How to use corpora for translation

How to use corpora for translation

1. Translation and technology: corpora, computer-assisted translation and machine translation

Translation is an operation that concerns texts: through interlinguistic translation, texts in one language are re-created for delivery in another language. In this chapter we will not be concerned with what this means exactly: issues of equivalence, accuracy, faithfulness, adequacy have occupied philosophers, sociologists, literary scholars and linguists for millennia. Here it is important to point out that, whatever the *skopos*, or purpose, of a translation task (Nord 1997), the *habitus*, or set of dispositions, of a translator (Bourdieu 1977), the socio-cultural *norms* operating at a specific point in time (Toury 1995), and any other constraint operating on this complex process and affecting its success conditions, translation at its most basic entails text understanding and text production. It is no surprise therefore that corpora, namely *text* collections, should be especially relevant to translation, both from the point of view of those who *translate*, and from the point of view of those who *study translation*. Certain types of corpora are more useful for the former, and others for the latter, although substantial overlap exists. The bulk of this chapter will describe translation-relevant types of corpora and the main ways in which they can be used to (learn to) translate, and to study translation. Before we concern ourselves with the foreground, however, it is important to position corpora and corpus use for translation against the wider background of related translation technology.

Virtually all professional translators nowadays, and also most non-professional translators and students of translation, are familiar with translation memories (TM). These resources, that lie at the core of computer-assisted translation (CAT) tools, consist of databases of aligned source text (ST) and target text (TT) segment pairs – where a segment is usually the size of a sentence. CAT tools provide automatic look-up facilities during translation, offering translators partial and complete matches retrieved from the TM whenever (a portion of) a segment they are translating is found in the database of previously translated segments. A related resource is that of *bitexts* (Melby et al 2015): complete ST-TT text pairs, aligned at segment level. Some CAT tools are able to query bitexts, offering users the added value of accessing whole-text contexts for the retrieved matches. Thanks to translation memories and bitexts, 'human knowledge and translation competence [are] captured in machine-processible format' (Melby et al 2015: 668; see also Section 2 on bitexts and parallel corpora).

The same recycling principle and the same textual resources also underlie current approaches to machine translation (MT) systems. *Statistical* machine translation and the more recent and highly successful *neural* machine translation differ substantially from the point of view of the computational techniques they use to process textual data. Yet they are similar in taking advantage of extremely large TMs or bitexts to produce their output. In this sense, both technologies are data- or corpus-driven (Forcada 2017), differently from previous standards that relied on grammar rules (so called *rule-based* approaches).

Corpora can thus be said to be the engine that has propelled the two major transformations we have witnessed since the 1990s in the translation world: CAT and, more recently, MT. However, this role has remained somewhat hidden since the main emphasis has been on the efficient retrieval of translation matches by more or less sophisticated algorithms. While responsibility for reviewing and approving suggestions by CAT tools and for post-editing machine-translated output is bound to remain with the translator, in CAT and MT it is the software that does most of the corpus-related work, and translators may be only vaguely aware of the inner workings of the technology they use daily. In the type of corpus work described in the remainder of this chapter, corpora and corpus users instead take centre stage; efficient retrieval is not a priority, and responsibility for querying corpora and for interpreting results remains with the user. A much wider variety of corpus types than the mere collection of ST-TT pairs thus becomes available. We review these in the next section.

2. Translation-relevant corpus types

Users and their needs

Translation-relevant, or translation-*driven* corpora, using Zanettin's (2012) term, are corpora 'which are created and/or used for some translation-related purpose' (Zanettin 2012:8). Adopting this broad definition, arguably any corpus can become translation-relevant, depending on one's purpose – the practical translation task or research questions one is addressing. In this section we sketch a corpus typology adopting first the viewpoint of the translation practitioner and student, and then that of the translation scholar, laying the bases for the explanation of applied and descriptive/theoretical translation-related purposes that concern us in 3 and 4 below.

Applied needs: corpora for translators and translation students

From the viewpoint of the translator and translation student, the most relevant corpora are bilingual comparable corpora and parallel corpora. Bilingual comparable corpora are collections of texts in two languages – the source and target languages of one's translation task –, that have been assembled adopting similar selection criteria. The relevant selection criteria should at least include similarity in topic and similarity in text type (or genre) both with respect to the source text at hand and with respect to one another. Similarity in topic is essential if one needs terminology and subject-matter information, e.g. what is the term for *bone regeneration* in language x? What is the difference between *bone regeneration* and *bone remodeling*? Similarity of text type is to be prioritized when familiarization with genre conventions is a priority (e.g., what verbs are used to present results in an academic paper? Is *bone (re)growth* a better solution than *bone regeneration* when translating a cosmetic dentistry website into English?). Of course, texts that tick both boxes would be ideal candidates for inclusion, but may not be easy to find when one is dealing with a very specialized subject matter: clarifying in a simple *readme* file how the corpus will be used and what criteria are therefore applied to text collection will save time when collecting texts, lead to a more useful resource when consulting it, and act as a memo for future use, when related reference needs emerge.

