THE ROLE OF ELECTRONIC CORPORA IN TRANSLATION TRAINING

Silvana Neshkovska[2]

Abstract: *Corpus linguistics has surely secured its position and status in the world of science nowadays. Its role in linguistic research, and consequently its implications for the linguistic theory and practice are practically indisputable today. Nevertheless, what started capturing researchers' attention in the last decades is the role that corpus linguistics has in the domain of translation studies and training. In fact, corpus linguistics has extended its influence so much that it is safe to claim that providing proper training to trainee translators and doing translation in general is inconceivable and inadmissible without taking full advantage of the benefits of corpora.*

The paper aims to take a closer look at the current research done on the role of corpus linguistics in the sphere of translation studies and translation training by examining closely some of the most recent and relevant studies which have dealt with this issue recently. More specifically, the aim of the paper is to offer an overview of the most salient findings and results obtained from these studies, and eventually to draw conclusions as to how future translators could apply these insights into their practical work in order to secure their competitiveness in the global labour market.

Key words: *corpora, corpus linguistics, translation training*

1. Introduction

In the contemporary world translation is in high demand. Consequently, priority should be given to high quality translation teaching in order to produce highly competent and skillful translators, adept at meeting the increasingly versatile and challenging demands of the global market.

Traditionally translation teaching has been teacher centred and text based. In other words, in such a traditional setting, teachers introduce and explain translation theories and then assign exercises to their students in order to evaluate their performance. This practically means that, generally speaking, students are mere passive recipients who engage in little or no creative thinking and have little or no interaction with their teacher or fellow students.

The aim of this paper is to shed some light on a rather novel method of translation teaching and doing translation. This method of teaching gives priority to students' participation in teaching by placing the focus on student-centred and

2. Ph.D. at "St. Kliment Ohridski" University, Faculty of Education, Bitola, Macedonia, e-mail: silvanakolevska@yahoo.com

autonomous learning. More specifically, the paper discusses the application of corpus linguistics in translation studies, by highlighting the benefits of corpus-based transition studies in the course of translation training as well as while doing translation in general.

Initially, the paper outlines the beginnings of compiling corpora and the emergence of corpus linguistics. Then, it sheds some light on the inception of corpus-based translation studies and their practical implications in the context of translation training and doing translation.

The salience of the paper rests on the fact that by tracing the development and application of corpora, it attempts to prove the validity for the claim that nowadays it is inconceivable and unacceptable to envision and realise the translation training of future translators without taking advantage of corpora and the opportunities that corpora proffer for advancing and alleviating the process of transferring linguistic material from a source language to a target language.

2. A historic overview of corpora and different types of corpora

When the term 'corpus' was originally introduced into the Latin language it meant 'body' (Niladri & Arulmozi, 2018). Nowadays, the term 'corpus' is associated with a collection of written texts or transcribed speech which can serve as a basis for linguistic analysis and description. Nevertheless, one has to acknowledge that compiling and utilising corpora for linguistic research is not a new endeavour at all. On the contrary, it has a long tradition, dating back to medieval times when a lot of scholars, mostly clergymen, were engaged in compiling and investigating corpora for various research purposes. Thus, for instance, there was a considerable tradition of corpus-based linguistic analyses of various kinds occurring in several main fields of scholarship such as biblical and literary studies, lexicography, dialect studies, language education studies and grammatical studies. Understandably, since compiling and analysing corpora was done manually, the researchers were going through a painstakingly long and time-consuming experience. Also, since the analysis of huge bodies of texts was done 'by hand', the analysis was prone to error and was not always exhaustive or easily replicable (Kennedy, 1998). From today's perspective, all these corpora can be referred to as *pre-electronic corpora* (Hofmann, 2004, cited in Lüdeling & Kytö, 2008).

As of 1960s, however, with the advent of computers and information technology, the terrain was set for a brand new type of corpora – *electronic corpora* (Hofmann, 2004, cited in Lüdeling & Kytö, 2008), which present a systematic, planned and structured collection of texts stored in an electronic database, specifically compiled for linguistic analysis. In the case of electronic corpora,

unlike in the pre-electronic corpora, the analysis is carried out at an incredible speed; electronic corpora provide total accountability, accurate replicability, statistical reliability and display an ability to handle huge amounts of data (Kennedy, 1998).

