiers-Louvain Échange de Corpus Informatisés* PLECI[4]; the French-Slovenian FraSloK parallel corpus, Mezeg, 2010; and the English-Spanish ACTRES parallel corpus, Izquierdo et al., 2008). A handful of directional parallel corpora cover a wider range of text types. Examples include the DPC for the language pairs Dutch-English and Dutch-French (Macken et al., 2011) and the *CroCo* corpus for German-English (Hansen-Schirra et al., 2012), with five and ten text types represented, respectively.

The directional parallel corpora featuring the four characteristics outlined above are relatively modest in size compared with monolingual reference corpora commonly used in corpus linguistics (they usually contain a few million words). This is even more striking for parallel corpora of interpreted language, in view of the many hurdles inherent in transcribing spoken data (Bernardini et al., 2018; Chap. 11). If more parallel data are needed, and provided translation direction and directness are not considered to be of particular relevance, researchers can turn to several non-directional parallel corpora (mainly of legislative and administrative texts) that are much larger than the parallel corpora discussed so far. These mega corpora are used widely in natural language processing, for example for data-driven machine translation. However, it is important to bear in mind that (i) in these corpora, translation direction is often unknown (i.e. the source and target languages are not clearly identified), and (ii) in many instances, the translation relationship between the parallel texts for a given language pair is indirect (either the translation is done through an intermediary, pivot language, or the parallel texts in a given pair are both translations from another, third language). Generally speaking, non-directional parallel corpus data

[3] https://uclouvain.be/en/research-institutes/ilc/cecl/must.html
[4] https://uclouvain.be/en/research-institutes/ilc/cecl/pleci.html

should be treated with caution. While their use makes sense in natural language processing research, it remains to be seen whether they can yield reliable insights into cross-linguistic differences.

### *12.2.3 Methods of Analysis in Cross-linguistic Research*

Parallel corpora are widely used in corpus-based contrastive linguistics and translation studies and they are starting to emerge as a useful source of data in typology as well (Levshina, 2016). As pointed out by Johansson (2007: 3), most contrastive scholars "have either explicitly or implicitly made use of translation as a means of establishing cross-linguistic relationships. […] As translation shows what elements may be associated across languages, it is fruitful to base a contrastive study on a comparison of original texts and their translations". In other words, parallel corpora can be used to study cross-linguistic correspondences (e.g. between lexical items, lexico-syntactic patterns or grammatical structures). The corpus methods used to achieve that goal are similar to those applied in monolingual corpus linguistics, such as concordances (Chap. 8) and co-occurrence data (Chap. 7).

Figure 12.1 provides a sample of bilingual concordances for the English phrase *no kidding* and its Italian equivalents in a corpus of subtitled films and series (OpenSubtitles2011, available in Sketch Engine; see Sect. 12.4). A cursory glance at Fig. 12.1 shows that Italian equivalents include *non scherzo* ('I am not kidding'), *sul serio* ('seriously') and *davvero* ('really'). The detailed analysis of English-Italian equivalences found in the corpus can act as a springboard for an in-depth contrastive analysis (e.g. what are the discursive and pragmatic functions of *no kidding* in scripted spoken English and which equivalent expressions are used in Italian to fulfill these functions?). Bilingual concordances are also widely used in translation studies to investigate the translation procedures used to render specific items (e.g. lexical innovations, proper names, culture-specific elements). For instance, on the basis of an Italian-to-German parallel corpus of tourist brochures, it is possible to determine whether translators adapt SL culture-bound items (e.g. *macchiato*, *caffè latte*) or whether they keep them in their translation (perhaps with an explanatory note), which reflects more general translation strategies towards domestication and foreignization.

Figure 12.2 shows a sample of a bilingual Word Sketch, i.e. a summary of the grammatical and collocational behaviors of equivalent words, for English *sustainability* and its French equivalent *durabilité* in parliamentary proceedings (Europarl). The bilingual Word Sketch makes it possible, among other things, to detect equivalent verbal collocates of the English and French nouns under scrutiny, such as *jeopardize/menacer* and *ensure/assurer*. This kind of co-occurrence analysis is particularly helpful for contrastive phraseology, applied translation studies (e.g. to raise trainee translators' awareness of phraseological equivalence) and bilingual lexicography.

| | OPUS2 English | | | OPUS2 Italian |
|---|---|---|---|---|
| OpenSubtit... | months! What' s up? Shut the fuck up! | No kidding | , I' m sorry. I' m sorry, man. She | OpenSubtit... Non scherzo, mi spiace. |
| OpenSubtit... | Disregard the guy' s insinuations. | No kidding | , lot of stuff on the ball. Sounds | OpenSubtit... Non scherzo, ha moltissimo talento. |
| OpenSubtit... | , my two little stars, Don and Lina. | No kidding | , folks, aren' t they great? All | OpenSubtit... Non sono straordinari, signori? |
| OpenSubtit... | ... Boss, we got one for you ... Yeah, | no kidding | ... a sedan! - I was afraid of this ... | OpenSubtit... Sì, non scherzo ... una berlina! |
| OpenSubtit... | could someone just -- You' re upset. | No kidding | I' m upset. My life is freaking over! | OpenSubtit... Sono turbata mica per scherzo. |
| OpenSubtit... | would be a good one for you. ... Yeah, | no kidding | . Does it also say, if you got no age, | OpenSubtit... Quella sarebbe un' ottima opportunita ' per te. ... si ', stai scherzando. |
| OpenSubtit... | ... No, but this time ... I mean it ... | No kidding | . Want some? We wouldn' t want to be | OpenSubtit... Vi assicuro ... Sul serio! |
| OpenSubtit... | when they were planting that bomb. | No kidding | . pardon! Pardon. S' il vous plaÃ®t. | OpenSubtit... Ma davvero? Pardon, pardon! |
| OpenSubtit... | to be your research assistant. Oh, | no kidding | . We are not going to have time for | OpenSubtit... Oh, non scherzare. |
| OpenSubtit... | ! Hey! We got to get out of here. Yeah, | no kidding | . Come on. Come on! Come on, hurry up! | OpenSubtit... Contaci. |
| OpenSubtit... | least you' re taking it lying down. | No kidding | , Cosmo. Did you ever see anything as | OpenSubtit... Non scherzare, Cosmo. |
| OpenSubtit... | can guess the rest. Wait a minute. | No kidding | . What' s your idea? I got a look at | OpenSubtit... Senza scherzi. |
| OpenSubtit... | . Shall I massage you again? Again? | No kidding | ! We ' d have to stay here forever. | OpenSubtit... Starai scherzando !! |
| OpenSubtit... | you his number, And he delivers. | No kidding | ? Listen, I wanted to say thank you. | OpenSubtit... Ma davvero? |
| OpenSubtit... | Stay Here. Oh, What Now? Broke Axle. | No Kidding | . Come On, Let' s Go. We' re Walking. | OpenSubtit... Non scherziamo. |
| OpenSubtit... | is Young Long talking to? Cancer: | No Kidding | ! I congratulate you on the success | OpenSubtit... Non scherzare! |
| OpenSubtit... | must have been a leak. - Really, | no kidding | ? - Unbelievable! - Is this the | OpenSubtit... - Sul serio, non scherzi? |
| OpenSubtit... | What " no kidding "? Nothing. I said " | No kidding | . " Not right now. Because. Not right | OpenSubtit... - Nulla, ho solo detto " non scherzare ". |
| OpenSubtit... | ) Get a little air. - (Andrew) | No kidding | . A little. It' s such a pretty night. | OpenSubtit... - Ma dai? |
| OpenSubtit... | . MOUSE: He is so weird. (LAUGHS) | No kidding | . That is so beautiful. I wonder what | OpenSubtit... Sul serio. |
| OpenSubtit... | . I was Treasurer! No, I know. | No kidding | , what about -- obviously, I don' t | OpenSubtit... - Davvero ... - Ovviamente, non ho l' esperienza di tua madre, in campagna elettorale, ma . |
| OpenSubtit... | business opportunity. Oh, | no kidding | . Yeah, it' s, uh ... Well, actually, | OpenSubtit... - Oh, ma non mi dire. |
| OpenSubtit... | Jesus. Hello. I' m Eva Molnar. Yeah, | no kidding | . Are you Bela Molnar? No. I used to be | OpenSubtit... Davvero? |
| OpenSubtit... | , either? I don' t want to lose her. | No kidding | ! Stop acting. What did you ask Osumi | OpenSubtit... Basta con le finzioni! |
| OpenSubtit... | , the girl was sexually assaulted. | No kidding | . And the weapon used was a dagger - A | OpenSubtit... E questo mi pare evidente. |
| OpenSubtit... | . Always getting things wrong. | No kidding | . Why' re you looking at me that way? | OpenSubtit... - Sempre a sbagliare tutto. |

