

Mikhail Mikhailov 2022. Text corpora, professional translators and translator training. *The Interpreter and Translator Trainer*, 16:2, 224-246.

### **Text corpora, professional translators and translator training**

Although machine translation software and CAT tools are commonly used both by professional translators and by those involved in the training of translators, the usefulness of electronic text corpora for these purposes is less widely known. Corpora of various types have become much easier to access during the last decade, and the main obstacle to their becoming a standard tool for translators is currently the inertia conservatism of both the industry and in universities. Translator training in universities can play an important role in promoting new working methods. To study to what extent corpora are present in university translator training programmes, a survey was carried out. The responses show that corpora are indeed becoming part of curricula, at least in EU countries. However, the role of corpora in these programmes is often peripheral. For example, compiling Do-It-Yourself corpora – a very important skill for translators – is still taught in only a few university programmes. For the most part, corpora are used mainly as a research instrument rather than as a tool in practical translation work.

Keywords: text corpora for translation, translator training, Do-It-Yourself corpora, LSP corpora, CAT tools.

## 1. Introduction

For a long time, text corpora have been primarily considered as a source of data for doing linguistic research (cf. Francis 1992, 17). Indeed, the first electronic corpora were developed to be used for frequency counts and for studying grammar (Ibid.). However, during the last decades, corpus-based approaches have been gaining ground in one area after another: grammar, semantics, lexicology and lexicography, contrastive studies, historical linguistics, dialectology, sociolinguistics and translation studies (Bowker and Pearson 2002, 11; Mikhailov and Cooper 2016, 1). At the same time, the more practical, non-research use of corpora has also been growing steadily.

Recently, the following major changes have taken place in the availability of corpora as an important resource:

- Numerous large corpora for many different languages are now available online. Open-source resources are becoming more and more common, and the fees for using commercial resources are becoming more moderate. It should be noted, however, that while the situation with monolingual corpora is fairly good, aligned parallel corpora are still far less common. The available corpus data is biased towards world languages (with English clearly dominating) and towards Language for General Purposes (LGP) rather than Language for Special Purposes (LSP) (see e.g. Frankenberg-Garcia 2015, 351–352; Mikhailov and Cooper 2016, 197–211).

- Some corpora are now collected from the web using data-mining technologies, and as a result it has become possible to compile corpora of over a billion running words (e.g. the corpora at SketchEngine, the BYU corpora, and Aranea).
- In addition to traditional dictionaries, there has appeared a new type of online dictionary service that uses archives of aligned parallel texts as their source, thus combining data from both dictionaries and parallel corpora (e.g. Linguee, GlosBe).
- Compiling and hosting one's own corpora has also become relatively easy, (e.g. by using the facilities offered by SketchEngine); and various tools for working with online corpora are available for non-commercial use (see Cerutti 2017).

These new developments make possible the large-scale use of corpora as a translating aid.

Usually, however, corpora are only mentioned in the literature in the context of research activities, whereas their use in practical translation work comes second, if mentioned at all (Bowker and Pearson 2002, 11). Presumably, it is the researchers who are supposed to compile the corpora they need. Later, they might make their corpora publicly available, and practitioners (editors, writers, teachers, translators, etc.) might also be given access to the data. Even so, demand was never expected to be high and little advice was provided to potential users (Bowker and Pearson 2002; Zanettin 2012).

It is also important to realise that a corpus is a much more complicated resource than a search engine or an online dictionary, and is difficult to master without special training and a background in linguistics (see e.g. Frankenberg-Garcia 2015, 354). As regards the use of corpora by translators, ready-made corpora can never satisfy all their needs. In contrast to other corpus users, translators often work with specialist texts in genres and on topics that are

poorly represented in large LGP corpora, and LSP corpora exist for only a limited number of text types and a limited number of languages.

As result, the role of corpora in translation activities is fairly modest. Several surveys performed in the U.K., Canada, France and elsewhere show that few translators regularly use corpora in their work, and many of them have never even heard of corpora (Kübler 2010, 63; Frankenberg-Garcia 2015, 353). The situation is better in Spain, but even there, around 40% of Spanish translators never use corpora (García-Izquierdo and Conde 2012; Gallego-Hernández 2015).