The first electronic corpora were compiled in the 1960s and 1970s. The Brown University Standard Corpus of Present-day American English (*The Brown Corpus*) and Lancaster-Oslo-Bergen Corpus of British English (*LOB*) were in fact the first samples of electronic corpora (Kennedy, 1998). The Brown Corpus was initiated in 1961 and was completed with remarkable speed in 1964. It consisted of approximately one million words and the samples were taken from a large number of text categories from both informative and imaginative prose, excluding verse and drama. The Lancaster-Oslo-Bergen Corpus, on the other hand, was compiled from 1970 to 1978. This corpus of written British English was compiled at the University of Lancaster and the University of Oslo. It also contained one million words of different genres of texts, but what is peculiar about this corpus is that all texts included in it were produced in 1961. In addition, because LOB was compiled one decade after the Brown corpus, the compilers were able to take an advantage of the developments in computer technology. In other words, apart from the general version of the corpus, they produced a partly analysed version with tags to each word and Key Words in Context Concordances (KWIC) (Hu, 2016).

In the 1980s it became obvious that the existing corpora were too small to meet the needs of researchers conducting lexical and semantic analysis. Fortunately, developments in technology for text capture and storage came at the right time and made bigger corpora ("mega-corpora") possible. Thus, by the 1990s corpora of millions words or more became available (Kruger et al., 2011). The Cobuild Corpus, the Longman Corpus Network and the British National Corpus (BNC) are some of them. The British National Corpus, for instance, contains 100 million words of contemporary British English (90 million words are from written texts and 10 million from spoken texts) (Kennedy, 1998). The Corpus of Contemporary American English (COCA) is, in fact, the largest freely-available corpus of English, which contains more than 560 million words of text and is equally divided among spoken texts, fiction, popular magazines, newspapers, and academic texts.

Unsurprisingly, all these existing corpora are not of the same type. Depending on the purpose they have been created for, we can distinguish between *general* and *specialised corpora*. General corpora, as their name suggests, are compiled for unspecified linguistic research and contain texts from different genres and domains. Linguists use them to research the grammar, vocabulary, etc. of a specific language. In contrast, specialised corpora, are designed with special research in mind. Further distinction can be made between *synchronic corpora*,

in which an attempt is made to present the language at a particular time, and *diachronic corpora*, in which a language is depicted over a certain period of time (e.g. Helsinki Corpus, ARCHER, aims at representing an earlier stage or earlier stages of a language). Additionally, some corpora are dubbed *regional corpora* as they represent one regional variety of a language (e.g. Wellington Corpus of Written New Zealand English) and they can be juxtaposed to *corpora containing more than one regional variety*. Then, there are *learner corpora* which aim at representing the language as produced by its learners (e.g. International Corpus of Learner English), which are contrasted with *native speaker corpora*. Distinction can also be made between *multilingual corpora* whose aim is to represent several, at least two, different languages, often with the same text types (for contrastive analyses), as opposed to *monolingual* or *bilingual corpora*. Finally, there exist *spoken corpora* which aim at representing spoken language (e.g. London-Lund Corpus of Spoken English) as well as *written corpora* which comprise only written texts, and *mixed corpora* which include both spoken and written texts.

3. Corpus linguistics

The emergence of all these corpora, from 1950s onwards made it possible for corpus linguistics (CL) to spring to life. According to Nilardi and Arulmozi (2018) corpus linguistics is not a new branch of linguistics; it is rather a new approach to language study which supplies samples and linguistic information for all the branches of linguistics. Crystal (1997) refers to CL as "a body of language texts both in written and spoken form … which being preserved in machine readable form, enables all kinds of linguistic description and analysis" (cited in Nilardi & Arulmozi, 2018). In other words, CL is based on the empirical study of "real life" language use, done with the help of specialised computer software. It is used for the investigation of many different types of linguistic questions (lexical, semantic, syntactic, etc.), and it has been shown that it has a great potential to yield highly interesting, fundamental, and often surprising new insights about language. Namely, CL in the recent decades has become widely used and has become fundamental in lexicography, textbooks writing and language teaching in particular. Furthermore, currently, electronic corpora are often used in the research conducted in sociolinguistics, psycholinguistics, language acquisition, semantics, pragmatics, stylistics, literary study, discourse analysis, forensic linguistics, computational linguistics, lexical studies, grammatical studies, translation studies, contrastive analysis, etc. (Laviosa, 2011).

Obviously, electronic corpora have earned an excellent reputation of being both objective and scientific due to the fact that they rely both on qualitative and quantitative analysis, and, as a result have been put to numerous theoretical and practical uses in a variety of scientific fields.