**Fig. 12.1** English *no kidding* and its Italian translation equivalents in the OpenSubtitles2011 corpus (OPUS2, Sketch Engine, Lexical Computing Ltd)

**sustainability** *(noun)* EUROPARL7, English freq = 2,490 (40.99 per million) **durabilité** EUROPARL7, French freq = 1,883 (28.24 per million)

Use another candidate translation: viabilité finance durable environnemental biocarburants pêche pérennité halieutique compétitivité
Click on collocates to access reciprocal bilingual search or find translated collocations

| object_of | | | verbs with "durabilité" as object | | |
|---|---|---|---|---|---|
| | | 25.14 | | | 19.33 |
| jeopardise | 13 | 7.37 | assurer | 71 | 7.61 |
| ensure | 135 | 7.19 | assurer la durabilité | | |
| ensure the sustainability of | | | garantir | 94 | 7.44 |
| measure | 10 | 6.58 | de garantir la durabilité | | |
| does not measure environmental sustainability | | | menacer | 10 | 7.03 |
| endanger | 5 | 6.43 | mesurer | 7 | 6.60 |
| guarantee | 43 | 6.41 | placer | 5 | 6.03 |
| guarantee the sustainability of | | | refléter | 4 | 5.34 |
| secure | 15 | 6.39 | accroître | 6 | 5.25 |
| enhance | 11 | 6.08 | promouvoir | 10 | 5.23 |
| threaten | 12 | 5.97 | promouvoir la durabilité | | |
| safeguard | 11 | 5.86 | favoriser | 7 | 5.20 |
| promote | 25 | 5.00 | évaluer | 4 | 5.10 |
| promote sustainability | | | maintenir | 5 | 4.77 |
| undermine | 6 | 4.98 | encourager | 6 | 4.42 |
| preserve | 4 | 4.93 | améliorer | 10 | 4.24 |
| assess | 6 | 4.91 | améliorer la durabilité | | |
| achieve | 33 | 4.79 | savoir | 5 | 4.05 |
| to achieve sustainability | | | atteindre | 5 | 4.03 |
| restore | 4 | 4.77 | soutenir | 7 | 3.16 |
| incorporate | 5 | 4.74 | concerner | 15 | 2.66 |
| maintain | 10 | 4.52 | ce qui concerne la durabilité | | |
| improve | 13 | 3.75 | | | |

| subject_of | | | adj_subject_of | | | verbs with "durabilité" as subject | | |
|---|---|---|---|---|---|---|---|---|
| | | 5.78 | | | 0.88 | | | 1.81 |
| mean | 6 | 3.26 | possible | 4 | 1.30 | devoir | 14 | 2.90 |
| require | 5 | 2.53 | important | 5 | 0.34 | la durabilité doit | | |
| be | 92 | 0.44 | | | | | | |
| sustainability is | | | | | | | | |

**Fig. 12.2** Sample of a bilingual Word Sketch for English *sustainability* and its French equivalent *durabilité* (Europarl7, Sketch Engine, Lexical Computing Ltd)

In the two examples mentioned above, we started with a given SL item (*no kidding*, *sustainability*) and examined its translation equivalents in the TL (Italian and French, respectively), i.e. going from source to target. Interestingly, this source-to-target approach is also used in monolingual corpus linguistics to examine the semantic, discursive and pragmatic features of source-language items (Noël, 2003). For example, Aijmer & Simon-Vandenbergen (2003) examine the meanings and functions of the English discourse particle *well* on the basis of its Swedish and Dutch translation equivalents in a parallel corpus of fictional texts.

An alternative method is to start off from a given item or structure in translated texts and examine its corresponding source-text items or structures, i.e. from target to source. Taking the same example as above, this would entail analyzing all occurrences of *sul serio* in Italian subtitles and identifying the English source items that have triggered their use. This target-to-source approach is quite common in translation studies. Delaere & De Sutter (2017), for example, rely on an English-to-Dutch parallel corpus to find out whether the English loanwords found in translated Dutch stem from their corresponding trigger words in the English source texts.

Naturally, these two approaches (source to target and target to source) can be combined if a more comprehensive picture of cross-linguistic correspondences is required. Indeed, many new insights can be gained by investigating a given item or structure in both source and target texts, so as to find out how it is commonly translated and which items in the other language have triggered its use in translation (e.g. Zufferey & Cartoni, 2012).

It is also possible, on the basis of parallel corpora, to work out what Altenberg has termed *mutual correspondence* (or *mutual translatability*), i.e. "the frequency with which different (grammatical, semantic and lexical) expressions are translated into each other" (Altenberg, 1999: 254). Mutual correspondence is calculated as follows, with *At* and *Bt* corresponding to the frequencies of the compared items *A* and *B* in the target texts (*t*), and *As* and *Bs* to their frequencies in the source texts (*s*):

$$mutual\ correspondence = \frac{(\mathrm{At} + \mathrm{Bt})\ x\ 100}{\mathrm{As} + \mathrm{Bs}}$$

If, say, a lexical item *A* is always translated with an item *B*, and vice versa, then items *A* and *B* have a mutual correspondence of 100%. If, on the contrary, *A* and *B* are never translated with each other, they display a mutual correspondence of 0%. In other words, this index makes it possible to assess the extent to which items are equivalent across languages: "the higher the mutual correspondence value is, the greater the equivalence between the compared items is likely to be" (Altenberg & Granger, 2002: 18). For example, Dupont & Zufferey (2017) find that in samples of 200 occurrences extracted from Europarl, the adverb pair *however/cependant* displays a mutual correspondence of 57% (*however* > *cependant*: 87/200, *cependant* > *however*: 140/200), while the *however/toutefois* pair has a lower correspondence score of 49% (*however* > *toutefois*: 80/200, *toutefois* > *however*: 114/200):

$$however/cependant = \frac{(87 + 140)\ x\ 100}{200 + 200}$$

$$however/toutefois = \frac{(80 + 114)\ x\ 100}{200 + 200}$$

The scores tend to indicate that in parliamentary proceedings, the *however/cependant* cross-linguistic equivalence is somewhat stronger than for *however* and *toutefois*.

So far, we have outlined different methods of parallel corpus analysis (from source to target, from target to source, mutual correspondence). However, it should be stressed that several types of corpora can be combined to reveal and disentangle cross-linguistic contrasts and translation-related phenomena (Bernardini, 2011; Johansson, 2007; Halverson, 2015). Two types of corpora are commonly used in cross-linguistic research in combination with parallel corpora: (i) bilingual/multilingual comparable corpora and (ii) monolingual comparable corpora. Their combined use with parallel corpora will be discussed in turn.

