4

How to Use Multilingual Corpora

In this chapter, we will discuss the main characteristics of multilingual corpora, as well as their different uses. First, we will discuss the advantages and disadvantages of two types of multilingual corpora, namely comparable corpora and parallel corpora. We will see that one of the great difficulties inherent in the use of comparable corpora is the need to define a neutral term of comparison, called *tertium comparationis*, which enables us to measure similarities and differences between languages. We will discuss the different possible terms of comparison, depending on the type of research question being considered. Parallel corpora make it possible to compare texts in their original language, with the corresponding translation into one or more languages. We will discuss the particularities of translations as a text genre and show that, due to these particularities, they cannot be used as if they were original language texts. In the rest of the chapter, we will illustrate the use of multilingual corpora in the fields of contrastive linguistics, translation and bilingual lexicography.

4.1. Comparable corpora and parallel corpora

Multilingual studies can be based on two types of corpus data. First of all, comparable corpora contain original texts in different languages. These corpora are built so as to make samples as similar as possible between languages, and to prevent comparison bias. For example, it would be inappropriate to compare French editorials with English dispatches, even though these two types of texts belong to the journalistic genre. Indeed, their many differences in communicational aims and content make them different

in nature, and such disparities could mask differences between languages. It is necessary to neutralize the differences in the type of data used in order to bring out the differences between languages. According to Johansson (1998), the parameters that need to be controlled in order to compare languages include:

– the time when the texts were written;

– their discursive genre (descriptive, argumentative, etc.);

– the type of audience targeted and their field (law, science, etc.).

For example, in order to study the linguistic differences between French and English, one possibility would be to create a comparable corpus of leading articles from journalistic sources with a similar political orientation, published during the same years.

Parallel corpora containing texts in one or more original languages, and their translations into one or more languages, represent the second type of multilingual corpora. It sometimes happens that parallel corpora contain only texts translated into different languages from another language that has not been included in the corpus, or it may occur that the original text cannot be identified among all the texts. As we will see later, the use of corpora in which source languages and target languages remain unidentified poses major problems for contrastive linguistics, due to the special status of translations as a discursive genre (see also Lauridsen (1996)). It would therefore be advisable to use parallel corpora in cases where source and target languages are clearly identified. These are called directional parallel corpora, which refer to the translation direction of source and target languages. Some parallel corpora are even bi-directional, where all the languages they contain are alternately source and target languages. These corpora are particularly valuable for contrastive analyses, since the equivalences between languages are often different in the two directions of translation (see section 4.4).

Both comparable and parallel corpora have many advantages, and also some disadvantages, which we will discuss in the rest of this section. First of all, we should point out that the use of these two types of corpora is not mutually exclusive. On the contrary, the disadvantages of one type can often be counterbalanced, at least partly, by the advantages of the other, and vice

versa. This is why many authors are in favor of carrying out contrastive studies on the basis of both comparable and parallel data, when available corpora and time allow for it. We will see examples of such studies later in this chapter.

The main advantage of comparable corpora is their great simplicity of access. *A priori*, it is possible to create a comparable corpus for any language pair, provided that digitized texts of a comparable nature are available in each language. In the case of languages that have already been the subject of numerous corpora, as is the case for European languages, Aijmer (2008) argues that it is often possible to compile a comparable corpus based on existing monolingual corpora. Another advantage of these corpora is that they only contain language samples originally produced in each of the languages, which guarantees their authenticity when compared to translations.

The major drawback of comparable corpora is that researchers have to find data that are highly similar in different languages, in order to avoid blurring comparisons, as we have already discussed. In addition, usage conventions may vary considerably between languages even when the same text genre exists in both, which makes them difficult to compare.

In addition to the difficulty of identifying suitable corpora, from a linguistic point of view, the main limitation regarding the use of comparable corpora is the need to find a neutral term of comparison, undeformed by the prism of either language. Finally, we should point out that while linguistic features can be identified when comparing words or syntactic structures in different languages, these traits are nonetheless difficult to annotate systematically. Indeed, they require a complex type of linguistic interpretation and analysis on the part of the annotator, which, in many cases, implies that the results of the annotation may differ when performed by several annotators (see Spooren and Degand (2010) for an in-depth discussion of this problem and Cartoni *et al*. (2013a) for an illustration in the field of connectives). In Chapter 7, we will discuss the difficulties associated with manual annotation and possible solutions to improve their reliability.

Unlike comparable corpora, the main advantage of parallel corpora is that they guarantee excellent comparability between languages, since the texts they contain are the same. These corpora make it possible to look for equivalences between words, syntactic structures and discursive phenomena,

without having to set points of comparison. As a result, comparing languages through the use of parallel corpora is greatly simplified in contrast to comparable corpora because annotators can keep a track of translation equivalents without having to annotate syntactic or semantic features. This method is called *translation spotting* in the literature (Véronis and Langlais 2000) and can also be carried out, in part, thanks to automatic tools. On the other hand, the main disadvantage of parallel corpora from the point of view of linguistics is that they contain only a small portion of original texts, whereas the rest of the material is made up of translations. As we will see later in section 4.3, using translations as a mirror of linguistic practices can also have its drawbacks.