4. Corpus-based translation studies

Translation studies, which despite its long tradition was established as a separate scientific discipline in the second half of the 20th century, was also quick to recognise the potential of corpora and corpus linguistics for its own purposes. This so-called marriage between descriptive translation studies and corpus linguistics is now known as corpus-based translation studies (Laviosa, 2011). Corpus-based translation studies is focused on investigating the nature of translation as a product and a process by means of corpora, based on the statistical analysis of the features of translated texts in relation to non-translated texts and source texts (Hu, 2016). More specifically, since the mid-1990s a great number of corpora were compiled and investigated purposefully to ascertain: the specific features of translated texts on syntactic, lexical, semantic, and textual levels; translator's style (i.e. translator's choices in the use of lexicon, syntactic structure, punctuation, discourse structures, etc.); translational norms (which are changeable and depend on the historical period in which a particular translation is done); translator training (allows students to better understand the regularities, the patterns of language transfers by observing large numbers of existing translation samples) and interpreting (insights into the features of interpreted texts, interpreting norms, strategies and methods) (Hu, 2016).

The publication of Baker's seminal paper entitled "Corpus Linguistics and Translation Studies: Implications and Applications" (1993) is believed to have instigated the emergence of corpus-based translation studies. Baker (1993, p. 243) predicted that the compilation of various types of corpora of both original and translated texts, together with the development of a corpus-driven methodology, would enable translation scholars to uncover "the nature of translated text as a mediated communicative event" through the investigation of what she then termed "universals" of translation, i.e. linguistic features that occur in translated texts and which are free from the influences of specific language pairs involved in the translation process. Baker (1993, p. 248) insisted that "translated texts record genuine communicative events and as such are neither inferior nor superior to other communicative events in any language."

Naturally, at the very core of corpus-based translation studies is the design and navigation of corpora created not only as sources for the retrieval of translation equivalents or improving the quality and efficiency of the final translation product, but also as repositories of data used to better understand translational processes and language behaviour. Bernardini (2003) distinguishes several distinct types of corpora compiled and used for the purposes of translation studies:

a) **Parallel corpora** comprise the source texts of a language and their target texts in another language, which are aligned at a certain level. In terms of the number

of the languages involved, a parallel corpus can be categorised as a bilingual parallel corpus or a multilingual parallel corpus. According to the direction of translation, however, parallel corpora can be divided into a unidirectional parallel corpus, a bidirectional parallel corpus, and a multidirectional parallel corpus. A unidirectional parallel corpus includes source texts of one language and their target texts into another language. A bidirectional parallel corpus includes the source texts of language A and their target texts in language B and the source texts of language B and their target texts in language A. A multidirectional parallel corpus includes the source texts of one language aligned with their translations of two or more languages.

b) **Comparable corpora** include texts that are comparable at different levels. A comparable corpus can be monolingual, bilingual, and multilingual. A monolingual comparable corpus is composed of the non-translated texts and translated texts in the same language. Texts in the two corpora are similar with regard to registration, language variation, and time span, and the size of the two sub-corpora is roughly the same. A bilingual or multilingual comparable corpus contains texts in two or more languages which are comparable but not in translational relationship to one another. The corpus of this kind is primarily used in contrastive studies between languages.

c) **Translational corpora** consist exclusively of texts translated from one or more languages into a certain language. Generally, a translational corpus is compiled for the investigation of features of translations, translational norms, translators' style, etc. However, it should be used hand in hand with a corpus which contains original texts.[1]

d) **Interpreting corpora** include texts transcribed orthographically from video or audio files with the purpose to investigate interpreting strategies, linguistic features of interpreted texts, interpreting norms, the cognitive process of interpreting, etc.[2]

Irrespective of the specific type of a corpus, as Hu (2016) remarks, it is important to bear in mind that the usage of corpora into translation teaching, has two major

1. The earliest and most influential translational corpus, which was started in 1996 and completed in 1999 and which offers a website for free use by the general public, is the Translational English Corpus (TEC). TEC has ten million words and consists of English biographies, novels, newspaper reports, and magazine articles translated from more than a dozen languages including French, German, Italian, Chinese, etc. TEC was designed to be comparable with the British National Corpus, and it was compiled for investigating the similarities and differences between translated and non-translated English texts (Hu, 2016).

2. The European Parliament Interpreting Corpus compiled by Bologna University is one such corpus. An interpreting comparable corpus collects transcribed interpreted speeches and non-interpreted speeches in the same language which are comparable. This kind of corpus is useful for studies of linguistic features of interpreted texts and interpreting norms.

advantages: (1) automatic extraction and analysis of data, and (2) automatic presentation of abundant translation examples. Both these features are crucial for advancing and enhancing the translation teaching process, and the process of doing translation as well.

