Chapter 1

Creating and using multilingual corpora in translation studies

Claudio Fantinuoli and Federico Zanettin

1 Introduction

Corpus linguistics has become a major paradigm and research methodology in translation theory and practice, with practical applications ranging from professional human translation to machine (assisted) translation and terminology. Corpus-based theoretical and descriptive research has investigated written and interpreted language, and topics such as translation universals and norms, ideology and individual translator style (Laviosa 2002; Olohan 2004; Zanettin 2012), while corpus-based tools and methods have entered the curricula at translation training institutions (Zanettin, Bernardini & Stewart 2003; Beeby, Rodríguez Inés & Sánchez-Gijón 2009). At the same time, taking advantage of advancements in terms of computational power and increasing availability of electronic texts, enormous progress has been made in the last 20 years or so as regards the development of applications for professional translators and machine translation system users (Coehn 2009; Brunette 2013).

The contributions to this volume, which are centred around seven European languages (Basque, Dutch, German, Greek, Italian, Spanish and English), add to the range of studies of corpus-based descriptive studies, and provide examples of some less explored applications of corpus analysis methods to translation research. The chapters, which are based on papers first presented at the 7th congress of the European Society of Translation Studies held in Germersheim in

Claudio Fantinuoli & Federico Zanettin. 2015. Creating and using multilingual corpora in translation studies. In Claudio Fantinuoli & Federico Zanettin (eds.), *New directions in corpus-based translation studies*, 1–11. Berlin: Language Science Press

July/August 2013[1], encompass a variety of research aims and methodologies, and vary as concerns corpus design and compilation, and the techniques used to analyze the data. Corpus-based research in descriptive translation studies critically depends on the availability of suitable tools and resources, and most articles in this volume focus on the creation of corpus resources which were not formerly available, and which, once created, will hopefully provide a basis for further research.

The first article, by Tatiana Serbina, Paula Niemietz and Stella Neumann, proposes a novel approach to the study of the translation process, which merges process and product data. The authors describe the development of a bilingual parallel translation corpus in which source texts and translations are aligned together with a record of the actions carried out by translators, for instance by inserting or deleting a character, clicking the mouse, or highlighting a segment of text. The second article, by Effie Mouka, Ioannis Saridakis and Angeliki Fotopoulou, is an attempt at using corpus techniques to implement a critical discourse approach to the analysis of translation based on Appraisal Theory. The authors describe the development of a trilingual parallel corpus of English, Greek and Spanish film subtitles, and the analysis focuses on racist discourse. The third article, by Naroa Zubillaga, Zuriñe Sanz and Ibon Uribarri, describes the developments of a trilingual parallel corpus of German, Basque and Spanish literary texts. Spanish texts, which were included when used as relay texts for translating from German into Basque, provide a means for the study of translation directness. In the following article Ekaterina Lapshinova-Koltunski uses a corpus which contains translations of the same source texts carried out using different methods of translation, namely, human, computer aided and fully automated. Her chapter provides an innovative contribution to the description of systematic variation in terms of translation features. Steven Doms investigates the strategies translators use to translate non-human agents in subject position when working from English into Dutch. Finally, Gianluca Pontrandolfo's study addresses the needs of practicing and training legal translators by proposing a trilingual comparable phraseological repertoire, based on COSPE, a 6-million word corpus of Spanish, Italian and English criminal judgments.

Rather than providing a summary of the articles, for which individual abstracts are available, we have chosen to briefly illustrate some of the issues involved in different stages of corpus construction and use as exemplified in the case studies included in this volume.

[1] All selected articles have undergone a rigorous double blind peer reviewing process, each being assessed by two reviewers.