Another practical problem associated with the use of such corpora is their limited availability. In fact, not all languages or discursive genres are regularly translated. In most cases, translations correspond to written genres, often related to the administrative or the literary field (Mauranen 1999).

In addition, language pairs that are regularly the subject of direct translations from one into the other are also limited. What is more, these corpora often include a single source language and a single target language, which makes it impossible to generalize results beyond that particular language pair.

In order to overcome certain limitations pertaining to parallel corpora, the ideal would be to work with a bi-directional corpus, where both languages are alternately source and target, since these corpora make it possible to combine the two types of multilingual data discussed above (comparable and parallel). Bi-directional corpora offer the possibility of studying equivalences in both translation directions through the use of parallel corpora.

In addition, these corpora can be used as comparable corpora, produced in very similar situations, when analyzing only the original language portions of the corpus, as illustrated in Figure 4.1.

Certain corpora fulfill these conditions, such as the Europarl Corpus, a corpus of debates at the European Parliament, where each member employs their own language and whose exchanges are later transcribed and translated (see Chapter 5 for a list of these corpora).

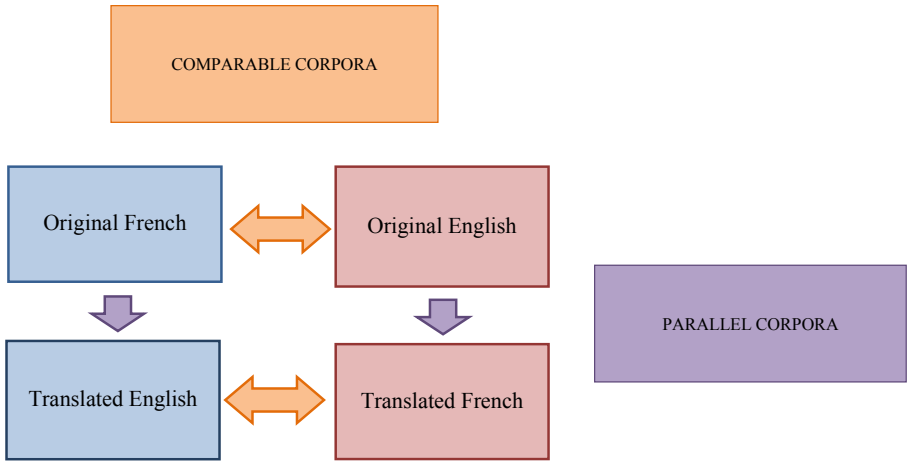

Figure 4.1. *Comparable and parallel corpora that can be retrieved from a bi-directional corpus. For a color version of this figure, see www.iste.co.uk/zufferery/corpus.zip*

4.2. Looking for a *tertium comparationis*

One of the main difficulties inherent in contrastive studies is to find a suitable point of comparison between languages. The problem is that, by nature, comparing two languages implies comparing systems that are partly incommensurable. Therefore, linguists are confronted with the challenge of finding common elements around which languages are close enough so as to be comparable. In fact, relevant differences between languages can only be observed insofar as the latter are compared on the basis of a similar concept or structure. If the objects compared differ in nature, then the differences observed will not be relevant. Let us take a practical example. Observing that mice are smaller than elephants is irrelevant to understanding the morphology of mice or elephants, since these are different animals. On the other hand, observing the differences in size between a Chihuahua and a Saint Bernard is relevant for understanding the different morphologies of dogs.

Contrastivists call this point of comparison between languages *tertium comparationis*. Such a point of comparison should be determined in a neutral manner in relation to the functioning of one language or the other, in order not to bias comparisons. For example, comparing the phonological system of French and German using a list of German phonemes as a starting point

would provide a biased comparison, since the comparison platform is not neutral but established on the basis of one of the language's categories. It is therefore necessary to choose a point of comparison that can be applied to both languages and, which is, as far as possible, neutral. For example, when trying to compare tense categories between German and English, Gast (2012) selected different time spheres along a time axis, including the *Past Tense*, the *Present Perfect*, the *Present Tense* and the *(will) Future*, independently from both languages. He then drew a line corresponding to the time interval that each tense category covered in each language. Through this comparison, he showed that the English *Past Tense* and the German *Präteritum* seem to cover similar time intervals, whereas the English *Present Perfect* and the German *Perfekt* do not have the same function. Thus, while the English *Present Perfect* only applies to events in the near past, the German *Perfekt* covers a wider range which also includes the distant past, as illustrated in Figure 4.2 (adapted from Gast 2012, Figure 5).

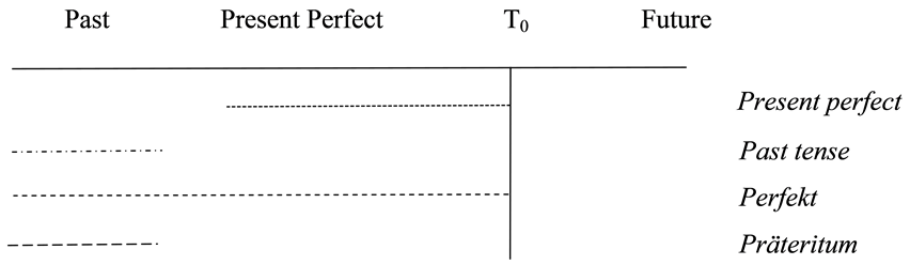

Figure 4.2. *Tertium comparationis for past tenses in English and German*

The choice of the *tertium comparationis* is all the more important since, depending on the chosen point of view, languages may appear to be rather similar or rather different. In his discussion, Krezeszowski (1990) picked the example of squares and rectangles. If these two shapes are compared in terms of their number of sides and angles, they will appear to be identical. However, if they are compared from the point of view of the length of their different sides, they will appear different. The same applies to languages. If French and German tense categories are compared from the point of view of the existence of different tenses for expressing the past, the present and the future, their tense categories will look quite similar. On the contrary, if the comparison concerns the possible uses of the present for designating different temporal references, these two languages will look quite different,