Bilingual comparable corpora of this kind do not need to be very large: corpora of about 100,000 words and about 20-40 texts per component may prove useful as a starting

point, and can be further refined and enlarged while carrying out the translation. They can be collected by searching the web and saving documents to separate text files with informative names, within separate folders for the two languages, and can be searched using stand-alone applications such as AntConc (Anthony 2019) (see Chapter X, this volume). An application like BootCaT (Zanchetta 2020) can speed up the collection of web texts, though some compromises on quality or control over contents may have to be made. A similar corpus building tool is available from within the commercial corpus manager and query application SketchEngine (Kilgarriff et al. 2014), which also allows upload of local corpora, provides part-of-speech tagging and lemmatization for many languages, and offers sophisticated search and display options. Depending on the time available and the characteristics of the translation task, the monolingual source language component of a comparable corpus may be dispensed with, while a specialized target language corpus remains, in most cases, indispensable.

A final note is needed on large/general monolingual corpora of the source and/or target language. These corpora are often available in the public domain and easily accessible through dedicated interfaces (such as the *KonText* corpus query interface, Machálek 2020). While they are unlikely to be of help with terminology and genre conventions, they do offer precious support, particularly when interpreting or rendering the creative vs. conventional, ironic or evaluative force of an expression (Partington 2017). Some publicly available general corpora were even constructed according to similar criteria for different languages, and may thus be considered, at a rather high level, comparable – this is the case, for instance, of the *Aranea* corpora (Benko 2014) or the *WaCky* corpora (Baroni et al. 2009).

Moving on to *parallel* corpora, these are collections of bitexts: in other words, source texts aligned segment-by-segment to their translations, or translations aligned to each other. We have already mentioned the role of bitexts and translation memories for automatic retrieval of equivalents in CAT tools and MT engines. But parallel corpora can also be queried through parallel concordancers to observe strategies and retrieve equivalents in context, thus tapping into the translation competence of fellow translators. For instance, the Italian-English Cambridge dictionary online (https://dictionary.cambridge.org/dictionary/italian-english/) lists *apparatus*, *device*, *system*, *mechanism*, *contrivance* and *gear* as equivalents of the Italian word *dispositivo*. A search of the Intercorp parallel corpus (Čermák 2019) returns *mechanism* but also *arrangement* as equivalents of *dispositivo* from administrative/legal texts, and shows how the different equivalents collocate with different adjectives and nouns (*dispositivo giuridico* / *legal mechanism*, *dispositivo di valutazione* / *evaluation arrangement*). There is no doubt that equivalents can be more easily retrieved from parallel than comparable corpora. It is no coincidence that students of translation are keen on using platforms like *Reverso Context* (https://context.reverso.net/) and *Linguee* (https://www.linguee.com), which provide results from TMs. Yet this ease of retrieval comes at a price. First of all, building a parallel corpus takes a long time for locating adequate text pairs and for aligning them. While automatic aligners exist that facilitate the process, painstaking manual correction is almost always needed to obtain a usable resource; a user-friendly application that can be used both for alignment and for correction is *Intertext Editor* (Vondřička 2016). Second, some expertise is needed to make sure that the alignment output format complies with the requirements of the

parallel concordancer of choice (such as ParaConc (2002), AntPConc (2017) and the parallel concordancing facility of the SketchEngine). In practice, building a parallel corpus for reference purposes in translation practice is hardly ever worth the effort. It makes more sense to familiarize oneself, on the one hand, with parallel concordancers and sources of aligned parallel texts (such as the *Opus corpus*, Tiedemann 2012), and, on the other, with self-contained platforms providing access to parallel corpora, such as the KonText corpus query platform.

A final note of caution concerns directionality of translation and the reliability of equivalents found in parallel corpora. Many sources of parallel corpora do not state explicitly the direction of translation (this is the case, for instance, with multilingual text production at the European Union), and even when we know what language is the source and what is the target in a bitext or TM, we cannot rule out that translations differ from related texts originating in the target context due to cultural reasons. To fully exploit the potential of parallel and comparable corpora, these should be used together: parallel corpora (from the public domain) may provide suggestions about translator strategies and translation equivalents, while (self-made) specialized comparable corpora of non-translated target language texts may be used to (dis)confirm the general currency of the choices made by translators (Bernardini and Zanettin 2004; Kenning 2010).

Descriptive/theoretical needs: corpora for translation scholars

The corpus types of greatest relevance to translation scholars are, unsurprisingly, those that include one or more translated components. We have already discussed one such

corpus type, the parallel one, in which translations are set alongside their STs or other translations of the same ST. In general, a prototypical parallel corpus such as the ones described in the previous section, and made of ST-TT pairs, is especially apt at investigating the (hypothesized) decisions made while translating, variously conceptualized as translation shifts (Catford 1965: 73), translation procedures (Vinay and Darbelnet 1995 (1958)) or transfer operations (Klaudy and Karoly 2005). However, depending on one's research questions, this design can be extended in several ways. Rather than text pairs, corpora may contain several target texts to each source: this is the case, for instance, with corpora of learner translations (Castagnoli et al. 2011) and corpora of literary classics for which several published translations exist (Malmkjær 2004). Another popular corpus design is the bilingual bidirectional one, exemplified by the ENPC for Norwegian/English (Johansson 2007), or COMPARA for Portuguese/English (Frankenberg-Garcia and Santos 2003). These corpora are made of two parallel subcorpora: STs in language A and their translations into language B, comparable STs in language B and their translations into A. As suggested by Johansson (2007: 38), analysing a corpus of this type is 'a kind of navigation, where new perspectives may be revealed depending upon the direction of exploration': comparing STs (or "originals") in two languages (as in traditional contrastive studies), STs and TTs in two directions (as in parallel corpora), and originals and translations in the same language(s).