Bilingual (or multilingual) comparable corpora are "collections of original [i.e. non-translated] texts in the languages compared" (Johansson, 2007: 5). The texts are strictly matched by criteria such as register, genre, text type, domain, subject matter, intended audience, time of publication, and size. Examples include KIAP, a comparable corpus of research articles in Norwegian, English, and French (Fløttum et al., 2013) and the *Multilingual Editorial Corpus*, a comparable corpus of newspaper editorials in English, Dutch, French, and Swedish.[5] Bilingual and multilingual comparable corpora usefully complement parallel corpora in that a given phenomenon can be studied cross-linguistically on the basis of comparable *original* texts, i.e. texts displaying no trace of source-language or source-text influence, unlike translations in parallel corpora. Corpus studies combining both types of corpora can start either with the bilingual/multilingual comparable analysis, before turning to the parallel corpus analysis, or the other way around, depending on the research questions to be tackled (see Johansson, 2007 for more details). Interestingly, bilingual comparable and parallel corpora can be combined in the same corpus framework, namely bidirectional parallel corpora whose two translation directions are truly comparable in terms of size, text types, etc. As shown in Fig. 12.3, for example, the ENPC can function both as a bidirectional parallel corpus (English originals > Norwegian translations and Norwegian originals > English translations; see black arrows) and as a bilingual comparable corpus (English originals and Norwegian originals; see white double arrow). Numerous parallel corpora are based on the ENPC model, such as the *English-Swedish Parallel Corpus* (ESPC), COMPARA and PLECI.

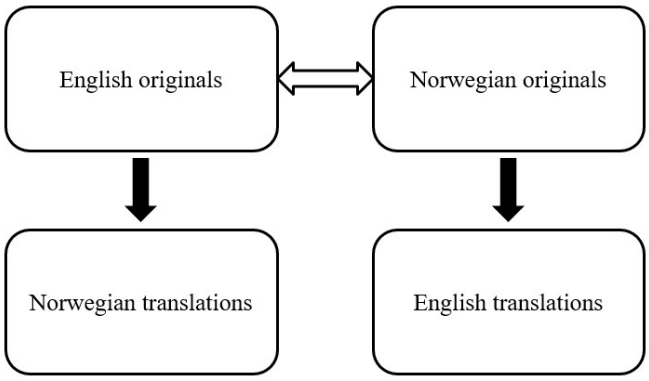


**Fig. 12.3** The model for the ENPC (based on Johansson, 2007: 11)

However, the main problem of the bidirectional ENPC model is that the selection of texts to be included in the corpus is limited to genres that are commonly translated in *both* directions (see Johansson, 2017: 12 on this issue). In other words, the number of genres and texts that can be included in the corpus is often limited (e.g. only fiction and non-fiction texts in the ENPC). As a result, to improve representativeness, the comparable, original components of bidirectional parallel corpora need to be supplemented with larger, multi-genre (reference) monolingual corpora of the languages investigated (see Fig. 12.4).

[5] https://uclouvain.be/en/research-institutes/ilc/cecl/mult-ed.html

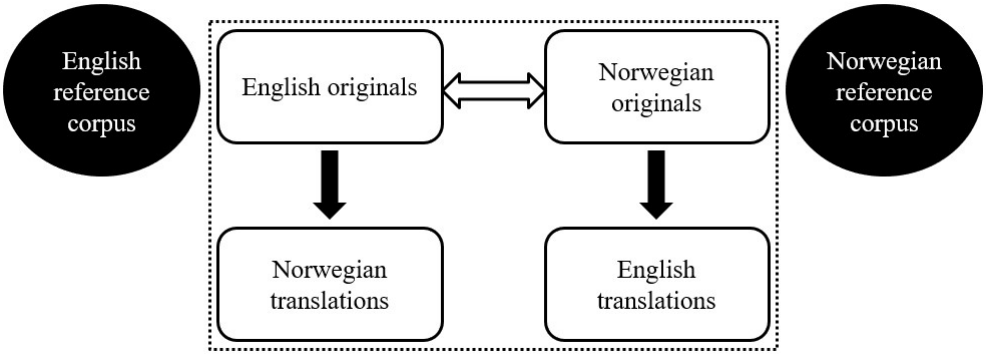


**Fig. 12.4** The model for the ENPC, with additional reference monolingual corpora

Parallel corpora can also be combined with monolingual comparable corpora, which include comparable translated and non-translated (i.e. original) texts in a given language (e.g. novels originally written in English alongside novels translated into English from a variety of source languages; see, for example, the 10-million-word *Translational English Corpus*[6]). Monolingual comparable corpora of translated and original texts are widely used in translation studies, with a view to identifying the major distinguishing features of translated language, when compared with original language production (the so-called *translation universals*, or *translation features/properties*, such as simplification, normalization and increased explicitness; cf. Baker, 1993, 1995). Parallel corpora, when combined with monolingual comparable corpora, are used to check for source-text and/or source-language influence. Cappelle & Loock (2013), for example, use parallel corpus data to find out whether the under-representation of existential *there* in English translated from French (as compared with non-translated English) stems from SL (French) interference. Parallel and monolingual comparable corpora can be integrated within the same overall corpus framework, as shown in Fig. 12.5.

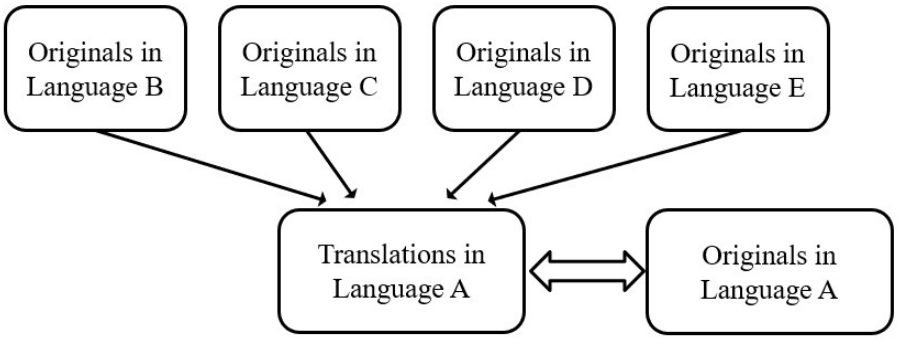


**Fig. 12.5** The model for a monolingual-comparable-cum-parallel corpus

### *12.2.4 Issues and Methodological Challenges*

#### 12.2.4.1 Issues and Challenges Specific to the Design of Parallel Corpora

[6] https://www.alc.manchester.ac.uk/translation-and-intercultural-studies/research/projects/translational-english-corpus-tec/

This section presents an overview of some of the main challenges specific to the design of parallel corpora (for a detailed discussion of more general issues, such as representativeness and balance, copyright clearance[7], and text encoding, see Chap. 1).

The first issue is text availability. As mentioned above, parallel corpora, especially bidirectional ones, tend to be modest in size and are often restricted to a small number of text types. One of the reasons for this is that for any given language pair ($L_X$ and $L_Y$), there is often some kind of asymmetry or imbalance between the two translation directions ($L_X > L_Y$ and $L_Y > L_X$). This imbalance can take several forms: either there are simply fewer texts translated in one direction than in the other (especially when the language pair involves a less "central", or more "peripheral", language), or certain text types are only (or more frequently) translated in one of the two directions. For example, as noted by Frankenberg-Garcia & Santos (2003: 75) in relation to the translation of tourist brochures for the English-Portuguese pair:

> [t]ourist brochures in Portuguese translation are practically non-existent: Portuguese-speaking tourists abroad are expected to get by in other, more widely known languages. In contrast, almost all material destined to be read by tourists in Portuguese-speaking countries comes with an English translation.