since in German the present form of the tense is used for expressing future events, something which in French is expressed in the future form.

The suitable *tertium comparationis* type for carrying out a study depends on the kind of linguistic elements compared (phonemes, syntactic structures, speech acts, etc.). A distinction can be made between the *tertium comparationis* based on linguistic forms and those based on linguistic function (Gast 2012). In terms of the comparison of functions, some focus on the formal correspondence between functions, for example on the existence of certain categories and syntactic functions, whereas others focus on their semantic equivalence, that is, on the similarity of meanings they make it possible to express (Chesterman 1998).

A *tertium comparationis* determined exclusively by formal criteria, however, is not appropriate, not even for comparing structural elements from different languages. As we have seen previously, English and German have two verbal tenses to refer to the past. Thus, from a structural point of view, we could say that these two languages are similar. However, uses between the two languages are quite different. Conversely, a language may lack a certain linguistic form but still express it through other means. For example, in some languages, speakers verbalize the source of information (which they have acquired either directly by their own perception or indirectly by inference or hearsay) by means of a verbal suffix. These languages have what is called an *evidential* verbal system. This is not the case in French, which does not have such suffixes in its verbal morphology. However, French speakers have other means of indicating sources of information in their statements, in particular by adding phrases such as *il paraît que* (it seems that), *j'en conclus que* (I conclude that) or *je vois que* (I see that). So, to infer from the absence of a suffix that the French language does not make it possible to express belief sources would therefore be wrong. That being said, the fact that languages express certain concepts by different means can, in certain cases, give rise to interesting differences, particularly at the age when these elements are acquired by children and the way in which speakers encode this information. The potential impact of such encoding differences on speaker's cognition is known as *linguistic relativism* (see Deutscher (2011) for an argument in favor of the existence of relativism and McWhorter (2016) for a refutation of such).

In many cases, a *tertium comparationis* based on semantic equivalence appears to be preferable to a *tertium comparationis* based on formal criteria.

However, Krezeszowski (1990) warns us against using translation equivalences as an index of semantic equivalence. As we will see in the next section, translations are not always texts entirely representative of a language. In section 4.3, we will provide some in-depth examples of studies that have used a *tertium comparationis* of the semantic type. Let us bear in mind that for certain research questions, notably in the field of pragmatics and discourse, semantic equivalence is not always appropriate. In fact, while two linguistic forms may be semantically equivalent between languages, they may not be used for achieving the same function. For example, in French, it is very frequent to formulate a request indirectly using a question that refers to the interlocutor's capacity, such as in *Peux-tu me passer le sel?* (Can you pass me the salt?). This same strategy is frequently used in other languages such as English and German, but is not universal. For Polish or Russian speakers, a request formulated in this way would not be understood since this typical association does not exist. As a matter of fact, more direct methods for formulating requests are preferred (Ogiermann 2009). Other differences between languages and cultures are discussed by Jaszczolt (2003) in her article on semantic and pragmatic equivalences between languages.

In summary, in addition to being based on corpora with high comparability, contrastive studies should use neutral points of comparison that make it possible to establish comparisons between linguistic phenomena across languages, which are as relevant and adequate as possible. Depending on the research question, the appropriate equivalence levels will be different.

4.3. Translations as a discursive genre

The main question raised by the use of parallel corpora concerns the status of translations and, more specifically, the possibility of using them as language samples. An important amount of research carried out since the 2000s has shown that translations represent a discursive genre in their own right, and that translations do not fully share the same properties as texts written in original language. This discursive genre is also sufficiently stable and different from others so as to be identifiable using machine learning algorithms for automatic text classification (Ozdowska 2009; Ilisei *et al*. 2010).

One of the reasons why translations represent a stylistic genre different from original texts is that these keep a certain imprint of the source language. Even if translators are language professionals, their lexical, grammatical and stylistic choices are still influenced by what they have to translate. For example, Zufferey and Cartoni (2012) observed that the causal connective *since* is used five times less in English to French translations than in original French texts sharing the same register. The reason for this is that this connective is specific to French and has no exact translation equivalent, even in close languages including English (Degand 2004; Pit 2007). This lack of equivalent means that English text translators are much less likely to use it than an author writing in French. There are many other examples of interference created by the source language in translations.