In this paper, I examine the role that corpora currently play in practical translation work from the point of view of researchers, translators and translator trainers; I shall then present the results of a small-scale survey; and finally, I shall discuss future prospects for corpora in the translation industry generally.

## **2. Corpora: types and modes of use**

First of all, it is necessary to distinguish corpora from other similar resources. Translators – and even some translator trainers – often confuse corpora with the World Wide Web and translation memories (TM). Of course, one can call any body of texts a corpus. However, such an approach is not very productive. It makes much more sense to treat the Internet and TMs as alternative types of language data, because their role in a translator’s work is different.

In this paper, the term ‘corpus’ will be used in a narrower sense. A text corpus is ‘a collection of naturally occurring language texts, chosen to characterize a state or variety of a language’ (Sinclair 1991: 171). Although a corpus can technically be an archive of raw texts, modern corpora are assumed to be either equipped with special tools for querying, or else

encoded and formatted so that they can be searched with existing standard corpus tools. Moreover, in most cases, corpora are expected to consist of well-formed texts that can be used as examples of how native speakers actually write or speak.<sup>1</sup> A corpus can be monolingual or multilingual. Multilingual corpora can be ‘parallel’ (i.e. versions of the same documents in different languages, or original texts vs. their translations in another language) or ‘comparable’ (i.e. collections of original texts in the same genre in two or more languages) (Mikhailov and Cooper 2016, 4–5).

The World Wide Web is not a specific collection of texts; it is a vast random miscellany of language data. The popular term ‘web as corpus’ does not actually refer to the WWW as a ready-made natural corpus, but as a source of data for the compilation of further, more specific corpora (see e.g. Baroni and Bernardini 2006). One cannot expect that all the texts found on the internet will be of high quality. Besides, except for the simplest searches, more sophisticated queries conducted with Google and other similar search engines take far too much time. Adam Kilgarriff (2007) has therefore criticized the use of Google in linguistic research, and to some extent this criticism can also be applied to the use of Google for checking technical terms, idioms or set phrases when translating.

Translation memories represent the language of a single writer or a group of writers, and they are commonly devoted to the same topics. In this respect TMs are indeed close to corpora, but the only TM functionality that is corpus-related is that of concordancing. TMs do not normally contain duplicate records which might create problems with calculating word frequencies. However, a translation memory (or, better still, aligned parallel texts from

1 Less typical are collections of texts produced by children, by language learners, and other specific groups, to study their language performance and to develop new methods of teaching, and the like.

previous TM projects) can be converted into a parallel corpus after the aligned texts have been lemmatized, morphologically and syntactically annotated, and prepared for use with a relevant corpus tool.<sup>2</sup>

In discussions about the uses of corpora in translation, it is often unclear what kinds of resources are being referred to: monolingual or multilingual corpora, general or specialized corpora, etc. In fact, corpora can be used at different stages of the translation process for different purposes. A tentative classification of the most typical applications of corpora in translating is shown in Table 1.

Table 1. Using corpora for translating.

Code	Object	Task	Type of Corpus			
			Monolingual LGP	Parallel LGP	Monolingual or comparable LSP	Parallel LSP
S1	Source text	Understanding the meaning & usage of lexical units	x		x	
S2	Source text	Understanding the meaning & usage of grammatical constructions	x		x	
S3	Source text	Understanding the meaning & usage of idioms and fixed phrases	x		x	
S4	Source text	Understanding the meaning & usage of special terms			x	
ST1	Source & target text	Finding & checking equivalents for lexical units	x	x	x	x
ST2	Source & target text	Finding & checking equivalents for grammatical constructions	x	x	x	x
ST3	Source & target text	Finding & checking equivalents for idioms and fixed phrases	x	x		
ST4	Source &	Finding & checking			x	x

2 In the recent years such tools have been made more available and easier to use, e.g. the users of Sketch Engine can create ad hoc parallel corpora and upload aligned texts. Sketch Engine performs lemmatization and morpho-syntactic analysis with integrated parsers.