The latter corpus type, known as *monolingual comparable*, has been central to corpus-based translation studies from the very beginning. These corpora are often a combination of an existing non-translated corpus and a translation corpus designed to be

comparable to the former. Examples are the English Comparable Corpus (Laviosa 1997), made of a subset of the British National Corpus and the purpose-built Translational English Corpus (TEC) for English, and the XJU Corpus of Translational Chinese (XCTC) plus the Lancaster Corpus of Mandarin Chinese (LCMC) for Chinese (Xiao and Hu 2015). Even though it might seem counterintuitive to exclude source texts from a corpus meant to study translation, Mona Baker (1993) in fact suggested that translations are first and foremost communicative events of relevance to the *target* language context. Shedding the 'longstanding obsession' (Baker 1993: 237) with source texts and with the myth of equivalence, translation studies, with the support of corpora, could concentrate on the linguistic patterns that are specific to translated texts with respect to comparable non-translated texts in the target language, and arrive at generalizations about more or less universal features of translated texts (a thorough description of which can be found in Laviosa 2002). For instance, several studies have hypothesized that translated texts are more explicit than non-translated texts (Baumgarten et al. 2008). One way to confirm this hypothesis is to compare the frequency of more or less explicit words and structures, such as connectives or premodified/postmodified noun phrases, in translated and non-translated texts.

A monolingual comparable corpus is thus no different from corpora used to study sociolinguistic or register variation, where components representing different language varieties are paired. As in comparative studies in general, the two subcorpora should ideally be comparable to each other along all dimensions but the one under study, in this case the translated/non-translated one. In reality, comparability across cultures is particularly tricky, and researchers adopting this corpus design should use extra care when interpreting their findings as being due to translation rather than any other

variable. Aware of the complexity of their object of study, translation scholars have called for triangulation of different data sources and methods (Malamatidou 2018; Serbina et al. 2015; Wang and Li 2020).

Monolingual comparable corpora provide a bird's-eye view of the general quantitative differences across translated and non-translated texts in the same language, but parallel corpora allow one to zoom in on choices made in distinct acts of translation. Triangulation of these two types of corpora thus fruitfully links the general and the particular, allowing one to argue, for instance, that *explicitness* in translated texts is the result of *explicitation* in the translation process.

3. In practice: how to use corpora to (learn to) translate

In this section we concentrate on the different ways in which corpora of the kinds described in Section 2 can be of use to translators, and provide one example, among the innumerable ones one could make. Many more examples can be found in the references provided in this section. These come mainly from classroom research and practice, since, as suggested in Section 1, corpora are to this day less widely used by professionals than other technological aids (Frérot 2016; Frankenberg-Garcia 2015), despite some signals that the situation may be changing (Gallego-Hernandez 2015).

Kübler and Aston (2010) subdivide an act of translation into three phases: documentation, in which translators familiarize themselves with the source text, the domain and the terminology; drafting, in which, chunk after chunk, the ST is

comprehended and recreated; and revision, in which all aspects of the TT are evaluated: its internal consistency and flow, its adequacy with respect to the ST, its acceptability for the target linguaculture and so on. Specialized source language corpora, including the ST used as a mini-corpus, can be used for understanding the ST domain: wordlists (lists of the most frequent words in the text/corpus), or keyword lists (lists of the most typical words in the text/corpus when compared to another corpus) are particularly useful for this purpose. Concordances from parallel corpora and bilingual comparable corpora can be used to generate and subsequently check hypotheses about target language equivalents in context. In particular, purpose-built corpora are required to check denotational correctness and register appropriateness of specialized terms and phrases (López-Rodríguez and Tercedor-Sánchez 2008), while large general corpora can assist with finding out about more subtle aspects of language use such as evaluation (Munday 2011, Stewart 2009), conventionality (Hoey 2011) and creativity (Philip 2009). Finally, to evaluate the internal and external consistency of the choices made in the TT we can resort to target language corpora, including the TT used as a mini-corpus. Experts recommending the use of corpora in the translation classroom stress the beneficial side effects of using corpora for reference purposes: 'corpora, because they can provide data which is not pre-digested […], allow translators to acquire and apply skills which are after all central to their trade – ones of text interpretation and evaluation' (Kübler and Aston 2009: 503). They also stress their potential for the development of autonomy in the learning and translation processes and of capacity for self-assessment (see also Frérot 2016; Giampieri 2020). For these purposes, learner corpora, particularly if annotated for errors and translation strategies, can offer a further

promising resource, both for teaching and for classroom research (Castagnoli et al. 2011).

Given the impossibility to illustrate all the different ways in which the corpora described in this chapter can help in the translation process, I will briefly describe a case study that brings together reference use and autonomous learning (a fuller account can be found in Frank et al forthcoming). The starting point for this case study is the Italian word *contaminazione* (that the Cambridge English/Italian dictionary translates as *contamination* or *pollution*), and its French dictionary equivalent *contamination*. The definitions provided by Italian and French monolingual dictionaries for the two words are virtually identical: the two words may refer to pollution and corruption, or to a mixture of literary or artistic forms. A translation student may thus conclude that the phrase *contaminazione di stili* (*mixture of styles*), common in Italian texts about fashion and design, can be appropriately translated as *contamination de styles*.