To sum up, "translations are heavily biased towards certain genres, but these biases are rarely symmetrical for any language pair" (Mauranen, 2005: 74). In addition, some widely translated text types may be hard to obtain, for obvious confidentiality reasons specific to translation projects carried out by translation agencies and freelance translators (e.g. legal texts or texts translated for internal use only). Finally, there are language pairs for which there are very few parallel texts available (cf., for example, Singh et al., 2000 on building an English-Punjabi parallel corpus). To compensate for data scarcity, there have been a number of initiatives since the early 2000s (Resnik & Smith, 2003) aiming to create mainly non-directional parallel corpora by crawling sites across the web (Chap. 15).

Obtaining detailed metadata is another challenge facing anyone wishing to compile a parallel corpus. In this respect, parallel corpora are clearly lagging behind compared with other corpus types, such as learner corpora, which are more richly documented (Chap. 13). Ideally, the following metadata should be collected (this list is non-exhaustive):

- Source text and target text: author(s)/translator(s), publisher, register, genre, text type, domain, format, mode, intended audience, communicative purpose, publication status, publication date, etc.
- Translation direction, including SL and TL (and their varieties)
- Translation directness: use of a pivot language or not
- Translation directionality: L2>L1 translation, L1>L2 translation, L2>L2 translation, etc.
- Translator: translator's status (professional, volunteer/amateur, student, etc.), translator's occupation, gender, nationality, country of residence, translation expertise (expert *vs*. novice), translation experience (which can be measured in many different ways, e.g. number of years' experience), language background (native and foreign languages), etc.
- Translation task: use of computer-aided translation tools (translation memories, terminological databases) and other tools and resources (dictionaries, forums, corpora, etc.), use of a translation brief (set of translation instructions, including, for instance, use of a specific style guide or in-house terminology), fee per word/line/hour, deadline/time constraints, etc.

[7] Unsurprisingly, it is far from easy to obtain copyright clearance for texts to be included in parallel corpora. For this reason, many parallel corpora are not publicly available (e.g. ENPC, PLECI, Raf Salkie's INTERSECT, P-ACTRES, CroCo).

- Revision/editorial intervention: self- and other-revision, types of revision (e.g. copyediting, monolingual *vs*. bilingual revision), etc.

It is also important to stress here that the concepts of *source language* and *source text* are becoming increasingly blurred. In today's world, some "source" documents are simultaneously drafted in several languages. In multilingual translation projects, there are also cases where there is no single "source" text, as translators translate a given text while accessing some of its already available translations (e.g. when confronted with an ambiguous passage).

Third, there is the issue of alignment, i.e. the process of matching corresponding segments in source and target texts (see Tiedemann, 2011). Software tools can be used to align parallel texts automatically at paragraph, sentence and word level (see, for instance, Hunalign, Varga et al., 2007; GIZA++, Och & Ney, 2003; fast_align, Dyer et al., 2013). Most directional corpora are aligned at sentence level. Different sources of information can be used to match sentences across parallel texts, such as sentence length (in words or characters, normalized by text length), word length, punctuation (e.g. quotation marks), and lexical anchors (e.g. cognates). Some aligners also rely on bilingual dictionaries. Sentence alignment is not a straightforward task, as translators often merge or split sentences when producing the target text. This is referred to as 2:1 and 1:2 alignment links, respectively (see examples in Table 12.1).

**Table 12.1** Splitting and merging source-text sentences in translation

| | | |
|---|---|---|
| 1:1 alignment | Hachez les feuilles de coriandre et mélangez au gingembre. | Chop the coriander leaves and mix with the ginger. |
| Splitting (1:2 alignment) | Hachez les feuilles de coriandre et mélangez au gingembre. | Chop the coriander leaves.<br>Mix with the ginger. |
| Merging (2:1 alignment) | Râpez le gingembre.<br>Coupez les feuilles de coriandre et mélangez au gingembre. | Grate the ginger, then chop the coriander leaves and mix with the ginger. |

As pointed out by Macken et al. (2011: 380), "[t]he performance of the individual alignment tools varies for different types of texts and language pairs and in order to guarantee high quality alignments, a manual verification step is needed". A good option is to use a tool that combines automatic sentence alignment and manual post-alignment correction options, such as the open-source desktop application *InterText editor* (Vondřička, 2014) or the Hypal interface (Obrusnik, 2014). One way of reducing this manual editing step is to combine the output of several aligners, as done for the DPC, where the corpus compilers combined the output of three aligners. The alignment links that were present in the output of at least two aligners were considered as reliable alignment links. All the other links were then checked manually (this shows that manual editing of automatically aligned texts is essential, even when the output of several aligners is combined). Aligners typically generate the following types of XML output: (i) one source-text file, one target-text file and one link file (linking up the source- and target-text segments), (ii) one source-text file and one target-text file, containing the same number of segments, or (iii) a TMX (*Translation Memory eXchange*) file.

Finally, yet another major challenge relating to the compilation of parallel corpora (or any other type of multilingual corpus) is multilingual linguistic annotation (e.g. lemmatization, morphosyntactic annotation, syntactic parsing, semantic tagging; Chap. 2). Johansson (2007: 306) rightly argues that "[t]o go beyond surface forms, we need linguistically annotated corpora that allow more sophisticated studies". However, multilingual annotation raises the following key questions, which echo the more general "universality *vs*. diversity" debate in linguistics (see, for example, Evans & Levinson, 2009):

> If corpora are annotated independently for each language, to what extent is the analysis comparable? If they are provided with some kind of language-neutral annotation (for parts of speech, syntax, etc.), to what extent do we miss language-specific characteristics? (Johansson, 2007: 306)

At present, no definite answers have been found to these questions. As a matter of fact, issues related to multilingual annotation (e.g. whether it should be language-specific or language-neutral, or, more generally, how cross-linguistic comparability can be achieved) have received relatively little attention in contrastive linguistics and translation studies (one notable exception is Neumann, 2013). The language-specific and language-neutral approaches are both used in parallel corpora, the former being more common. In the language-specific approach, researchers rely either on separate annotation tools (one per language involved) or on one single tool that is available for several languages, such as the TreeTagger (Schmid, 1994) or FreeLing (Padró & Stanilovsky, 2012) POS taggers. However, it is important to bear in mind that in these multilingual annotation tools, (i) the annotation systems are not designed to be cross-linguistically comparable: some tags are language-specific (e.g. the RP tag used for English adverbial particles) while, unsurprisingly, "shared" tags display language-specific features (e.g. the TreeTagger JJ tag used for English adjectives does not correspond fully to what the French ADJ tag covers), and (ii) precision and recall ratios (Chap. 2) differ across languages (e.g. for the TreeTagger, they tend to be higher for English than for French). These two factors can potentially jeopardize the contrastive comparability of annotated multilingual data. Great care should therefore be taken when analyzing annotated data in cross-linguistic research (see, for example, Neumann, 2013 and Evert & Neumann, 2017 on the English-German language pair). An interesting language-neutral approach, suggested in Rosen (2010), consists in using an abstract, interlingual hierarchy of linguistic categories mapped to language-specific tags. In the same vein, some researchers have proposed "universal" tagsets, which include tags that accommodate language-specific parts-of-speech (see, for example, Benko, 2016; the MULTEXT-East project[8], with its harmonized morphosyntactic annotation system for sixteen languages; Erjavec's SPOOK specifications[9], with harmonized tagsets for English, French, German, Italian, and Slovenian).

The multilingual annotation of existing parallel corpora is still very basic, being mostly limited to lemmatization and POS tagging. Syntactic annotation will probably become more standard in years to come, given recent advances in multilingual parsing (e.g. Bojar et al., 2012; Volk et al., 2015; Augustinus et al., 2016 on parallel treebanks; see also the Universal Dependencies project[10]).