2 Corpus design

The initial thrust to descriptive corpus-based studies (CBS) in translation came in the 1990s, when researchers and scholars saw in large corpora of monolingual texts an opportunity to further a target oriented approach to the study of translation, based on the systemic comparison and contrast between translated and non-translated texts in the target language (Baker 1993). In the wake of the first studies based on the Translation English Corpus (TEC) (Laviosa 1997) various other corpora of translated texts were compiled and used in conjunction with comparable corpora of non-translated texts. Descriptive translation research using multilingual corpora progressed more slowly, primarily because of lack of suitable resources. Pioneering projects such as the English Norwegian Parallel Corpus (ENPC), set up in the 1990s under the guidance of Stig Johansson (see e.g. Johansson 2007) and later expanded into the Oslo Multilingual Corpus, which involved more than one language and issues of bitextual annotation and alignment, were a productive source of studies in contrastive linguistics and translation, but they were not easily replicable because the creation of such resources is more time consuming and technically complex than that of monolingual corpora.[2] Thus, research was initially mostly restricted to small scale projects, often involving a single text pair, and non re-usable resources. However, the last few years have seen the development of some robust multilingual and parallel corpus projects, which can and have been used as resources in a number of descriptive translation studies. Two of these corpora, the Dutch Parallel Corpus (Rura, Vandeweghe & Perez 2008) and the German-English CroCo Corpus (Hansen-Schirra, Neumann & Steiner 2013) are in fact sources of data for two of the articles contained in this volume. Other corpora used in the studies in this volume were instead newly created as re-usable resources.

Typically, a distinction is made between (bi- or multi-lingual) parallel corpora, said to contain source and target texts, and comparable corpora, defined as corpora created according to similar design criteria. However, not only is the terminology somewhat unstable (Zanettin 2012: 149) but the distinction between the two types of corpora is not always clear cut. First, parallel corpora do not

[2] Given the advances in parallel corpus processing behind developments in statistical machine translations, it may appear somewhat surprising that they have not benefited descriptive research more decisively. However, while descriptive and pedagogic research depends on manual analysis and requires data of high quality, research in statistical machine translation privileges automation and data quantity, and thus tools and data developed for machine translation (including alignment techniques and tools, and aligned data), are usually not suitable or available for descriptive translation studies research.

necessarily contain translations. For instance, the largest multilingual parallel corpora publicly available, Europarl and Acquis Communautaire, created by the activity of European Institutions, contain all originals in a legal sense. Second, comparable corpora may have varying degrees of similarity and contain not only "original" texts but also translations. Third, various "hybrid texts" exist in which "translated" text is intermingled with "comparable" text, very similar in terms of subject matter, register etc., but not a translation which can be traced to "parallel" source text. Examples include news translation and text crowdsourcing (e.g. Wikipedia articles in multiple languages), which are generated through "transediting" (Stetting 1989) practices and are thus partly "original writing" and partly translation, possibly from multiple sources.

It may thus be useful to consider the attribute "parallel" or "comparable" as referring to a type of corpus architecture, rather than to the status of the texts as concerns translation. Parallel corpora can thus be thought of as corpora in which two or more components are aligned, that is, are subdivided into compositional and sequential units (of differing extent and nature) which are linked and can thus be retrieved as pairs (or triplets, etc.). On the other hand, comparable corpora can be thought of as corpora which are compared on the whole on the basis of assumed similarity.

A distinctive feature of the corpora described in this volume is their complexity, as most corpora contain more than two subcorpora, often in different languages, and in some cases together with different types of data. Serbina, Niemietz and Neumann's keystroke logged corpus contains original texts and translations, together with the intermediate versions of the unfolding translation process. The corpus is based on keystroke logging and eye-tracking data recorded during translation, editing and post-editing experiments. The log of keystrokes is seen as an intermediate version between source and final translation. The corpus created by Mouka, Saridakis and Fotopoulou is a multilingual and multimodal corpus comprising five films in English together with English, Greek and Spanish subtitles. The films were selected for their related subject matter and contain a significant amount of conversation carried out in interracial communities, and feature several instances of racist discourse. Zubillaga, Sanz and Uribarri describe the design and compilation of Aleuska, a multilingual parallel corpus of translations from German to Basque. The corpus, which collates three subcorpora of literary and philosophical texts, was collected after meticulous bibliographic research. Translation into a minority language, such as Basque, is a complex phenomenon, and this complexity is reflected in the design of the corpus, which includes a subcorpus of Spanish texts used as a relay language in the translation process.