To check if these two phrases are in fact good translation equivalents of each other, we need to first establish if the semantic preferences and prosodies of *contaminazione* and *contamination* are similar (Sinclair 2004). To this aim we look up the two lemmas in two large corpora of Italian and French, the *Araneum Italicum Maius* and the *Araneum Francogallicum Maius* (Benko 2014), that were built approximately at the same time using similar procedures, and can be consulted through the KonText corpus platform (among others). First of all, we observe that the frequencies of the two lemmas are very similar (8.81 and 8.86 per million words respectively). After browsing a few screenfuls of randomly sorted concordance lines to get an informal impression, we obtain lists of

collocates of the two words using the platform default parameters (a span of five words to the left and right, with minimum frequency in the corpus of five, and minimum frequency in the span of three). We then group collocates into semantic sets (semantic preferences), and assess whether the evaluation conveyed is positive or negative. Notice that there is no corpus tool for this: the grouping is done manually, in a word processor or spreadsheet application.

French			Italian		
Collocate	Frequency	logDice	Collocate	Frequency	logDice
radioactive *[radioactive.fem.sing]*	185	9.006	linguaggi *[languages]*	16	4.399
croisée *[cross.fem.sing.]*	180	8.462	sotterranee *[underground.fem.sing]*	7	4.104
virus	290	7.912	falde *[acquifers]*	7	3.979
bactérienne *[bacterial.fem.sing.]*	84	7.794	contaminazione *[contamination]*	8	3.892
VIH *[HIV]*	177	7.779	jazz	13	3.830
microbienne *[microbial.fem.sing]*	61	7.448	generi *[genres]*	11	3.746
prévenir *[prevent]*	193	7.332	falda *[acquifer]*	5	3.588
risque *[risk]*	821	7.248	artistiche *[artistic.fem.plur]*	10	3.588
risques *[risks]*	499	7.216	potabili *[drinkable.plur]*	4	3.572

sols *[soils]*	135	7.213	laicità *[secularism]*	5	3.501
microbiologique *[microbiological]*	47	7.114	Chernobyl	4	3.470
contamination *[contamination]*	83	7.097	acque *[waters]*	21	3.470
fécale *[faecal]*	44	7.045	microrganismi *[microorganisms]*	4	3.278
chimique *[chemical]*	101	7.036	manipolazione *[manipulation]*	5	3.238
souterraines *[underground.fem.plur]*	59	7.021	espressive *[expressive]*	4	3.131
accidentelle *[accidental.fem.sing]*	49	6.990	minzione *[urination]*	3	3.127
croisées *[cross.fem.plur.]*	52	6.928	IPA *[PAH]*	3	3.059
éviter *[avoid.inf]*	478	6.928	sedimenti *[sediments]*	3	3.031
eaux *[waters]*	227	6.904	arti *[arts]*	9	3.012
OGM *[MGO]*	79	6.892	colte *[learned.fem.plur.]*	3	3.010

Table 1. The top 20 collocates of *contamination* and *contaminazione* in two web corpora of French and Italian [with English glosses in square brackets]

French *contamination* has semantic preferences for words denoting dangers (*risque*, *prévenir*), contamination agents (*OGM*, *VIH*), contaminated substances (*eaux*, *sols*), and types of contamination (*croisées*, *chimique*). The semantic prosody is therefore

consistently negative, at least judged from a non-technical point of view. Among the top 20 collocates of Italian *contaminazione*, some refer to contaminated substances (*acque*, *sedimenti*) and agents (*microrganismi*), but types of contamination and dangers are absent, and the only potentially negative collocate is *manipolazione*. Instead, several words are related to creativity and artistic expression (*linguaggi*, *jazz*, *artistiche*, *espressive*, *arti*), which express positive evaluation. Through a much more extensive analysis, Frank et al. are able to confirm that *contaminazione* as mixture of artistic expressions cannot be translated as *contamination*, despite the dictionary definitions. They then search the French corpus for collocates of words related to arts, literature and culture (the French equivalents of words that collocate with *contaminazione* in Italian), and obtain a list of potentially more appropriate equivalents, such as *mélange*, *échange*, *rencontre* and *carrefour*. While such a list could be more easily obtained from a parallel corpus (if available), its reliability might be called into question, given that translators themselves might not be aware of the differences highlighted by the above corpus analysis. Indeed, the starting point for this study was Mélanie Frank's doubts about the appropriateness of the phrase *contamination de styles*. Mélanie, a French student of specialized translation and professional translator, then went on to investigate this hunch with a corpus-assisted study that finally grew into the cited paper, thus transforming an instance of reference use into an autonomous learning experience.

4. In practice: How to use corpora to study translation

In Section 2.2 we have mentioned several corpus types used to study translation and, in passing, some of the research objectives that can be pursued thanks to these corpora.

Adopting the monolingual comparable corpus design, one of the major undertakings in corpus-based translation study has been the attempt to find empirical proof for the existence of typical (or universal) features of translated language. Translated texts have been suggested to be simpler, more explicit, more proper (or conventional), more similar to each other and richer in target-language specific linguistic structures than comparable non-translated texts (Laviosa 2002). While not always conclusive, this evidence has been used to support generalizations about underlying socio-cognitive mechanisms at work in the translation process, as well as in other kinds of bilingual processing (Halverson 2017; Lanstyak and Heltai 2012).