In addition to these influences, which vary from one source language to another, some authors have hypothesized that translations are so similar (to the point of making up a stylistic genre of its own) due to certain effects related to the translation process itself. These effects might reflect translation universals rather than the variable effects pertaining to the languages involved (Baker 1993; Laviosa-Braithwaite 2009). Over the past 20 years, several potential universals have been discussed in the literature. One of these concerns the tendency of translations to be lexically and syntactically simpler than original texts in the target language (Laviosa-Braithwaite 1997). Another universal concerns their tendency to contain a more standardized, less inventive use of the language than original texts (Baker 1993). In the literature, this universal has been linked to the desire of translators to conform to the standards of the target language as much as possible, in order to produce correct texts, something which hinders their creativity in comparison to authors writing in their original language, who can take more liberties. Finally, another universal concerns the tendency of translations to be more explicit than original texts and, more specifically, to contain a greater number of cohesive markers (Blum-Kulka 1986). In the literature, this universal has been explained by the translators' desire to optimize the readability of translated texts by making explicit the type of coherence relations linking discourse segments. In section 4.5, we will present a study that empirically tested the existence of an explicitation universal by means of a parallel corpus.

For all of these reasons, it is important not to use a parallel corpus as if it were a comparable corpus only containing excerpts in the original language. Despite this limitation, parallel corpora represent valuable resources for a

number of research questions in contrastive linguistics. As a matter of fact, these are the only data that make it possible to establish equivalences between words or expressions in two languages. In addition, comparing words and their translations makes it possible to better understand the whole semantic field that each word or expression covers in a language. In this case, translations act as a mirror reflecting certain properties of the source language (Noël 2003), which are not always visible in monolingual studies. For example, thanks to this technique, Cartoni *et al*. (2013a) identified six different meanings for the English connective *while*, each corresponding to specific translation equivalents in French. The result of such analysis using the *translation spotting* technique also revealed that numerous occurrences of *while* simultaneously expressed a temporal and a contrastive relation, matching the connective *alors que* in French. In contrast, in dictionaries, the temporal and contrastive meanings of *while* are generally presented separately and appear to be mutually exclusive.

To conclude, in order to limit the bias introduced by the use of translations, it is desirable to use bi-directional corpora as far as possible, as well as to study language equivalences in the two directions of translation. We will see examples of such corpora later in this chapter. We will illustrate the fact that these corpora can help us to work simultaneously on comparable and parallel data, and thus exploit the advantages of each, while limiting their bias.

4.4. Multilingual corpora and contrastive linguistics

In its beginnings in the 1950s, contrastive linguistics emerged as a discipline aiming to compare two or more languages with the aim of improving language teaching methods. Indeed, linguists working on language teaching had long observed that mistakes made by learners were often linked to transfers from their mother tongue. This observation justified the systematic study of differences between languages in order to better understand the risk of making mistakes in different learner populations (see, in particular, Lado (1957)). However, many studies in the field of language learning quickly showed that learner mistakes were by far not always associated with differences between their first and their second languages. On the one hand, in certain cases, gaps between languages should lead to transfer effects, which nonetheless do not take place. Conversely, learners produce numerous mistakes, which cannot be explained through transfer phenomena (see Ortega (2014, Chapter 3)) for a detailed discussion of this).

These new data led to a relative abandonment of contrastive studies for several decades. The situation has changed a great deal since the 1990s, thanks to the arrival of corpus linguistics, which made it possible to empirically compare linguistic systems. The data provided by these contrastive corpus-based studies are not only useful in theoretical linguistics for understanding how languages work, but may be helpful for other applications, notably for the development of tools such as bilingual dictionaries (see section 4.6). In this section, we will present a sample of studies which illustrate the usefulness of corpora for carrying out contrastive linguistic studies.

The first case study that we will discuss concerns the French–English language pair and, more specifically, how the verbs *faire* in French and *make* in English work, both of which can be used in causative constructions such as *faire rire* or *make believe*. On an intuitive level, it may seem that these verbs share a similar meaning and perform equivalent functions in both languages. However, by means of an empirical study of both comparable and parallel data, Gilquin (2008) showed that these two verbs are not equivalent.

Gilquin's study is based on the PLECI bi-directional parallel corpus, which contains newspaper articles and fictional texts in English and French. This corpus can be useful both as a comparable corpus and as a parallel corpus, as illustrated in Figure 4.1. Within this corpus, all the occurrences of the verbs *faire* and *make* were retrieved automatically using a bilingual concordancer (see Chapter 5, section 5.7). Since these two verbs have a host of other non-causative uses, the occurrences had to be sorted manually in order to retain only the causative constructions. The data obtained included 109 occurrences of the verb *make* and 355 occurrences of the verb *faire*. In order to establish the similarities and differences between these two verbs, it was necessary to annotate a set of potentially relevant syntactic and semantic features that could represent a suitable *tertium comparationis*. Gilquin chose to annotate the type of subject of the causal construction (animate vs. inanimate, nominal vs. pronominal), as well as the type of infinitive verb used as a complement of *faire* or *make* (volitional vs. non-volitional, transitive vs. intransitive).