Code	Object	Task	Type of Corpus			
			Monolingual LGP	Parallel LGP	Monolingual or comparable LSP	Parallel LSP
	target text	equivalents for special terms				
ST5	Source & target text	Finding & checking translations of quotations, equivalents for proper names and other culturally-bound items	x	x		
T1	Target text	Checking the use of lexical items	x		x	
T2	Target text	Checking grammar	x		x	
T3	Target text	Checking orthography and punctuation	x		x	
T4	Target text	Checking style & register	x		x	
T5	Target text	Checking genre and text type conventions			x	

The activities listed under S1–S4 and T1–T5 apply to any language service: copy-editing, proof-reading, writing (in one’s native language, but especially in a non-native language), and translating. The difference with other language services is that translators analyse texts in the source language (S1–S4 above) and edit texts in the target language (T1–T5), while other writers and editors do not usually switch languages. Non-translators are happy with monolingual corpora and in most cases rely on large general corpora. For instance, a copy editor would not need to compile his or her own corpus, but would simply check the use of rare words and expressions, specific constructions, etc. using precompiled corpora. In contrast, the needs of translators are more extensive. They often need corpora of both general and specialist texts, monolingual, comparable or parallel corpora, and even corpora they have compiled themselves. This is because the nature of a translator’s work

involves the production of texts in different genres and on varying – and often unfamiliar – topics.

### **3. Using corpora: from research to practical translation work**

#### ***3.1. Universities***

Back in the early 1990s, researchers working in translation studies were optimistic about the prospect of using text corpora both as a reference instrument for translating and as a research tool (Baker 1995 and 1999; Bowker and Pearson 2002; Gallego-Hernández 2012; Zanettin 2012). Four thematic conferences titled *Corpus Use and Learning to Translate* (CULT) were organized (Bertinoro in 1997 and 2000; Barcelona in 2004; Alicante in 2015) and numerous papers, books, and edited volumes were published on the topic (Bowker and Pearson 2002; Bernardini et al. 2003; Beeby et al. 2009, etc.).

Nevertheless, it is difficult to imagine that ready-made parallel or comparable corpora will ever be available for a wide variety of specialist fields and language pairs. In order to have access to such corpora, therefore, translators must first compile them themselves, either alone or in cooperation with other translators or researchers. Federico Zanettin (2002) calls such self-made collections of texts *Do-It-Yourself corpora* (DIY); Krista Varantola (2003) uses the term, *disposable corpora*, thus highlighting the temporary and non-research nature of such resources; and other researchers use the term *ad hoc corpora* (e.g. Corpas Pastor 2001).

Many researchers advocate DIY corpora in translator training (TT), with an emphasis on comparable LSP corpora (Corpas Pastor 2001; Corpas Pastor and Seghiri 2009; Sánchez-Gijón 2009; Gallego-Hernández 2012; Frankenberg-Garcia 2015; Veiga Díaz 2016;

Makowska 2016; Biel 2017). Natalie Kübler (2010, 67) stresses that the compilation of DIY corpora should be included in any translator training program.

### ***3.2. The Translation Industry***

The main problem when using corpora in translation in the 1990s was the insufficient IT proficiency of the translators themselves (see e.g. Varantola 2003, 66–67). Working with DIY corpora required many skills that translators still lacked at that time: scanning and OCR, knowing how to convert files into different formats, the merging and splitting of files, the use of advanced Find/Replace techniques, knowing how to store and archive data, etc. Another challenge was the accessibility of technical equipment: computers, scanners, and the relevant software were not always available. Later, in the 2000s, the situation began to improve in many countries. However, trends in the translation industry did not change at the same pace. This time the problem was not in the availability of resources, but in the inertia of professional translators who had been trained in a different way and were used to working with other tools in a completely different environment.

A field study carried out at the beginning of the 2000's by the Finnish researchers Anna Mauranen and Riitta Jääskeläinen (2005) in a large company in Finland revealed the extent of the problem: the company's translators were simply not interested in using corpora in their work, and they were not willing to spend time compiling text archives or to learn to use new software.

Most people would agree that it is hard to persuade a professional to start using new tools. As Maeve Olohan (2004, 176) points out:

One of the important points often made by translation professionals when confronted with the range of electronic resources available is that they see the potential usefulness of the data and tools but are unlikely to have the time, first, to acquaint themselves with the software etc., and second, to focus in such depth on specific aspects of their translation tasks .