#### 12.2.4.2 Issues and Challenges Specific to the Analysis of Parallel Corpora

Clearly, compared with monolingual corpora, parallel corpora are lagging behind in terms of size (representativeness is also an issue, as small corpora tend to represent relatively few authors and translators/interpreters). Low-frequency linguistic phenomena may be hard to analyze on the basis of parallel corpora, for sheer lack of sufficient data that would allow reliable generalizations. Researchers in contrastive linguistics and translation studies are therefore often forced to combine several parallel corpora to extract a reasonable amount of data, but this approach raises a number of problems. One is that several confounding variables may be intertwined in the various corpora used, which in turn hinders the interpretability of the results. In Lefer & Grabar (2015), for instance, we relied on two parallel corpora, i.e. verbatim reports of parliamentary debates (Europarl) and interlingual subtitles of oral presentations (TED Talks), so as to investigate the translation of rather infrequent lexical items, namely evaluative prefixes (e.g. *over-* and *super-*). We found marked and seemingly insightful differences

[8] http://nl.ijs.si/ME/V4/msd/html/index.html
[9] http://nl.ijs.si/spook/msd/html-en/
[10] http://universaldependencies.org/

between the translation procedures used in Europarl and TED Talks but were forced to recognize that it was impossible to assess to what extent the observed differences were due to *source-text genre* (parliamentary debates *vs*. oral presentations), *translation modality* (written translation *vs*. subtitling) or *translator expertise* (professional translators *vs*. non-professional volunteers) or, for that matter, a combination of some or all of these factors.

Another issue, also directly related to the interpretability of the results, is the cross-linguistic comparability (or lack thereof) of genres and text types in bidirectional parallel corpora (such as the ENPC, the DPC and CroCo) (see Neumann, 2013). Matching genres or text types cross-linguistically is "by no means straightforward" (Johansson, 2007: 12). We may indeed wonder whether the observed differences reflect genuine cross-linguistic contrasts and/or translation-specific features or whether they are due to fundamental cross-linguistic differences between supposedly similar genres or text types (e.g. research articles or newspaper opinion articles) (cf. Fløttum et al., 2013 on medical research articles in Norwegian). This question cannot be overlooked.

It is also worth pointing out that most parallel corpora are poorly meta-documented (source and target texts and languages, translator, translation task, editorial intervention, etc.), which, unfortunately, can lead researchers to jump to hasty conclusions as regards both cross-linguistic contrasts ("this pattern is due to differences between the two language systems under scrutiny") and features of translated language ("this is inherent in the translation process").

One final point to be made in this section is that parallel corpora (even those whose texts have all been translated by highly-skilled professionals) contain infelicities and even translation errors (to err is human, after all). Researchers may therefore feel uncomfortable with some of the data extracted from parallel corpora. Rather than sweeping erroneous items under the carpet, when in doubt it is probably safer to acknowledge these seemingly infelicitous or erroneous data explicitly. Moreover, looking on the bright side, these infelicities and errors can prove to be highly valuable in applied fields such as bilingual lexicography, foreign language teaching or translator training. In Granger & Lefer (2016), we suggest using them to devise corpus-based exercises, such as the detection and correction of erroneous translations or the translation of sentences containing error-prone items.

**Representative studies**

**Dupont, Maïté, and Sandrine Zufferey. 2017. Methodological issues in the use of directional parallel corpora. A case study of English and French concessive connectives. *International Journal of Corpus Linguistics* 22(2): 270–297.**

In their study, Dupont & Zufferey make an important methodological contribution to the field of corpus-based contrastive linguistics by examining three factors that can potentially affect the nature of the cross-linguistic correspondences found in parallel corpora: register, translation direction and translator expertise. More specifically, they compare three registers (news, parliamentary proceedings, and TED Talks) in two translation directions (from English into French, and vice versa), examining three types of translator expertise (they compare professional, semi-professional and amateur translators). Their study is particularly innovative in that relatively few contrastive corpus studies to date have taken into consideration these influencing factors (especially translation direction and translator expertise), focusing almost exclusively on the source and target linguistic systems under scrutiny. By assuming that the correspondences extracted from parallel corpora are mainly (or solely) due to similarities and differences between the source and target languages, researchers fail to acknowledge the inherently multidimensional nature of translation. In this study, Dupont & Zufferey investigate the translation equivalences between English and French adverbial connectives expressing concession (e.g. *yet*,

*however*, *nonetheless*) across three parallel corpora (PLECI news, Europarl Direct and TED Talk Corpus). Their results indicate that translation choices (and hence, observed cross-linguistic correspondences) depend on the three factors investigated.

**Delaere, Isabelle, and Gert De Sutter. 2017. Variability of English Loanword Use in Belgian Dutch Translations: Measuring the Effect of Source Language, Register, and Editorial Intervention. In *Empirical Translation Studies: New Methodological and Theoretical Traditions*, eds. Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, 81-112. Berlin/Boston: De Gruyter Mouton.**

Delaere & De Sutter's study is situated in the field of corpus-based translation studies. The authors explore three factors that can impact on the linguistic traits of translated language, namely source-language influence, register, and editorial intervention (i.e. revision). They do so through an analysis of English loanwords (*vs*. their endogenous variants) in translated and original Belgian Dutch (e.g. *research & development vs*. *onderzoek en ontwikkeling*). Loanword use is related to a widely investigated topic in translation studies, viz. the normalization hypothesis, which states that translated text is more standard than non-translated text. The starting-point hypothesis of Delaere & De Sutter's study is that overall, translators make more use of endogenous lexemes (a conservative option compared with the use of loanwords), than do non-translators (writers). Relying on the *Dutch Parallel Corpus*, the authors combine two approaches in their study: monolingual comparable (Dutch translated from English and French, alongside original Dutch) and parallel (English to Dutch). As is often the case in corpus-based translation studies, parallel data are used with a view to identifying the source-text items/structures that have triggered the use of a given item/structure in the translations (in this case, the presence of a trigger term in the English source texts, such as *unit*, *job*, or *team*). The authors apply multivariate statistics (profile-based correspondence analysis and logistic regression analysis) to measure the effect of the three factors investigated on the variability of English loanword use. The logistic regression analysis reveals that the effect of register is so strong that it cancels out the effect of source language. Their study convincingly illustrates the need to adopt multifactorial research designs in corpus-based translation studies, as these make it possible to go beyond the monofactorial designs where, typically, only the "translation status" variable is considered (translated *vs*. non-translated).

**Representative corpora**

The *Dutch Parallel Corpus* (Macken et al., 2011) is a 10-million-word bidirectional Dutch-French and Dutch-English parallel corpus (Dutch being the central language). The DPC includes five text types: administrative texts (e.g. proceedings of parliamentary debates, minutes of meetings, and annual reports), instructive texts (e.g. manuals), literature (e.g. novels, essays, and biographies), journalistic texts (news reporting articles and comment articles) and texts for external communication purposes (e.g. press releases and scientific texts). The DPC also features rich metadata, such as publisher, translation direction, author or translator of the text, domain, keywords and intended audience. The corpus is fully aligned at sentence level and is lemmatized and part-of-speech tagged. Unlike many similar corpora, the DPC is available to the research community, thanks to its full copyright clearance.

To date, Europarl (Koehn, 2005) is one of the few parallel corpora to have been used widely in both corpus-based contrastive/translation studies and natural language processing. It contains the proceedings (verbatim reports) of the European Parliament sessions in 21 languages. Its seventh version, released in

2012 by Koehn, includes data from 1996 to 2011[11] and amounts to 600+ million words. Europarl contains two types of European Parliament official reports, viz. written-up versions of spontaneous, impromptu speeches and edited versions of prepared (written-to-be-spoken) speeches. Europarl files contain some metadata tags, such as the speaker's name and the language in which the speech was originally delivered[12]. The main problem, however, is that in part of the corpus, LANGUAGE tags are either missing or inconsistent across corpus files. To solve this problem, Cartoni & Meyer (2012) have homogenized LANGUAGE tags across all corpus files. Thanks to this approach, they have been able to extract *directional* Europarl subcorpora, i.e. subcorpora where the source and target languages are clearly identified (see <https://www.idiap.ch/dataset/europarl-direct>).