Singling out one hypothesized typical feature of translation, namely explicitness, in this section we show how to carry out a simple comparison aiming to confirm whether translated texts are more explicit than corresponding non-translated (or original) texts in the same language. For reasons of space, the comparison will be limited to two equivalent part-of-speech (POS) patterns: noun phrases premodified by another noun (Noun – Noun sequences), and noun phrases post-modified by a prepositional phrase (Noun – preposition – Noun sequences). According to Biber et al. (1999: 588), 'premodifiers are consistently more condensed than postmodifiers, [and] are much less explicit in identifying the meaning relationship that exists between the modifier and head noun'. This is especially true of nominal premodification, since the meaning relations holding between the two nouns cannot easily be reconstructed if one is not already familiar with the meaning of the phrase as a whole. Consider the phrase "food fight", which out of context could easily be understood to mean "fight over/for food", rather than "fight using food (as a weapon)".

If translators make meanings that are implicit in the source text explicit more often than they make explicit meanings implicit (Klaudy and Karoly 2005), then translated texts will contain more noun phrases with nominal postmodification than comparable non-translated texts, and fewer noun phrases with nominal premodification. Notice that this is a very simplified comparison, that only takes into account two basic noun phrase structures, ignoring alternative structures or structural variation. Furthermore, it is misleading to speak of "structures" in the first place, since the corpus used for this study is lemmatized and POS-tagged, but not syntactically parsed. We in fact compare the frequencies of POS sequences *approximating* the two structures of interest, on the assumption that false positives (sequences matching the query but not the target structure, such as "way bankers" in the phrase "the irresponsible way bankers acted") are similarly frequent in translated and untranslated texts. Keeping these limitations in mind, we can proceed with our analysis.

To limit the variables at play, we restrict our study to a single register (journalistic commentaries from the Intercorp v.12 corpus) and a single source language (Italian). Table 2 shows corpus size in tokens (words, numbers, punctuation marks etc.) and number of occurrences of the two patterns in the two subcorpora. The Log-likelihood significance values and effect size values in table 2 are calculated using Rayson's LL wizard available from http://ucrel.lancs.ac.uk/llwizard.html.

	Original English	English translated from Italian	Log-Likelihood	Bayes factor (BIC)
Number of tokens	152,228	184,465	---	---
Noun – Noun sequences	2,692	2,599	68.23 ($p < 0.0001$)	55.5
Noun – Preposition – Noun sequences	1,558	2,088	9.09 ($p < 0.01$)	-3.63

Table 2. Statistical significance and effect size values for the comparison of the frequencies of the two patterns in original and translated English

Using the original English data as a point of comparison, our results suggest that the compact Noun – Noun sequences are under-represented in English commentaries translated from Italian, while the more explicit Noun – Preposition – Noun sequences are over-represented. Indeed, Log-Likelihood results are statistically significant for both comparisons. Yet effect size results confirm the under-representation of Noun – Noun sequences only, while the negative value obtained for Noun – Preposition – Noun sequences is to be interpreted as confirming the null hypothesis: there is no evidence that the frequencies of this pattern in original and translated English differ. One would tentatively conclude that translators avoid the condensed, implicit Noun – Noun sequence, while they use the more explicit Noun – Preposition – Noun alternative to a similar extent as authors of original English texts.

It should not be forgotten, however, that the translated texts have Italian as a source language. The observed lower frequency of the condensed premodified structure could

be due to the fact that a corresponding Italian structure does not exist, therefore there is no prompt for the English one. This explanation would be coherent with Tirkkonen-Condit's (2004) *unique items hypothesis*, as well as with Halverson's (2017) *gravitational pull hypothesis*, but would be unrelated to the explicitness/explicitation, hypothesis (Blum-Kulka 1986). In order to investigate our hypothesis further, we could therefore vary the source language of the translated texts used in the comparison, for instance substituting Italian with German, a language where a corresponding structure exists, and in which therefore the unique items and gravitational pull hypotheses would not predict lower frequencies of the target structure in translated than original English.

5. Summing up and looking ahead: The future of corpora for translation

The focus in this chapter has been on corpus use for translation practice, teaching and research. I have argued that corpora are a central component of the main technological innovations that have boosted change in translation practice in the last three decades: first Computer-Assisted Translation, and then Machine Translation. Yet general uptake by professional translators has been slower, probably due to the perceived complexity of corpus querying and analysis. The bulk of the chapter has focused on types of corpus resources and ways of exploiting them for translation-related purposes. As concerns the former, I have provided a typology distinguishing bilingual comparable corpora (of use mainly in translation practice), monolingual comparable corpora (of use main in translation research), and bilingual parallel corpora, that are relevant for both practice and description/theory. Moving on to corpus use for translation, the subject of the practice-oriented case study has been the use of bilingual comparable corpora for

reference purposes in translation from Italian into French (based on Frank et al forthcoming). Here I showed how a rather straightforward reference use (checking if two look-alike words in different languages are good translation equivalents in context) can turn into a more structured learning experience thanks to the rich evidence offered by corpora. My research-oriented example focused on frequency of nominal premodification and postmodification in a monolingual comparable corpus of translated and original English texts. The main aim in this case was methodological: I pointed out that alternative explanations are often possible for observed differences, and that data triangulation may be needed to arrive at sounder generalizations.