Lapshinova-Koltunski's VARiation in TRAnslation (VARTRA) corpus comprises five sets of translations of the same source texts carried out using different translation methods, together with the source texts and a set of comparable German originals. The first subcorpus of translations is a selection extracted from the Cross-linguistic Corpus (CroCo) (Hansen-Schirra, Neumann & Steiner 2013), which contains human translations together with their source texts from various registers of written language. Since CroCo is a bidirectional corpus, it also contains a set of comparable source texts in German (and their English translations, which however were not needed for this investigation). The second set of German translations contains texts produced by translators with the help of Computer Assisted Translation (CAT) tools, while each of the three remaining subcorpora contains the output of a different machine translation system. The last two articles in this collection focus on corpus analysis rather than on the design and construction of the corpora used, which are described extensively elsewhere. However, it is clear that results are as good as the criteria which guided the creation of the corpora from which they are derived. Doms draws his data from the Dutch Parallel Corpus (DPC), a balanced 10 million word corpus of English, French and Dutch originals and translations, while the data analyzed by Pontrandolfo come from the COrpus de Sentencias PEnales (COSPE), a carefully constructed specialized corpus of legal discourse. COSPE is a trilingual comparable corpus and does not contain translations, though its Italian, English and Spanish subcorpora are extremely similar from the point of view of domain, genre and register.

3 Annotation and alignment

The enrichment of a corpus with linguistic and extra-linguistic annotation may play a decisive part in descriptive studies based on corpora of translations, and are of particular concern to the first four articles, in which research implementation relies to a large extent on annotation. Issues of annotation and alignment come to the fore in the study by Sebine, Niemetz and Neumann, who show how both process and product data can be annotated in XML format in order to query the corpus for various features and recurring patterns. The keylogged data provided by the Translog software are pre-processed to represent individual keystroke logging events as linguistic structures, and these process units are then aligned with source and target text units. All process data, even material that does not appear in the final translation product, is preserved, under the assumption that all intermediate steps are meaningful to an understanding of the translation process.

Bringing together approaches from descriptive translation studies and critical discourse linguistics, Mouka, Saridakis and Fotopoulou address the topic of racism in multimedia translation by creating a time-aligned corpus of film dialogues, and attempting to code and classify instances of racist discourse in English subtitles and their translations in multiple languages. The authors devise a taxonomy of racism-related utterances in the light of Appraisal Theory (Martin & White 2005), and use the ELAN and GATE applications to apply multiple layers of XML, TEI conformant annotation to the multimodal and multilingual corpus. Racism-related utterances in the source and target languages are classified in order to allow for the analysis of register shifts in translation. The subtitles are aligned together into the trilingual parallel corpus as well as synchronized with the audiovisual data, allowing access to the wider context for every utterance retrieved.

Zubillaga, Sanz and Uribarri had to face the challenge of working with a minority language, Basque, for which scarce computational linguistics resources are available, and had therefore to develop their own tools. Research into literary translations from German into Basque involves direct translations from German into Basque but also indirect translation, carried out by going through a Spanish version. In order to observe both texts in the case of direct translations and all three texts for indirect translations, Zubillaga, Sanz and Uribarri have aligned their XML annotated parallel trilingual corpus at sentence level, using a project specific alignment tool.