### 12.3 Critical Assessment and Future Directions

As shown above, anyone wishing to design and compile a directional parallel corpus faces a number of key issues, such as parallel text availability (especially in terms of text-type variety), access to source text-, translator- and translation task-related metadata, automatic sentence alignment, and linguistic annotation. Relying on existing parallel corpus resources poses its own challenges as well, as present-day parallel corpora tend to be quite small and/or poorly meta-documented and typically cover relatively few text types. Notwithstanding these issues and challenges, parallel corpus research to date has yielded invaluable empirical insights into cross-linguistic contrasts and translation.

There are many hopes and expectations for tomorrow's parallel corpora. There are three ways in which headway can be made in the not too distant future. The first two are related to the design of new parallel corpora, while the third is concerned with a rapprochement between natural language processing and cross-linguistic studies.

First, it is high time we started collecting richer metadata, notably in terms of SL/TL, source and target texts, translator, translation task, and editorial intervention. This will make it possible to adopt multifactorial research designs and use advanced quantitative methods in contrastive linguistics and translation studies much more systematically, thereby furthering our understanding of cross-linguistic contrasts and of the translation product in general.

Second, whenever possible, we should go beyond the inclusion of translated novels, news, and international organizations' legal and administrative texts, and strive for the inclusion of more genres and text types, especially those that are dominant in today's translation market, to which corpus compilers have had limited access to date, for obvious reasons of confidentiality and/or copyright clearance. This also entails compiling corpora representing different translation modalities (e.g. audiovisual translation, interpreting) and translation methods, such as computer-aided translation and post-editing of machine-translated output, as translation from scratch is increasingly rarer today (one

[11] The practice of translating the European Parliament proceedings into all EU languages was ceased in the second half of 2011. The verbatim reports of the plenary sittings are still made available on the European Parliament website, but the written-up versions of the speeches are only published in the languages in which the speeches were delivered.

[12] In this respect, it is important to stress that English is increasingly used as a lingua franca at the European Parliament. In other words, some of the speeches originally delivered in English are in fact given by non-native speakers of English (the same holds, albeit to a lesser extent, for other languages, such as French). This is not a trivial issue, as recent research indicates that the use of English as a Lingua Franca can have a considerable impact on translators' (and interpreters') outputs (see Albl-Mikasa, 2017 for an overview of English as a Lingua Franca in translation and interpreting).

notable exception is literary translation). Including different versions of the same translation would also prove to be rewarding (e.g. draft, unedited, and edited versions of the translated text).

Finally, we need to cross-fertilize insights from natural language processing and corpus-based cross-linguistic studies. This "bridging the gap" can go both ways. On the one hand, cross-linguistic research should take full stock of recent advances in natural language processing, for tasks such as automatic alignment and multilingual annotation. Significant progress has been made in recent years in these areas, but parallel corpora, especially those compiled by research teams of corpus linguists, have not yet fully benefited from these new developments. At present, for instance, very few parallel corpora are syntactically parsed or semantically annotated. On the other hand, natural language processing researchers involved in parallel corpus compilation projects could try to document, whenever possible, meta-information that is of paramount importance to contrastive linguists and translation scholars, such as translation direction (from $L_X$ to $L_Y$, or vice versa) and directness (use of a pivot language or not). In turn, taking this meta-information into account may very well help significantly improve the overall performance of data-driven machine translation systems and other tools relying on data extracted from parallel corpora.

Even though it is quite difficult to predict future developments with any certainty, especially in view of the fact that translation practices are changing dramatically (e.g. human post-editing of machine-translated texts is increasingly common in the translation industry), it is safe to say that compiling and analyzing parallel corpora will prove to be an exciting and rewarding enterprise for many years to come.

### 12.4 Tools and Resources

*Query tools*

*Sketch Engine* by Lexical Computing Ltd is undoubtedly the most powerful tool available to linguists, translation scholars, and lexicographers to analyze bilingual and multilingual parallel corpora. The Sketch Engine interface offers powerful functionality, such as bilingual Word Sketches and automatic bilingual terminology extraction. It contains several ready-to-use sentence-aligned, lemmatized, and POS-tagged parallel corpora, such as DGT-Translation Memory, Eur-Lex, Europarl7 and OPUS2. It is also possible to upload your own parallel corpora in various formats (including XML-based formats used in the translation industry, such as TMX *Translation Memory eXchange* and XLIFF *XML Localization Interchange File Format*), and exploit them in Sketch Engine. A free, simpler version of the tool, *NoSketchEngine*, is freely available to the research community (<https://nlp.fi.muni.cz/trac/noske>).

There are also a number of multilingual parallel concordancers specifically designed for the extraction of data from parallel corpora, such as:

- Anthony's *AntPConc* (available from: <http://www.laurenceanthony.net/software/antpconc/>), a freely available parallel corpus analysis toolkit for concordancing and text analysis using line-break aligned, UTF-8 encoded text files
- Barlow's *ParaConc* (<http://www.athel.com/para.html>), a multilingual concordancer with the following functionality: semi-automatic alignment of parallel texts, parallel searches, automatic identification of translation candidates (called *Hot Words*) and collocate extraction

*Resources*

- OPUS project (Tiedemann, 2012 & 2016), a large collection of freely available parallel corpora: its current version covers 200 languages and language variants and contains over 28 billion tokens, and the collection is constantly growing, in terms of both coverage and size. Compared with other non-directional parallel corpora, OPUS has two major advantages: (i) rather than being restricted to administrative and legal texts (mainly EU and UN), it covers a relatively wide range of other genres and text types, such as user-contributed movie and TV show subtitles, software localization, and multilingual wikis; (ii) a number of poorly-resourced and non-EU language pairs are well represented (albeit often through an indirect translation relationship; e.g. in the $L_X$-$L_Y$ language pair, the two languages $L_X$ and $L_Y$ are both translations from the source language $L_Z$). <http://opus.nlpl.eu/>
- ParaCrawl (*Web-Scale Parallel Corpora for Official European Languages*): parallel corpora for various languages paired with English, created by crawling websites. <https://paracrawl.eu/index.html>
- CLARIN's Key Resource Families – parallel corpora (Fišer et al., 2018): many parallel corpora can be downloaded from the CLARIN webpage. <https://www.clarin.eu/resource-families/parallel-corpora>

*Surveys of available parallel corpora*

A large number of parallel corpora have been mentioned or discussed in this chapter, but it was outside the scope of the present overview to list all available parallel corpora. As a matter of fact, there is as yet no up-to-date digital database documenting all existing parallel corpora (be they bilingual or multilingual, directional or non-directional, developed for cross-linguistic research and/or natural language processing). However, there are some promising initiatives in this direction, such as Mikhailov & Cooper's (2016) survey, the "Universal Catalogue" of the *European Language Resources Association* (ELRA) (<http://www.elra.info/en/catalogues/universal-catalogue/>), CLARIN's overview of parallel corpora (<https://www.clarin.eu/resource-families/parallel-corpora>), and the TransBank project (<https://transbank.info/>).

### 12.5 References for Further Reading

Johansson, Stig. 2007. *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies*. Amsterdam/Philadelphia: John Benjamins.

Johansson's monograph is a must-read for anyone interested in corpus-based contrastive linguistics. The book provides a highly readable introduction to corpus design and use in contrastive linguistics. It also offers a range of exemplary case studies contrasting lexis, syntax, and discourse on the basis of parallel corpus data.

Mikhailov, Mikhail, and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Studies. A guide for research*. London/New York: Routledge.

In this accessible guide for research, Mikhailov & Cooper provide detailed information on parallel corpus compilation and describe a wide range of search procedures that are commonly used in corpus-based contrastive and translation studies. The book also offers a useful survey of some of the available parallel corpora.