The results revealed some similarities between the two languages. First, the distribution of occurrences between nominal and pronominal subjects was very similar. Second, the two verbs were mainly complemented by verbs describing concrete actions such as *partir* rather than existential verbs

like *exist*. Despite these similarities, significant differences were also observed. To begin with, in terms of frequency, the verb *faire* appeared four times more frequently in texts in original French compared to *make* in original English texts, and this was a first indicator that the role of each is not the same in both languages. Furthermore, verbs used with *make* were much more limited than those used with *faire*. The four most frequent verbs in English (*feel*, *look*, *work* and *think*) represented 25% of the occurrences. By contrast, in French, 12 different verbs were needed to reach this same proportion of occurrences. Conversely, some of the uses of the verb *make* seem much more atypical than the verb *faire*. For example, the verb *make* was mainly used in relation to inanimate subjects, which was not the case with *faire*. In sum, although the two verbs have a partly convergent semantic profile, each of them also has frequent uses that are not found in the other language in a similar proportion.

These semantic differences indicate that the verb *make* might not be the best translation choice for *faire*, and the other way around. In order to empirically determine the percentage of correspondences between two words, a mutual correspondence (MC) value can be calculated. This value takes into account the number of translations by the supposed equivalent word compared to the total number of occurrences, in both directions of translation (Altenberg 1999, p. 254). This value is calculated based on the number of occurrences of the two words in translations, which are respectively denoted as A_t and B_t, and then divided by the number of occurrences of these same words in the original texts, denoted as A_s and B_s, and then multiplied by 100 to get a percentage:

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

In the case of the pair made of *faire/make*, the MC value was 15.4%. Such a low value tends to confirm that these two words are not equivalent. In most cases, the causative construction *faire* + *infinitif* in French is translated by an English verb carrying the notion of causality, also called the synthetic causative. For example, the expression *faire taire* is often translated using the verb *to silence*. In the case of the verb *make*, its most frequent translations are the verb *make* as well as paraphrases, for English expressions that cannot be literally translated into French. For example, the sentence "it was the very intensity of her devotion that had ***made her give***

him a softness of upbringing…" was translated by adding "***Par un excès de tendresse****, Lady O'Connell l'éleva avec une faiblesse*...".

In a nutshell, this study made it possible to show that two words which may seem close, and which are often described as translation equivalents in reference tools such as bilingual dictionaries, are in fact partially different from each other. Furthermore, these differences can only emerge on the basis of a quantitative corpus study, which highlights the differences in frequency and context of use.

The second study we present in this section was devoted to the analysis of the different factors that influence translations in parallel corpora. To do this, Dupont and Zufferey (2017) compared the way in which concessive connectives are often treated as translation equivalents in bilingual dictionaries, namely: *however*, *yet*, *nevertheless* and *nonetheless* in English and respectively *pourtant*, *toutefois*, *néanmoins* and *cependant* in French. The authors specifically studied the role of three factors in the observed equivalences: the translation direction (French–English or English–French), the stylistic genre (journalistic texts or parliamentary debates) and the translators' degree of expertise (non-professional volunteers, journalists or qualified translators). For this study, the occurrences of the eight above-mentioned connectives were drawn from three parallel corpora (Europarl for the parliamentary debate genre, a corpus of newspaper articles and the TED corpus of online conferences; see Chapter 5 for a description of these corpora). These occurrences were then manually disambiguated in order to remove occurrences which had not been used as a concessive connective, for example when the connective *yet* was used to indicate a temporal relation.

The results showed that in original texts, the frequency of connectives often vary depending on language register, particularly in English, where the four connectives vary significantly. In French, only the connective *pourtant* varied significantly between journalistic texts and parliamentary debates. An analysis of translations also showed differences between the two genres. For French connectives, the typical translations in the journalistic genre were either the generic connective *but*, or there was an outright absence of a connective in the translation. In the parliamentary debate genre, more specific connectives were used: the connective *however* was a frequent translation for the four French connectives, not to mention *yet* as the translation of *pourtant* and *nevertheless* for *néanmoins*. Such a tendency to omit connectives in the journalistic genre can also be found in English. This

observation can no doubt be explained by the concern for efficacy in this genre, which tends to limit the amount of words used. The other translations were more variable than in the French–English direction.

We can see that the direction of translation is an important factor to take into account when establishing equivalences between languages. Differences between stylistic genres were also visible in the MC values between connectives. Indeed, these values were very low in the journalistic genre, oscillating between 14% and 27%, against 33% and 57% in the parliamentary debate genre, which reflected the above-mentioned more specific translation choices.

The last variation factor analyzed in this study referred to the translator's level of expertise. On the one hand, European Union translators are qualified professionals. On the other hand, the translations provided for TED conferences are carried out by volunteers. Finally, journalistic text translations are generally carried out by journalists, who are language professionals but not translation professionals. For the English–French pair (remember that the TED corpus is unidirectional), these variations enabled the authors to compare the impact of this variable on the translations under scrutiny. The comparison revealed that translation choices were systematically less varied in the TED corpus than in other corpora. The number of zero translations was also significantly lower. This trend reflected the fact that amateur translators are more likely than others to use the source text as a guide and to avoid structural changes as much as possible (see also Lefer and Grabar (2015) for a similar conclusion). In other words, their translations are often more literal than those of professional translators.

In summary, this study showed that the type of equivalences observed between languages can be variable across discourse genres. However, contrary to what happens in monolingual studies, contrastive studies are often performed on data from a single genre – due to the scarcity of multilingual corpora – which does not always make it possible to compare different genres. This study also showed that equivalences between languages should be considered separately for the two translation directions. Finally, the degree of expertise of translators also plays a role in their translation choices, and this factor should therefore be taken into account in the study of parallel corpora.