The features chosen for comparative analysis in Lapshinova-Koltunski's chapter were obtained on the basis of automatic linguistic annotation. All subcorpora were tokenised, lemmatised, tagged with part of speech information, and segmented into syntactic chunks and sentences, and were then encoded in a format compatible with the IMS Open Corpus Workbench corpus management and query tool. Though the set of translations extracted from the CroCo corpus are aligned with their source texts, the five subcorpora of translations are not aligned between them since this annotation level is not necessary for the extraction of the operationalisations used in this study. In this respect, then, VARTRA is treated as a comparable rather than as a parallel corpus.

Dom's data are a collection of parallel concordances drawn from the Dutch Parallel Corpus, and annotation and alignment at sentence level are clearly prerequisites for the type of investigation conducted. Pontrandolfo's COSPE contains criminal judgements in different languages by different judicial systems, and therefore the texts in the three subcorpora cannot be aligned. However, as shown by Pontrandolfo, both researchers and translators can benefit from research based on corpora which are neither linguistically annotated nor aligned.

4 Corpus analysis

Sebine, Niemetz and Neumann offer several examples of possible data queries and discuss how linguistically informed quantitative analyses of the translation process data can be performed. They show how the analysis of the intermediate versions of the unfolding text during the translation process can be used to trace the development of the linguistic phenomena found in the final product. Mouka, Saridakis and Fotopoulou use the apparatus of systemic-functional linguistics to trace register shifts in instances of racist discourse in films translated from English into Greek and Spanish. They also avail themselves of large comparable monolingual corpora in English and Greek as a backdrop against which to evaluate original and translated utterances in their corpus. Zubillaga, Sanz and Uribarri provide a preliminary exploration of the type of searches that can be performed using the Aleuska corpus using the accompanying search engine. They frame their search hypothesis within Toury's (1995) translation laws, finding evidence both of standardisation and interference, in direct as well as in indirect translation.

Lapshinova-Koltunski's chapter is one of the first investigations which compares corpora obtained through different methods of translation to test a theoretical hypothesis rather than to evaluate the performance of machine translation systems. The subcorpora are queried using regular expressions based on part of speech annotation which retrieve words belonging to specific word classes or phrase types. These lexicogrammatical patterns, together with word count statistics, are used as indicators of four hypothesized translation specific features, namely simplification, explicitation, normalisation vs. "shining through", and convergence. While these features have been amply investigated in the literature, the novelty of Lapshinova-Koltunski's study is that the comparison takes into account not only variation between translated and non-translated texts, but also with respect to the method of translation. Preliminary results show interesting patterns of variation for the features under analysis.

Doms analyses 338 parallel concordances containing instances of the English verbs *give* and *show* with an agent as their subject, and their Dutch translations. The analysis was carried out manually by filtering out from search results unwanted instances such as passive and idiomatic constructions, and by distinguishing between human and non-human agents. First, the author provides a discussion of the prototypical features of agents which perform the action with particular verbs, and an overview of the different constraints which certain verbs pose on the use of human and non-human agents in English and Dutch, respectively.

He then zooms in on the two verbs under analysis, and discusses the data from the corpus. Since sentences with action verbs like *give* or *show* and non-human agents are less frequently attested in Dutch than in English, the expectation is that translators will not (always) translate English non-human agents as subjects of *give* and *show* with Dutch non-human agents as subjects of the Dutch cognates of *give* and *show*, *geven* and *tonen*, respectively. Doms describes the choices made by the translators both on a syntactic and semantic level, comparing the translation data with the source-text sentences to verify whether these source-text verbs give rise to different solutions, showing how the translators decided between either primed translations with non-human agents and translations without non-human agents, but with specific Dutch syntactic and semantic patterns which differ from those in the English source texts.

Pontrandolfo presents the results of an empirical study of LSP phraseological units in a specific domain (criminal law) and type of legal genre (criminal judgments), approaching contrastive phraseology both from a quantitative and a qualitative perspective. He describes how four categories of phraseological units, namely complex prepositions, lexical doublets and triplets, lexical collocations and routine formulae, were extracted from the corpus using a mix of manual and automatic techniques. He shows how formulaic language, which plays a pivotal role in judicial discourse, can be analyzed and compared across three languages by means of concordancing software. The final goal of Pontrandolfo's research is to provide a resource for legal translators, as well as for legal experts, which can help them develop their phraseological competence through exposure to real, authentic (con)texts in which these phraseological units are used.