Zanettin, Federico. (2012). *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*. London/New York: Routledge.

Zanettin's coursebook is a practical introduction to descriptive and applied corpus-based translation studies. In addition to providing clear background information on the study of translation features in the field, it offers a wealth of useful information on translation-driven (including parallel) corpus design, encoding, annotation, and analysis. Each chapter is enriched with insightful case studies and hands-on tasks.

**12.6 References**


Aijmer, Karin, and Anne-Marie Simon-Vandenbergen. 2003. The discourse particle *well* and its equivalents in Swedish and Dutch. *Linguistics* 41(6): 1123–1161.

Albl-Mikasa, Michaela. 2017. ELF and translation/interpreting. In *The Routledge Handbook of English as a Lingua Franca*, eds. Jennifer Jenkins, Will Baker, and Martin Dewey, 369-384. London/New York: Routledge.

Altenberg, Bengt. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences. In *Out of corpora. Studies in honour of Stig Johansson*, eds. Hilde Hasselgård, and Signe Oksefjell, 249-268. Amsterdam: Rodopi.

Altenberg, Bengt, and Sylviane Granger. 2002. Recent trends in cross-linguistic lexical studies. In *Lexis in Contrast. Corpus-based Approaches*, eds. Bengt Altenberg, and Sylviane Granger, 3-48. Amsterdam/Philadelphia: John Benjamins.

Assis Rosa, Alexandra, Hanna Pięta, and Rita Bueno Maia. 2017. Theoretical, methodological and terminological issues regarding indirect translation: An overview. *Translation Studies* 10(2): 113–132.

Augustinus, Liesbeth, Vincent Vandeghinste, and Tom Vanallemeersch. 2016. Poly-GrETEL: Cross-Lingual Example-based Querying of Syntactic Constructions. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 3549-3554. European Language Resources Association (ELRA).

Baisa, Vít, Jan Michelfeit, Marek Medved, and Miloš Jakubíček. 2016. European Union Language Resources in Sketch Engine. In *Proceedings of tenth International Conference on Language Resources and Evaluation (LREC'16)*. European Language Resources Association (ELRA).

Baker, Mona. 1993. Corpus Linguistics and Translation Studies. Implications and Applications. In *Text and Technology. In Honour of John Sinclair*, eds. Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233-250. Amsterdam: John Benjamins.

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

Baños, Rocío, Silvia Bruti, and Serenella Zanotti. eds. 2013. *Corpus linguistics and Audiovisual Translation: In search of an integrated approach*. Special issue of *Perspectives* 21(4).

Beeby Lonsdale, Allison. 2009. Directionality. In *Routledge Encyclopedia of Translation Studies*, eds. Mona Baker, and Gabriela Saldanha, 84-88. Abingdon: Routledge.

Benko, Vladimír. 2016. Two years of *Aranea*: Increasing counts and tuning the pipeline. In *Proceedings of 10th International Conference on Language Resources and Evaluation (LREC'16)*, 4245-4248. European Language Resources Association (ELRA).

Bernardini, Silvia. 2011. Monolingual comparable corpora and parallel corpora in the search for features of translated language. *SYNAPS* 26: 2–13.

Bernardini, Silvia, Adriano Ferraresi, Mariachiara Russo, Camille Collard, and Bart Defrancq. 2018. Building Interpreting and Intermodal Corpora: A *How-to* for a Formidable Task. In *Making Way in Corpus-Based Interpreting Studies*, eds. Mariachiara Russo, Claudio Bendazzoli, and Bart Defrancq, 21-42. Springer.

Bojar, Ondrej, Zdenek Žabokrtský, Ondrej Dušek, Petra Galuščáková, Martin Majliš, David Marecek, Jiri Maršík, Michal Novák, Martin Popel, and Ales Tamchyna 2012. The joy of parallelism with CzEng 1.0. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, 3921-3928. European Language Resources Association (ELRA).

Bowker, Lynne, and Peter Bennison. 2003. Student Translation Archive and Student Translation Tracking System. Design, Development and Application. In *Corpora in Translator Education*, eds. Federico Zanettin, Silvia Bernardini, and Dominic Stewart, 103-117. Manchester: St. Jerome Publishing.

Cappelle, Bert, and Rudy Loock. 2013. Is there interference of usage constraints? A frequency study of existential *there is* and its French equivalent *il y a* in translated vs. non-translated texts. *Target* 25(2): 252–275.

Cartoni, Bruno, and Thomas Meyer. 2012. Extracting Directional and Comparable Corpora from a Multilingual Corpus for Translation Studies. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC-2012)*, 2132-2137. European Language Resources Association (ELRA).

Cartoni, Bruno, Sandrine Zufferey, and Thomas Meyer. 2013. Using the Europarl corpus for cross-linguistic research. *Belgian Journal of Linguistics* 27: 23–42.

Castagnoli, Sara, Dragos Ciobanu, Natalie Kübler, Kerstin Kunz, and Alexandra Volanschi. 2011. Designing a Learner Translator Corpus for Training Purposes. In *Corpora, Language, Teaching, and Resources: From Theory to Practice*, ed. Natalie Kübler, 221-248. Bern: Peter Lang.

Čermák, František, and Alexandr Rosen. 2012. The case of InterCorp, a multilingual parallel corpus. *International Journal of Corpus Linguistics* 13(3): 411–427.

Cettolo, Mauro, Christian Girardi, and Marcello Federico. 2012. WIT³: Web Inventory of Transcribed and Translated Talks. In *Proceedings of EAMT*, 261-268.

Delaere, Isabelle, and Gert De Sutter. 2017. Variability of English Loanword Use in Belgian Dutch Translations: Measuring the Effect of Source Language, Register, and Editorial Intervention. In *Empirical Translation Studies: New Methodological and Theoretical Traditions*, eds. Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, 81-112. Berlin/Boston: De Gruyter Mouton.

De Sutter, Gert, Marie-Aude Lefer, and Isabelle Delaere. eds. 2017. *Empirical Translation Studies: New Methodological and Theoretical Traditions*. Berlin/Boston: De Gruyter Mouton.

Dupont, Maïté, and Sandrine Zufferey. 2017. Methodological issues in the use of directional parallel corpora. A case study of English and French concessive connectives. *International Journal of Corpus Linguistics* 22(2): 270–297

Dyer, Chris, Victor Chahuneau, and Noah A. Smith. 2013. A Simple, Fast, and Effective Reparameterization of IBM Model 2. In *Proceedings of NAACL-HLT 2013*, 644-648.

Espunya, Anna. 2014. The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation* 48(1): 33–43.

Evans, Nicholas, and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences* 32: 429–492.

Evert, Stefan, and Stella Neumann. 2017.The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In *Empirical Translation Studies: New Methodological and Theoretical Traditions*, eds. Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, 47-80. Berlin/Boston: De Gruyter Mouton.

Ferraresi, Adriano, and Silvia Bernardini. forthcoming. Building EPTIC: A many-sided, multi-purpose corpus of EU Parliament proceedings. In *Parallel Corpora: Creation and Application*, eds. Maria Teresa Sánchez Nieto, and Irene Doval. Amsterdam/Philadelphia: John Benjamins.

Fišer, Darja, Jakob Lenardič, and Tomaž Erjavec. 2018. CLARIN's Key Resource Families. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, 1320-1325.

Fløttum, Kjersti, Trine Dahl, Anders Alvsåker Didriksen, and Anje Müller Gjesdal. 2013. KIAP – reflections on a complex corpus. *Bergen Language and Linguistics Studies* 3(1): 137–150.

Frankenberg-Garcia, Ana, and Diana Santos. 2003. Introducing COMPARA: the Portuguese-English Parallel Corpus. In *Corpora in Translator Education*, eds. Federico Zanettin, Silvia Bernardini, Silvia, and Dominic Stewart, 71-87. Manchester: St. Jerome Publishing.