5. The usage of corpora in translation training

Considering the fact that virtually all translators nowadays use computers in their everyday work and process texts electronically, it goes without saying that they should be proficient in using corpora for their own specific purposes. Namely, their familiarity with corpora for translation purposes should be instigated even when they still have the status of translator trainees; they need to be taught how to utilise ready-made corpora but also how to compile their own corpora.

Hu (2016) is one of the researchers who recognises the salience of corpora in translation training, claiming that corpora today are "valuable resources/aids not only for translators but also for translation trainees".

The advantage of corpora, according to Beeby et al. (2009), is that they present repositories which can help students fill their knowledge gaps, and which can be used in translators training and second language acquisition either as a means for autonomous learning or a source of materials for classroom use. Thus, for instance, Cosme (2006) provides an overview of corpus-based translation tasks and specific instances that can be used in class. He identifies three kinds of tasks: awareness raising tasks; translation enhancement tasks and production. Beeby et al. (2009) also point out that by means of corpora, translator trainees become more and more aware of the typical mistakes or errors that they make and that different types of corpora lend themselves to different kinds of pedagogic exploration, depending on whether they are monolingual, multilingual, parallel, comparable, general or subject-specific etc. In that context, Bernardini et al. (2003) outline that parallel corpora, for instance, help translators opt for natural, native-like terms and phrases in particular communicative situations. This is very important since trainee translators are also second language learners of a specific language. Also parallel corpora which contain original texts and their translations offer learners the possibility to observe what strategies translators appear to privilege; how they adopt and localise something; omit something; directly transfer something from SL to TL, etc. This observation and analysis helps translator trainees to start developing their own strategies, as well as to realise that different solutions can be appropriate in different situations, text types and registers. For instance, by observing how professional translators have dealt with culture specific terms which can be particularly tricky, they come to realise that depending on factors such as who commissioned the translation, what the purpose of the translated text is, who the target audience is, what

the publisher's guidelines are, etc. they must be equipped with a number of translation strategies in order to render the source text correctly into the target language. According to Pearson (2003), a parallel corpus is useful in revealing the translation strategies adopted by professional translators and in helping students establish their own translation principles, while a comparable corpus can help the translator check whether the terminologies and collocations in translations conform to the norms of target languages and cultures and whether solutions to translation problems are appropriate. Fernandes (2000) (in Hu, 2016) stressed that the use of a parallel corpus can help student translators compare their own works with translations by professionals to find out why certain decisions made in the translation process are ill-advised.

According to Zanettin (1998, p. 618-621), the functions of comparable corpora in translator training lie in three aspects: a) the trainees can evaluate the behaviour of similar textual units in respective languages and select proper target-language equivalents for the source-language words, which are compatible with the linguistic and stylistic norms of the target language, b) a comparable corpus can be used to inform translators of related expressions and terminology concerning specialised research fields, c) a comparable corpus helps students to testify the interrelationship between languages, carry out linguistic comparison, and find out similarities between different languages.

Moreover, both Bernardini et al. (2003) and Beeby et al. (2009) seem to agree that corpora of students' translations are essential as they allow learners to observe their own performance and progress over time, as well as provide a means for identifying areas of difficulties that could be integrated into the curriculum and discussed in depth in class.

The use of corpora also seems to be particularly useful for the improvement of the translator's autonomy and flexibility in translation (Monzó, 2003). Thanks to the use of corpora in translation teaching, students involve themselves in the learning process by collecting and evaluating texts, extracting terminologies, and establishing correspondence between different languages, which coincides with the highly advocated principles of "autonomy," "motivation," and "authenticity" and the idea that translator education is "a process of socialization in a professional community." Moreover, the use of corpora in translation teaching provides opportunities for the development of the students' innovation ability and problem-solving abilities. As Bernardini et al. (2003) contend, the greatest pedagogical value of corpora lies in their "thought-provoking" rather than "question-answering" potential. In other words, students should be trained to develop their own hypothesis about textual data and to devise their own strategies for extracting information from corpora and eventually decide on the interpretation of the data they have found in the corpus.

Finally, apart from using ready-made corpora, translation trainees should be involved in creating their own corpora as well. The benefits of that can be multifarious depending on what their purpose is. For instance, students can compile a corpus of their own translated texts which can be used to study the features of translations done by students and track students' learning process so as to make it more efficient. Besides, students can compile a disposable corpus, which is created for a specific translation task only. In order to do that successfully they need to be acquainted with all the factors that they need to take into consideration in the process of compiling the corpus such as the types of genres to be included in the corpus; the length not just of the corpus but of the samples to be included in it; the proportion of speech vs. writing that will be included; the educational level, gender, and dialect backgrounds of speakers and writers included in the corpus; and the types of contexts from which samples will be taken, etc. (Hu, 2016). Also, they need to explore how existing corpora function in terms of tagging, annotations, concordancing, etc.