4.5. Parallel corpora and translation studies

Translation studies is the scientific study of the processes at work in translation, as well as the factors that influence their realization. While translation is a practical and applied discipline, translation studies (translatology or traductology) is a theoretical science. As in the case of contrastive studies, translation studies has benefited from the availability of multilingual corpora, as well as theoretical and methodological advances in corpus linguistics. As we will see in this section, the use of large multilingual corpora makes it possible to carry out quantitative studies on different language pairs simultaneously and, therefore, go beyond the isolated observations that can be made on the basis of individual practice. Later, we will see that the use of the empirical methodology ingrained in corpus analysis can also work as a guide for the translator when it comes to making certain translation choices.

The first study that we present in this section looked into the existence of translation universals. As discussed previously, translations differ in several ways from original texts produced in one language. Translation studies specialists have suggested that a portion of these specificities can stem from the existence of translation universals, that is, from phenomena specifically pertaining to the translation process. One of these universals concerns the supposed propensity of translations to be more explicit (explicitation phenomenon), in terms of cohesion markers, than original texts. This hypothesis has been partly confirmed through corpus studies, performed on a single language pair and limited to one translation direction. Due to these limitations, these studies cannot be generalized to all translations.

In order to overcome this limitation, Zufferey and Cartoni (2014) used the multilingual corpus of parliamentary debates, Europarl, in order to determine whether explicitation phenomena were evenly observable when different variation parameters such as the source and target languages were tested, while keeping the factors of stylistic genre and translation quality constant across language pairs. The main advantage of using the Europarl corpus to carry out this study is that all languages are alternately source and target, and the texts contained in each portion of the corpus deal with very similar subjects, and they were produced under highly similar conditions (parliamentary debates), which guarantees their comparability.

The explicitation hypothesis relates to the number of cohesion markers present in translations, which is assumed to be higher than in original texts. Among these, Zufferey and Cartoni chose to focus on the category of causal connectives. Indeed, their use is very frequent, and often optional. In other words, they can be omitted without creating comprehension difficulties (Murray 1997), something which makes them perfect candidates for testing explicitation phenomena. If translators tend to make optional cohesion markers explicit then the number of causal connectives should significantly increase in translations, when compared to original texts. The methodological challenge of this study consisted of identifying those cases where a causal connective had been added in a translation. To achieve this, the authors looked for occurrences of the four French causal connectives (*parce que*, *car*, *puisque* and *étant donné que*) in the corpus section containing translated French, and checked whether an equivalent connective was also present in the source text, thus carrying out a form of reverse *translation spotting*. If no causal connective was present in the source text, then this would be an example of explicitation. This technique made it possible to count the number of connectives added to French translations from four different source languages: English, German, Spanish and Italian. The results showed that there were many cases of explicitation in translations (connectives had been added despite the lack of any source language indicator in about 7% of the cases), but this rate did not vary significantly depending on the source language. The authors then changed the target language, looking for the three causal connectives, *because*, *since* and *given that* added in English texts translated from French, German, Italian and Spanish. Once again, they observed the recurrent presence of explicitation phenomena but this rate did not vary, regardless of the target language. These results provided a first hint of evidence that explicitation was indeed a regular phenomenon in translations, regardless of the language pair involved.

Furthermore, the authors were able to observe that the explicitation rate varied significantly depending on the causal connective in question. On the one hand, some connectives like *parce que* in French and *because* in English gave rise to very few explicitation cases. On the other hand, causal connectives like *puisque* in French and *given that* in English gave rise to many explicitation cases. The authors attributed this gap to the different semantic profiles of connectives. Those that give rise to explicitation are typically used for introducing a cause presented as already known or easily inferred by the interlocutor, unlike the other connectives which are used for

announcing a new cause for the interlocutor (see Zufferey and Cartoni (2012) for a contrastive study of causal connectives in French and English). This observation reinforces the idea that explicitation reflects the desire of translators to improve text readability, by explicitly showing readers that a piece of information is considered by the author to be already known by the audience or easily accessible.

The second study that we will discuss does not deal with the analysis of translations themselves but with the stylistic analysis of the source text, namely the search for recurring *patterns* and monitoring how these *patterns* are translated. Čermáková (2015) studied the recurring stylistic elements in John Irving's novel *A Widow for One Year* and the way in which these were translated into Finnish and Czech. As we saw in Chapter 3 (section 3.6), corpus linguistics provides analytical tools that are very useful for the study of literary texts. In particular, they help to identify the keywords of a text and to analyze them in context. In the case of literary translation, the author argued that a preliminary stylistic analysis of the translatable material made it possible to identify certain recurring *patterns* that were not easily identifiable through qualitative research, and justified the need to treat them in a systematic manner. Using a concordancer, she analyzed the repeated sequences of words in Irving's novel and found eight-word sequences that were repeated at least three times; 27 sets were identified. She also generated a list of keywords in the novel, using the *British National Corpus* as a reference corpus. A comparison between the word sequences and the keywords revealed that most of the recurring word sequences contained or referred to a keyword (see Chapter 5, section 5.7 on methods for generating a list of keywords).