Granger, Sylviane, and Marie-Aude Lefer. 2016. From general to learners' bilingual dictionaries: Towards a more effective fulfillment of advanced learners' phraseological needs. *International Journal of Lexicography* 29(3): 279–295.

Granger, Sylviane, Jacques Lerot, and Stephanie Petch-Tyson. eds. 2003. *Corpus-based Approaches to Contrastive Linguistics and Translation Studies*. Amsterdam/New York: Rodopi.

Halverson, Sandra L. 2015. The status of contrastive data in Translation Studies. *Across Languages and Cultures* 16(2): 163–185.

Hansen-Schirra, Silvia, Stella Neumann, and Erich Steiner. 2012. *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. Berlin: De Gruyter.

Izquierdo, Marlén, Knut Hofland, and Øystein Reigem. 2008. The ACTRES parallel corpus: an English–Spanish translation corpus. *Corpora* 3(1): 31–41.

Jimenez Hurtado, Catalina, and Silvia Soler Gallego. 2013. Multimodality, translation and accessibility: a corpus-based study of audio description. *Perspectives* 21(4): 577–594.

Johansson, Stig. 2007. *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies*. Amsterdam/Philadelphia: John Benjamins.

Koehn, Philip. 2005. Europarl: A Parallel Corpus for Statistical Machine Translation. In *Proceedings of MT Summit X*, 79-86.

Kruger, Alet, Kim Wallmach, and Jeremy Munday. (eds). 2011. *Corpus-Based Translation Studies. Research and Applications*. London/New York: Bloomsbury.

Kutuzov, Andrey, and Maria Kunilovskaya. 2014. Russian Learner Translator Corpus. Design, Research Potential and Applications. In *Text, Speech and Dialogue. TSD 2014*, eds. Pert Sojka, Ales Horák, Ivan Kopeček, and Karel Pala, 315-323. Springer.

Lapshinova-Koltunski, Ekaterina. 2017. Exploratory analysis of dimensions influencing variation in translation. The case of text register and translation method. In *Empirical Translation Studies: New Methodological and Theoretical Traditions*, eds. Gert De Sutter, Marie-Aude Lefer, and Isabelle Delaere, 207-234. Berlin/Boston: De Gruyter Mouton.

Lefer, Marie-Aude, and Natalia Grabar. 2015. *Super-creative* and *over-bureaucratic*: A cross-genre corpus-based study on the use and translation of evaluative prefixation in TED talks and EU parliamentary debates. *Across Languages and Cultures* 16(2): 187–208.

Levshina, Natalia. 2016. Verbs of letting in Germanic and Romance languages: A quantitative investigation based on a parallel corpus of film subtitles. *Languages in Contrast* 16(1): 84–117.

Macken, Lieve, Orphée De Clercq, and Hans Paulussen. 2011. Dutch Parallel Corpus: A Balanced Copyright-cleared Parallel Corpus. *Meta* 56(2): 374–390.

Mauranen, Anna. 2005. Contrasting languages and varieties with translational corpora. *Languages in Contrast* 5(1): 73–92.

Meurant, Laurence, Maxime Gobert, and Anthony Cleve. 2016. Modelling a Parallel Corpus of French and French Belgian Sign Language. In *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, 4236-4240.

Mezeg, Adriana. 2010. Compiling and Using a French-Slovenian Parallel Corpus. In *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies (UCCTS 2010)*, ed. Richard Xiao, 1-27. Ormskirk: Edge Hill University.

Mikhailov, Mikhail, and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Studies. A guide for research*. London/New York: Routledge.

Neumann, Stella. 2013. *Contrastive Register Variation. A Quantitative Approach to the Comparison of English and German*. Berlin/Boston: De Gruyter Mouton.

Noël, Dirk. 2003. Translations as evidence for semantics: An illustration. *Linguistics*, 41(4): 757-785.

Obrusnik, Adam. 2014. Hypal: A User-Friendly Tool for Automatic Parallel Text Alignment and Error Tagging. *Eleventh International Conference Teaching and Language Corpora*, Lancaster, 20-23 July 2014, 67-69.

Och, Franz Josef, and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics* 29(1): 19–51.

Padró, Lluís, and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards Wider Multilinguality. In *Proceedings of the Language Resources and Evaluation Conference* (LREC-2012). European Language Resources Association (ELRA).

Resnik, Philip, and Noah A. Smith. 2003. The Web as a parallel corpus. *Computational Linguistics* 29(3): 349–380.

Rosen, Alexandr. 2010. Mediating between Incompatible Tagsets. *NEALT Proceedings Series* 10: 53–62.

Russo, Mariachiara, Claudio Bendazzoli, and Annalisa Sandrelli. 2006. Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: Extended analysis of EPIC. *Forum* 4(1): 221–254.

Russo, Mariachiara, Claudio Bendazzoli, and Bart Defrancq. eds. 2018. *Making Way in Corpus-Based Interpreting Studies*. Springer.

Schmid, Helmut. 1994. Probabilistic Part-of-Speech Tagging Using Decision Trees. In *Proceedings of International Conference on New Methods in Language Processing*.

Singh, Sukhdave, Tony McEnery, and Paul Baker. 2000. Building a parallel corpus of English/Panjabi. In *Parallel Text Processing. Alignment and Use of Translation Corpora*, ed. Jean Véronis, 335-346. Kluwer Academic Publishers.

Tiedemann, Jörg. 2011. *Bitext Alignment*. Morgan & Claypool Publishers.

Tiedemann, Jörg. 2012. Parallel Data, Tools and Interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, 2214-2218.

Tiedemann, Jörg. 2016. OPUS – Parallel Corpora for Everyone. *Baltic Journal of Modern Computing* 4(2): 384.

Toury, Gideon. 2012. *Descriptive Translation Studies – And Beyond*. Amsterdam/Philadelphia: John Benjamins.

Uzar, Rafal S. 2002. A corpus methodology for analysing translation. *Cadernos de Tradução* 9(1): 235–263.

Varga, Dániel, Péter Halácsy, András Kornai, Viktor Nagy, László Németh, and Viktor Trón. 2007. Parallel corpora for medium density languages. In *Recent Advances in Natural Language Processing IV: Selected papers from RANLP 2005*, eds. Nicolas Nicolov, Kalina Bontcheva, Kalina, Galia Angelova, and Ruslan Mitkov, 247-258. Amsterdam & Philadelphia: John Benjamins.

Volk, Martin, Anne Ghring, Annette Rios, Torsten Marek, and Yvonne Samuelsson. 2015. *SMULTRON (version 4.0) – The Stockholm MULtilingual parallel TReebank. An English-French-German-Quechua-Spanish-Swedish parallel treebank with sub-sentential alignments*. Institute of Computational Linguistics, University of Zurich.

Vondřička, Pavel. 2014. Aligning parallel texts with InterText. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC 2014)*, 1875-1879.

Waldenfels, Ruprecht von. 2011. Recent developments in ParaSol: Breadth for depth and XSLT based web concordancing with CWB. In *Natural Language Processing, Multilinguality. Proceedings of Slovko 2011, Modra, Slovakia, 20-21 October 2011*, eds. Daniela Majchráková, Daniela, and Rodovan Garabík, 156-162. Bratislava: Tribun EU.

Xiao, Richard. ed. 2010. *Using Corpora in Contrastive and Translation Studies*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Zanettin, Federico. 2012. *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*. London/New York: Routledge.

Ziemski, Michał, Marcin Junczys-Dowmunt, and Bruno Pouliquen. 2016. The United Nations Parallel Corpus v1.0. *Language Resources and Evaluation (LREC'16)*.

Zufferey, Sandrine, and Bruno Cartoni. 2012. English and French causal connectives in contrast. *Languages in Contrast* 12(2): 232–250.