Hu (2016) also notes that given the differences among students, particularly in terms of the extent to which they understand what is taught, the learning materials which vary in difficulty are extracted from the corpus and used for the analysis of translation strategies and methods by the students. Specifically, the students with high language proficiency can be assigned to analyse more complex statistics and translate more challenging texts, while those with lower language proficiency can be asked to extract and analyse texts comparatively easier to understand or investigate the translation of a single word or syntactic structure. In this way, translation teaching can be tailored to students' aptitude, and students' translation competence can thus be improved more effectively.

6. Conclusion

As a means of cross-cultural communication, translation serves as a bridge for people who speak different languages, enabling them to understand each other. Over the past decades, with the increase of global trade, cross-border immigration, globalisation, and the widespread application of the mass media, translation activities have been growing exponentially. As mediators in cross-cultural communication, translators play an increasingly important role. On the one hand, a translator has to cope with the transfer of avalanches of new information and new concepts across languages and cultures. On the other hand, a translator is often required to complete a translation task within a short period of time, during which a tiny error may cause grave consequences. Therefore, translation teaching or translator training is particularly important in the modern era when translation plays an increasingly important role. In the last decades, corpora have been increasingly used in establishing corpus-based mode of translation teaching.

Translator training can benefit a great deal from what corpora and corpus linguistics have to offer to it. Or as Bernandini et al. (2003) put it "the final goal is to make students better language professionals in an environment where computational facilities for processing texts have become the rule rather than the exception."

Finally, it is important to note that, in order to implement this novel corpus-based mode of translation teaching, as Hu (2016) rightfully remarks, the existing textbooks, pedagogy, and syllabus for the translation courses have to undergo an adequate revision, in view of enacting a shift of the students' role from a passive one to an active one. Moreover, one should not lose sight of the fact that since the compilation and use of corpora involve the use of software tools and statistical analysis, translation teachers should also be equipped with adequate ICT skills as well. It is therefore evident that the application of corpora in translation training promises to make translation teaching not only more objective and efficient but also much more independent and autonomous, thus, leading to the creation of well-versed future translators capable of facing the translation challenges of the modern era head-on.

References:

Baker, M. (1993). Corpus linguistics and translation studies – Implications and applications. In M. Baker, G. Francis & E. Tognini-Bonelli (Eds), *Text and technology*, *in honour of John Sinclair*, (pp. 233-252), Philadelphia and Amsterdam: John Benjamins.

Beeby, A., Rodríguez-Inés, P., & Sánchez-Gijón, P. (2009). Corpus use and translating: Corpus use for learning to translate and learning corpus use to translate. Amsterdam: Benjamins.

Bernardini, S., Stewart, D. & Zanettin, F. (2003). Corpora in translator education: An introduction. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 1-14). Beijing: Foreign Language Teaching and Research Press.

Cosme, C. (2006). Clause combining across languages: A corpus-based study of English-French translation shifts. *Languages in Contrast* 6(1):71-108, DOI: 10.1075/lic.6.1.04cos.

Hu, K. (2016). *Introducing corpus-based translation studies*. Berlin, Heidelberg: Shanghai Jiao Tong University Press, Shanhai and Springer-Verlag.

Kennedy, G. (1998). *An introduction to corpus linguistics*. London and New York: Longman.

Kruger, A., Wallmach, K. & Munday, J. (2011). *Corpus-based translation studies research and applications*. London and New York: Continuum International Publishing Group.

Laviosa, S. (2011). Corpus-based translation studies: Where does it come from? Where is it going? In A. Kruger, K. Wallmach & J. Monday (Eds.), *Corpus-based translation studies: research and application*. London and New York: Continuum Advances in Translation Studies.

Lüdeling, A., & Kytö, M. (2008). *Corpus linguistics,* Volume 1. Berlin: Walter de Gruyter GmbH & Co. KG.

Monzó, E. (2003). Corpus-based teaching: The use of original and translated texts in the training of legal translators. *Translation Journal* 7, 1-3.

Niladri, S., D., & Arulmozi, S. (2018). *History, features, and typology of language corpora*. Singapore: Springer Nature Singapore Pte Ltd.

Pearson, J. (2003). Using parallel texts in the translator training environment. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education*, (pp. 15-24). Manchester: St Jerome.

Zanettin, F. (1998). Bilingual comparable corpora and the training of translators. *Meta* 4, 616-630.