Looking ahead, there are various ways in which corpora and corpus methods are likely to further contribute to translation teaching, practice and research in the near future. As advances in technology make certain routine translation tasks amenable to Machine Translation treatment, human translation is likely to become more and more specialized, and to require even higher levels of expertise than was previously the case. At the same time, translators may lose access to translation memories, if post-editing of machine translated text is carried out outside of computer-assisted translation environments. Faced with such a fast-changing, highly technologized and specialized professional environment, the need for carefully constructed, documented and curated reference materials is likely to increase, as is the need to learn how to efficiently obtain information from such materials while translating, post-editing, or doing terminology work.

Research-wise, the field is moving in several interesting directions and a number of fascinating hypotheses are being explored. I have already mentioned, albeit in passing, triangulation of data and methods. A related development concerns the creation of corpora of simultaneous and consecutive interpreting, and the corpus-based comparison of interpreted and translated language (Shlesinger and Ordan 2012, Bernardini et al. 2016; Pan 2019). Finally, the relevance of translation to other kinds of discourses is actively being explored. Researchers conceptualizing translation as a type of bilingual language production are trying to single out similarities and differences with respect to contact language varieties and second language use (Kruger and van Rooy 2016; Kolehmainen et al 2014); at the same time, use of translation data in corpus-assisted critical discourse analysis is proposed, to 'escape the contagious nature of dominant discourses' in the search for 'an alternative language with which to argue against established institutional rhetoric' (Baker 2019: 1).

After almost three decades of work in corpus-based translation studies, the centrality of corpus methods to translation is undisputable. Corpora and corpus analysis have improved our understanding of translational behaviour and made translation ever more relevant to linguistic research. At the same time, corpus-based technological innovation in the form of Machine Translation has progressed at such a pace that it is already replacing human translation, at least for specific purposes and in specific settings. As fast and inexpensive – but not necessarily reliable, or creative – translation is provided by machines, the ability to use the different kinds of corpora discussed in this contribution, for translation learning, practice and research, will become even more

important than in the past, to endow human translators with the knowledge they need to outperform machines in that most human task of building bridges between cultures.

Further reading

Beeby, A., Rodríguez Inés, P. and Sánchez-Gijón, P. (eds.) (2009) *Corpus Use and Translating. Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Amsterdam: Benjamins. (This volume is a collection of papers on different aspects of corpus use in the classroom, including reports on corpus use by learners, corpus construction, use of specialized and general corpora, and their use for evaluation purposes. It will be of interest to translator trainers and trainees and researchers in applied linguistics, corpus linguistics and translation studies).

Zanettin, F. (2012) *Translation-Driven Corpora. Corpus Resources for Descriptive and Applied Translation Studies*, Oxford: Routledge. (This handbook covers corpus design, encoding and analysis, with a special focus on multilingual corpora and translation-oriented research questions, providing extensive exemplification and activities).

Ji, M., Oakes, M., Li, D. and Hareide, L. (2016) *Corpus Methodologies Explained. An empirical approach to translation studies*. Oxford and New York: Routledge. (The five chapters that, together with the introduction, make up this volume, investigate some of the central topics in descriptive corpus-based translation studies – machine translation, linguistic variation, style and universals – providing thorough descriptions of relevant theoretical background and methods).

Hu, K. and Kim K. H. (eds) (2020) *Corpus-based Translation and Interpreting Studies in Chinese Contexts. Present and Future*. London: Palgrave Macmillan. (This edited collection makes corpus-based translation studies involving the Chinese language and culture accessible also to non-Chinese speaking researchers. Its four parts cover central themes in descriptive translation studies (translation norms and universals, interpreting, equivalence, and style), as well as touching on the neighbouring fields of critical discourse analysis and cognitive research).

References

Anthony, L. (2017). *AntPConc* (Version 1.2.1) [Computer Software]. Tokyo: Waseda University. Available from https://www.laurenceanthony.net/software.

Anthony, L. (2019) *AntConc* (Version 3.5.8) [Computer Software]. Tokyo: Waseda University. Available from https://www.laurenceanthony.net/software.

Baker, M. (1993) 'Corpus Linguistics and Translation Studies: Implications and Applications', in M. Baker, G. Francis and E. Tognini-Bonelli (eds.) *Text and Technology: in Honour of John Sinclair*. Amsterdam: Benjamins, pp: 233–250.

Baker, M. (2019) 'Rehumanizing the Migrant: The Translated Past as a Resource for Refashioning the Contemporary Discourse of the (Radical) Left', *Palgrave Communications* 6(12): 1–16.

Barlow, M. (2002) ‘ParaConc: Concordance Software for Multilingual Parallel Corpora’, in proceedings of *LREC-2002: Third International Conference on Language Resources and Evaluation*, Las Palmas: ELRA, pp. 20–24.

Baroni, M., Bernardini, S., Ferraresi, A. and Zanchetta, E. (2009) ‘The WaCky Wide Web: A Collection of Very Large Linguistically Processed Web-Crawled Corpora’, *Language Resources and Evaluation* 43(3): 209–26.

Baumgarten, N., Meyer, B. and Özçetin, D. (2008) ‘Explicitness in Translation and Interpreting. A Review and some Empirical Evidence (of an Elusive Concept)’, *Across Languages and Cultures* 9(2): 177–203.