By analyzing the recurring sequences and the keywords they contained, the author was able to show that these repetitions played a particularly important stylistic role in the novel (which also contains many more repeated sequences than other works by the author), and that these repetitions should be maintained in the translation in order to preserve the spirit of the text. In fact, these sequences made reference in part to the titles of other literary works, and helped to grasp certain intertextuality elements. However, an analysis of the translations of these 27 recurring sequences, both by the Finnish translator and by the Czech translator, showed that they were mostly neutralized by stylistic choices avoiding repetitions. The tendency of translators to avoid repetition is also one of the recurring trends identified in translations, and some translation theorists point to various techniques for

achieving that result (Ben-Ari 1998). However, as in the case of Irving's novel, the presence of repetitions may be an integral part of the work's style and erasing them would certainly involve a form of stylistic loss.

In this way, we can see how corpus linguistics can provide translators with tools that may help them adapt their translation choices on the basis of a better identification of the recurrent linguistic properties at work in a text.

4.6. Parallel corpora and bilingual dictionaries

We have already discussed the importance of corpora for monolingual lexicography in Chapter 3 (section 3.5). In this section, we will refer more specifically to the role of parallel corpora in the creation of bilingual dictionaries. Bilingual dictionaries are essential for foreign-language learners but they are also controversial among language professionals, especially translators. The latter, in particular, criticize bilingual dictionaries for the limited list of equivalences that they provide and the lack of context, which often prevents users from making an appropriate distinction between the different meanings of a word or expression. Finally, as monolingual dictionaries, these dictionaries do not provide any indication regarding the frequency of the different meanings, apart from the order in which they are listed.

To some extent, equivalences between languages obtained through the use of parallel corpora respond to such criticisms. Corpora provide access to a broad context and offer a greater variety of equivalences than dictionaries. What is more, these can be easily classified by frequency, and differentiated according to the textual genres under consideration. In addition, computerized word alignment techniques make it possible to automatically produce bilingual dictionaries (see, for example, McEwan *et al.* (2002)). These same techniques also inspired online dictionaries such as *Linguee*[1] bilingual dictionaries, a resource which is based entirely on parallel corpora drawn from the Internet. The huge advantage of these resources is the diversity of translations they offer and the broad context that accompanies each of them. However, as translations are automatically identified, their accuracy is not guaranteed but requires a critical evaluation on the part of users.

1 Available at: https://www.linguee.com/.

In order to illustrate the importance of corpus data for providing more suitable translation equivalents than those of bilingual dictionaries, in this section, we will discuss a study concerning partially equivalent word pairs in French and in English. Cummins and Desjardins (2002) studied the different meanings of the words *population* in French and *population* in English, as well as fixed expressions such as *plus au moins* in French and *more or less* in English. The authors found that bilingual French–English dictionaries listed these words and expressions so as to convey the idea that these were completely equivalent. Then, resorting to the main monolingual dictionaries, they set up a list of their possible meanings in French and in English. When comparing the two lists, the authors realized that some of the meanings could not be found in the other language. For example, only the French use the term *population* in an emotional sense and in political contexts, and the expression *plus au moins* with a euphemistic sense.

At a second stage of the study, the authors looked for occurrences of these words in French–English comparable corpora. They chose 100 occurrences in each language and annotated them with the different meanings listed in monolingual dictionaries. They confirmed that some of the meanings frequently found in the corpus could not be adequately translated by their "equivalent". For example, 75% of the occurrences of *plus au moins* in the corpus should have been translated using expressions such as *pretty much* or *somewhat* in English, rather than using the expression *more or less*. The authors concluded that bilingual dictionaries do not provide enough information for helping users access the correct translation equivalents.

Many other studies have compared the translation equivalents provided by bilingual dictionaries with equivalents observed in parallel corpora. These studies invariably highlight a discrepancy between the translation equivalents found in dictionaries and in corpus data. In most cases, the equivalents provided by dictionaries are much more limited than the equivalents found empirically, or vice versa, dictionaries sometimes list equivalents that are completely absent from corpus data. We will work on two examples by way of illustration. Degand (2004) studied the causal connectives *puisque* in French and *aangezian* in Dutch, which are treated as equivalent in bilingual dictionaries. However, *puisque* was only translated as *aangezian* in 42% of the occurrences in parallel corpora. An even more striking result, *aangezian* was only translated as *puisque* in 8% of the cases. In another contrastive study on the French and English causal connectives,

Zufferey and Cartoni (2012) found that *puisque* was translated as *since* in only 43% of cases and that *since* was translated as *puisque* in only 23% of cases, whereas these two connectives are usually presented as equivalent in bilingual dictionaries. These examples illustrate the need to integrate corpus-based data in bilingual dictionaries in the future, in order to provide users with a more empirically based view of equivalences between languages.

4.7. Conclusion

In this chapter, we have discussed the different uses of comparable and parallel multilingual corpora. We have seen that their advantages and disadvantages are often complementary, and that it is useful to combine these two types of resources in contrastive linguistics. The study of translation often relies on parallel corpora, but can also make use of comparable corpora of texts translated into different languages, without considering the source language. In the field of translation studies, one of the major aims of such studies is to analyze the features of the translated language, with the purpose of looking for translation universals. We have also shown that corpus analysis methods can be useful for uncovering recurring patterns in a source text and to better adapt the strategies used for its translation. Finally, we argued that parallel corpora have become indispensable resources for the creation of bilingual dictionaries, since they provide rich lists of translation equivalents accompanied by their contexts of use, as well as information concerning their frequency in various genres.