5 Conclusions

Corpus-based translation studies have steadily grown as a disciplinary sub-category since the first studies began to appear more than twenty years ago. A bibliometric analysis of data extracted from the Translation Studies Abstracts Online database shows that in the last ten years or so about 1 out of 10 publications in the field has been concerned with or informed by corpus linguistics methods (Zanettin, Saldanha & Harding 2015). The contributions to this volume show that the area keeps evolving, as it constantly opens up to different frameworks and approaches, from Appraisal Theory to process-oriented analysis, and encompasses multiple translation settings, including (indirect) literary translation, machine (assisted)-translation and the practical work of professional legal translators (and interpreters). Finally, the studies included in the volume expand

the range of application of corpus applications not only in terms of corpus design and methodologies, but also in terms of the tools used to accomplish the research tasks outlined. Corpus-based research critically depends on the availability of suitable tools and resources, and in order to cope properly with the challenges posed by increasingly complex and varied research settings, generally available data sources and out of the box software can be usefully complemented by tools tailored to the needs of specific research purposes. In this sense, a stronger tie between technical expertise and sound methodological practice may be key to exploring new directions in corpus-based translation studies.

References

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis & Elena Tognini-Bonelli (eds.), *Text and technology: In honour of John Sinclair*, 233–250. Amsterdam: John Benjamins.

Beeby, Allison, Patricia Rodríguez Inés & Pilar Sánchez-Gijón. 2009. *Corpus use and translating: Corpus use for learning to translate and learning corpus use to translate*. Amsterdam: John Benjamins.

Brunette, Louise. 2013. Machine translation and the working methods of translators. *Special issue of JosTrans* (19). 2–7.

Coehn, Philipp. 2009. *Statistical machine translation*. Cambridge: Cambridge University Press.

Hansen-Schirra, Silvia, Stella Neumann & Erich Steiner. 2013. *Cross-linguistic corpora for the study of translations. Insights from the language pair English-German*. Berlin: de Gruyter.

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

Laviosa, Sara. 1997. How comparable can "comparable corpora" be? *Target* 9(2). 289–319.

Laviosa, Sara. 2002. *Corpus-based translation studies: Theory, findings, applications*. Amsterdam: Rodopi.

Martin, James Robert & Peter R. R. White. 2005. *The language of evaluation: Appraisal in English*. London: Palgrave Macmillan.

Olohan, Maeve. 2004. *Introducing corpora in translation studies*. London: Routledge.

Rura, Lidia, Willy Vandeweghe & Maribel M. Perez. 2008. Designing a parallel corpus as a multifunctional translator's aid. In *Proceedings of the XVIII FIT World Congress*. Shanghai.

Stetting, Karen. 1989. Transediting – A new term for coping with the grey area between editing and translating. In Graham Caie, Kirsten Haastrup & Arnt Lykke Jakobsen (eds.), *Proceedings from the fourth nordic conference for english studies*, 371–382. Copenhagen: University of Copenhagen.

Toury, Gideon. 1995. *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.

Zanettin, Federico. 2012. *Translation-driven corpora: Corpus resources for descriptive and applied translation studies*. Manchester: St. Jerome Publishing.

Zanettin, Federico, Silvia Bernardini & Dominic Stewart (eds.). 2003. *Corpora in translator education*. Manchester: St. Jerome Publishing.

Zanettin, Federico, Gabriela Saldanha & Sue-Ann Harding. 2015. Sketching landscapes in translation studies. A bibliographic study. *Perspectives: Studies in Translatology* 23(2). 1–22.