Benko, V. (2014) ‘Compatible Sketch Grammars for Comparable Corpora’. In A. Abel, C. Vettori and N. Ralli (eds.), *Proceedings of the XVI EURALEX International Congress: The User In Focus*, Bolzano/Bozen: Eurac Research, pp. 417–30.

Bernardini, S., Ferraresi, A., and Miličević, M. (2016) ‘From EPIC to EPTIC — Exploring Simplification in Interpreting and Translation from an Intermodal Perspective’, *Meta* 28(1): 61–86.

Bernardini, S. and Zanettin F. (2004) ‘When is a Universal not a Universal? Some Limits of Current Corpus-based Methodologies for the Investigation of Translation

Universals', in A. Mauranen and P. Kujamäki (eds.) *Translation Universals: Do they exist?* Amsterdam: Benjamins, pp. 51–62.

Biber, D., Johansson, S., Leech, G., Conrad, S., and Finegan, E. (1999) *Longman Grammar of Spoken and Written English*, London and New York: Longman.

Blum-Kulka, S. (1986) 'Shifts of Cohesion and Coherence in Translation' in J. House and S. Blum-Kulka (eds) *Interlingual and Intercultural Communication*, Tübingen: Narr, pp 17–35.

Bourdieu, P. (1977) *Outline of a Theory of Practice* (Translated by Richard Nice), Cambridge: Cambridge University Press.

Castagnoli, S., Ciobanu, D., Kunz, K. and Kübler, N. (2011) 'Designing a Learner Translator Corpus for Training Purposes' in N. Kübler (ed.) *Corpora, Language, Teaching and Resources: from Theory to Practice*, Bern: Peter Lang, pp 221–248.

Catford, J.C. (1965) *A Linguistic Theory of Translation*. Oxford: Oxford University Press.

Forcada, M. (2017) 'Making Sense of Neural Machine Translation', *Translation Spaces* 6(2): 291–309.

Frank, N., F. Bartolesi, S. Bernardini and A. Partington (forthcoming) 'Is

Contamination Good or Bad? A Corpus-assisted Case Study in Translating Evaluative Prosody', in A. Ferraresi and R. Pederzoli (eds.) *Mediazioni Special Issue.*

Frankenberg-Garcia, A. (2015) 'Training translators to use corpora hands-on: challenges and reactions by a group of 13 students at a UK university', *Corpora* 10(2), online: https://www.euppublishing.com/doi/full/10.3366/cor.2015.0081.

Frankenberg-Garcia, A. and D. Santos (2003) 'Introducing COMPARA: the Portuguese-English parallel corpus', in D. Stewart, F. Zanettin and S. Bernardini (eds.) *Corpora in Translator Education*, Manchester: St Jerome, pp. 71–87.

Frérot, C. (2016) 'Corpora and Corpus Technology for Translation Purposes and Academic Environments. Major Achievements and New Perspectives' *Cadernos de Tradução* 36(1), pp. 36–61.

Gallego-Hernández, D. (2015) 'The Use of Corpora as Translation Resources: A Study Based on a Survey of Spanish Professional Translators', *Perspectives* 23(3): 375–91.

Giampieri, P. (2020) 'Volcanic Experiences: Comparing Non-corpus-based Translations with Corpus-based Translations in Translation Training', *Perspectives,* online: https://www.tandfonline.com/doi/full/10.1080/0907676X.2019.1705361.

Halverson, S. (2017) 'Gravitational pull in translation. Testing a revised model', in G. De Sutter, M.-A. Lefer and I. Delaere (eds.) *Empirical Translation Studies*

New Methodological and Theoretical Traditions, Berlin: De Gruyter Mouton, pp. 9–46.

Hoey, M. (2011) 'Lexical Priming and Translation' in A. Kruger, K. Wallmach and J. Munday (eds.) *Corpus-Based Translation Studies. Research and Applications*. London: Continuum, pp. 153–168.

Johansson, S. (2007) 'Seeing through Multilingual Corpora. On the use of corpora in contrastive studies', Amsterdam: Benjamins.

Kenning, M.-M. (2010) 'What are Parallel and Comparable Corpora and How can We use Them?' in A. O'Keeffe and M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*, London: Routledge, pp. 487–500.

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P. and Suchomel, V. (2014) 'The Sketch Engine: ten years on', *Lexicography*, 1:7-36. Available from: https://www.sketchengine.eu/.

Klaudy, K. and Károly, K. (2005) 'Implicitation in Translation: Empirical Evidence for Operational Asymmetry in Translation', *Across Languages and Cultures* 6(1): 13–29.

Kolehmainen, L., Meriläinen, L. and Riionheimo, H. (2014) 'Interlingual Reduction: Evidence from Language Contacts, Translation and Second Language Acquisition', in H. Paulasto, L. Meriläinen, H. Riionheimo and M. Kok (eds), *Language Contacts at the Crossroads of Disciplines*, Newcastle: Cambridge Scholars Publishing, pp. 3–32.

Kruger, H. and van Rooy, B. (2016) ‘Constrained language: A Multidimensional Analysis of Translated English and a Non-native Indigenised Variety of English, *English World-Wide* 37(1): 26–57.