4.8. Revision questions and answer key

4.8.1. *Questions*

1) What type of multilingual corpus (comparable or parallel) seems most suitable for studying the two research questions stated below?

a) What are the similarities and differences between the causal connectives *porque* in Spanish, *parce que* in French and *perché* in Italian?

b) How are the European elections reported in the press in Germany, France and the United Kingdom?

2) What would be a good *tertium comparationis* for the following two research subjects?

a) Comparison of the German and the French consonant systems.

b) Speech acts of thanking in French and Chinese.

3) Why can we say that translations are a full-fledged text genre?

4) What are the parameters to take into account in order to carry out a contrastive study on the use of the indefinite pronouns *on* in French and *man* in German?

5) How could we test the supposed translation universal according to which translations are simpler than original texts by means of a parallel corpora study?

6) What types of equivalences are most likely to be insufficiently dealt with in bilingual dictionaries?

4.8.2. *Answer key*

1) a) In order to study the **similarities and differences between the causal connectives *porque* in Spanish, *parce que* in French and *perché* in Italian**, the use of a parallel corpus offers great advantages. Indeed, such a corpus makes it possible to establish the degree of mutual correspondences between these connectives, by counting the number of times that they can be translated by each other. Nonetheless, the use of this method also involves the risk of having a distorted vision of the functioning of these connectives, due to the translation prism. This study should therefore be supplemented by a semantic and pragmatic analysis on how these connectives work in the original language, by means of comparable corpora. For instance, the use of these connectives could be compared only in the source language section of the parallel corpus.

b) Conversely, to study **the way in which European elections are reported in the press in Germany, France and England**, the use of comparable corpora seems the most judicious choice. Indeed, for this study, it is important to have access to texts that were originally produced in each language, in such a way that they reflect both the linguistic structures of each language and bring out potentially different discourses regarding the same event. A parallel corpus, containing translations, would not be able to meet these two objectives.

2) a) The **comparison of the consonant system in German and French** can be done on the basis of formal rather than functional equivalences. In particular, consonants can be compared on the basis of their articulatory features.

b) In order to compare **speech acts of thanking in French and Chinese**, a *tertium comparationis* based on pragmatic equivalence is necessary. It is not only the words or expressions that should be compared, but also their illocutionary force, that is, the communicative intention of the speaker.

3) Several reasons have been given in the literature for explaining the linguistic specificities of translations. The first type of explanation concerns the influence of the source language, which inevitably leaves traces in translations. Even if translators are language professionals, they are inevitably influenced by the words and linguistic structures they have to translate, which leads them to make different lexical and syntactic choices than those of a speaker writing in their mother tongue. The second category for explaining translation specificities is of a general nature and is based on the supposed existence of translation universals (linguistic phenomena resulting from the very process of translation), regardless of the source and target languages involved. These universals include simplification, explicitation and standardization. All these processes reflect the pedagogical role of translators, who (unconsciously) try to improve the readability of texts.

4) First of all, this study should be carried out by means of a parallel corpus, in order to determine to what extent these two pronouns are translation equivalents or not. More specifically, a bi-directional parallel corpus should be used, since the equivalences are often variable depending on the direction of translation. This analysis of translations should be supplemented by a study on comparable corpora, made up of the two original language sections from the parallel corpus. For this analysis, the important point would be to establish which comparison factors would best highlight their common points and their differences. In this case, the possible factors could be the tense and aspect of the verb following the pronoun, etc. Finally, this study should, wherever possible, include two different discourse genres, in order to measure the extent of the variations between them.

5) The simplification universal implies that translations should be simpler linguistically than the original texts of the same discursive genre. Various lexical and syntactic factors, easily measurable, could contribute to this simplicity. For example, lexical simplicity implies that the number of different words should be smaller than in an original text. This can be measured thanks to the type/token ratio (see Chapter 8). Syntactic simplicity is measured, for example, by the average length of sentences. The number of words used per sentence can also be calculated, even on a corpus that has not been subjected to syntactic annotations.

6) The most problematic equivalence cases for bilingual dictionaries are partial equivalences, just as those we discussed in this chapter for expressions such as *plus au moins* and *more or less*. In these cases, the formal proximity and the identification of certain cases in which these expressions are equivalent may suggest that these words are completely equivalent, when actually they are not. Conversely, false cognates, where meanings are completely different between languages despite a formal resemblance, are easier to identify, since their meaning clearly appears to be different.

4.9. Further reading

Kenning (2010) provides a concise presentation of the similarities and differences between comparable and parallel corpora. The book edited by Sharoff *et al.* (2016) contains many chapters dedicated to the construction, evaluation and use of comparable corpora. Johansson (2007) is an essential reference on the use of multilingual corpora in contrastive linguistics. The question of translation universals is discussed in detail in the work by Mauranen and Kujamäki (2004). The different uses of multilingual corpora for contrastive linguistics and the study of translations are discussed in an accessible way by Mikhailov and Cooper (2016).