Kübler, N. and Aston, G. (2010) ‘Using Corpora in Translation’, in A. O’Keeffe and M. McCarthy (eds.) *The Routledge Handbook of Corpus Linguistics*, London: Routledge, pp. 501–515.

Lanstyák, I. and Heltai, P. (2012) ‘Universals in Language Contact and Translation’, *Across Languages and Cultures* 13(1):99–121.

Laviosa, S. (1997) ‘How Comparable Can 'Comparable Corpora' Be?’, *Target* 9(2): 287–317

Laviosa, S. (2002) *Corpus-based Translation Studies: Theory, Findings, Applications*. Amsterdam: Rodopi.

López-Rodríguez, C.I. and Tercedor-Sánchez, M.I. (2008) ‘Corpora and Students' Autonomy in Scientific and Technical Translation training’, *JoSTrans. The Journal of Specialised Translation* 9, online: https://www.jostrans.org/issue09/art_lopez_tercedor.php.

Machálek, T. (2020) ‘KonText: Advanced and Flexible Corpus Query Interface’, in N.

Calzolari, F. Béchet, P. Blache, et al. (eds.) *Proceedings of the 12th Language Resources and Evaluation Conference*, Marseille: ELRA, pp. 7003–08.

Malamatidou, S. (2018) 'Corpus Triangulation. Combining Data and Methods in Corpus-Based Translation Studies', Oxford: Taylor and Francis.

Malmkjær, K. (2004) 'Translational Stylistics: Dulcken's Translations of Hans Christian Andersen', *Language and Literature: International Journal of Stylistics* 13(1): 13–24.

Melby, A.K., Lommel, A. and Morado Vázquez, L. (2015) 'Bitext', in Chan, S. (ed), *Routledge Encyclopedia of Translation Technology*, Oxford: Routledge, pp. 409–24.

Munday, J. (2011) 'Looming Large: A Cross-Linguistic Analysis of Semantic Prosodies in Comparable Reference Corpora', in A. Kruger, K. Wallmach and J. Munday (eds.) *Corpus-Based Translation Studies. Research and Applications*. London: Continuum, pp. 169–186.

Nord, C. (1997) *Translating as Purposeful Activity*. Manchester: St Jerome.

Pan, J. (2019) 'The Chinese/English Political Interpreting Corpus (CEPIC): A New Electronic Resource for Translators and Interpreters', in I. Temnikova, C. Orasan, G. Corpas Pastor and R. Mitkov (eds.) *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, Varna, Bulgaria, pp.

82–88.

Partington, A. (2017) ‘Evaluative Clash, Evaluative Cohesion and How we Actually Read Evaluation in Texts’, *Journal of Pragmatics* 117:190–203.

Philip, G. (2009) ‘Arriving at equivalence. Making a case for comparable general reference corpora in translation studies’, in A. Beeby, P. Rodríguez Inés and P. Sánchez-Gijón (eds.) *Corpus Use and Translating. Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Amsterdam: Benjamins, pp. 59–73.

Serbina, T., Niemietz, P. and Neumann, S. (2015) ‘Development of a Keystroke Logged Translation Corpus’, in C. Fantinuoli and F. Zanettin (eds.) *New directions in corpus-based translation studies*, Berlin: Language Science Press, pp. 11–33.

Shlesinger, M., and Ordan, N. (2012) ‘More Spoken or More Translated? Exploring a Known Unknown of Simultaneous Interpreting’, *Target* 24(1): 43–60.

Sinclair, J. McH. (2004) *Trust the Text*. London: Routledge.

Stewart, D. (2009) ‘Safeguarding the Lexicogrammatical Environment: Translating Semantic Prosody’, in A. Beeby, P. Rodríguez Inés and P. Sánchez-Gijón (eds.) *Corpus Use and Translating. Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Amsterdam: Benjamins, pp. 29–46.

Tiedemann, J. (2012) ‘Parallel Data, Tools and Interfaces in OPUS’, in N. Calzolari, K. Choukri, T. Declerck et al. (eds.) *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC'2012)*, Istanbul: ELRA, pp. 2214–18.

Tirkkonen-Condit, S. (2004) ‘Unique Items — Over- or Under-represented in Translated Language?’ in A. Mauranen and P. Kujamäki (eds.) *Translation Universals: Do they exist?* Amsterdam: Benjamins, pp. 177–184.

Toury, G. (1995) *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

Vinay, J.-P. and Darbelnet, J. (1995) *Comparative Stylistics of French and English*. Translated by J.C. Sager. Amsterdam: Benjamins.

Vondřička, P. (2016) *Intertext Editor* (Version 1.5). Prague: Charles University. Available from: https://wanthalf.saga.cz/.

Wang, Q. and Li, D. (2020) ‘Looking for translator’s fingerprints: a corpus-based study on Chinese translations of Ulysses’, in Hu, K. (ed.) *Corpus-based translation and interpreting studies in the Chinese context: Present and future*. London: Palgrave Macmillan, pp. 155–179-

Xiao, R. and Hu, X. (2015) *Corpus-Based Studies of Translational Chinese in English-Chinese Translation*, Berlin and Heidelberg: Springer-Verlag.

Zanchetta, E. (2020) *BootCaT. Simple Utilities to Bootstrap Corpora And Terms from the Web (version 1.3).* Forlì: University of Bologna. Available from https://bootcat.dipintra.it/?section=download

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