For this reason, although the major technical obstacles in using corpora in a translator's everyday work have been removed, corpora are still peripheral among the technical aids used by translators, being clearly outmatched by translation memories. TMs are included in many translator training programmes; they are known in the industry, and translators are strongly encouraged – and sometimes even forced – to use them (cf. LeBlanc 2013, Frankenberg-Garcia 2015, 354-355). More recently, however, the introduction of corpora in translator training at different universities has gone well and students have shown considerable interest (see e.g. Corpas Pastor 2001; Corpas Pastor and Seghiri 2009; Sánchez-Gijón 2009; Rodríguez Inés 2008 and 2009; Kübler 2003 and 2010; Frankenberg-Garcia 2015; Makowska 2016; Veiga Díaz 2016; Biel 2017).

### ***3.3. Translator trainers***

Translator trainers started to be aware of the importance of IT skills in the work of translators by the end of the first decade of this century. In 2009, the European Master's in Translation network (EMT), and one year later the PACTE group, defined the essential competences of a translator. Along with knowledge of (at least two) languages, together with thematic, intercultural and technological competences, the list included a competence in information mining: 'Knowing how to use tools and search engines effectively (e.g. terminology software, electronic corpora, electronic dictionaries)' (EMT 2009; PACTE 2010). In a new framework

of translator competences, drawn up in 2017, the EMT also mentions corpora (EMT 2017).

To become an everyday routine, any new working method requires time. People in the field must get used to it, and the workers must become convinced that it is more efficient than traditional methods. The relevant equipment should also be provided and training should be organized in universities. It is natural, of course, that the current situation should be dictated to a large extent by the older generations who were trained to translate without corpora, without CAT tools, and often even without a computer. This explains the negative results found in surveys on the use of corpora that were mentioned above. But the situation will change if the use of corpora is actively promoted in university curricula. If this happens, in the next ten years the translators might supplement search engines such as Google with corpora, which are more reliable and more suitable for language queries .

Even so, it will be an uphill task, because a large percentage of practising translators do not have degrees in translation. They have often studied foreign languages, literature, journalism, and sometimes have no university degree at all.<sup>3</sup>

#### **4. Corpora in the classroom: a survey of the use of corpora in translator training**

To be able to forecast the developments in the translation industry during the next decade, we must first establish the extent to which corpora are integrated into university programmes in translator training today.

Exploring university curricula online does not yield many results. Quite often the web pages of M.A. programmes (if they have any web pages at all) have only a short synopsis of

3 This issue is widely discussed among translator trainers and can be observed indirectly in many surveys, e.g. Mikhailov 2015: 96.

their curricula and courses, without much detail. It is therefore difficult to know whether the use of corpora is included in practical courses or not, unless a separate course on corpus linguistics is mentioned in the course descriptions. For this reason, doing a survey was the only way to discover the real state of affairs.

In 2017, Andrew Rothwell and Tomáš Svoboda (2019) carried out a survey on the use of computer technologies in translator training, and the responses obtained demonstrated that the integration of IT in Translator Training (TT) curricula is gradually increasing. Moreover, although TM systems clearly dominated, many respondents also mentioned the importance of corpora (p. 40). My own survey focused on corpora in particular, the aim being to provide more detail on their role in university curricula.

#### ***4.1. The structure of the questionnaire***

To conduct the survey, I developed an online questionnaire that was placed on the server of Tampere University. The respondents' answers were later downloaded in table form for further analysis.

The questionnaire<sup>4</sup> began with an introduction in which I explained the purpose of the survey, the kind of respondents I needed, and the way in which the data was to be handled.

When developing the questionnaire, I took pains to make it as short as possible, so that it should not take more than 10 minutes to complete. Altogether, the questionnaire consisted of 21 questions covering the following topics:

4 The questionnaire and the data set are available online at [https://puolukka.rd.tuni.fi/corpora\\_and\\_translators/](https://puolukka.rd.tuni.fi/corpora_and_translators/)

- general information on the university programme: country, city, size of the programme, working languages, specializations;
- place and role of text corpora in the curriculum;
- corpus functionalities being used in the teaching;
- possible future plans for expanding the use of corpora in the programme.

The language of the questionnaire was English.

The survey began in September 2018 and continued until April 2020. Altogether, respondents from 91 programmes in 38 countries took part, with each respondent representing a different M.A. programme.

Of these 91 initial respondents, 72 said that they included corpus studies in their curricula.<sup>5</sup> This does not necessarily mean, however, that corpora are popular; it might rather mean that the programmes that do not use corpora were less interested in taking part in the survey. Because it would have been difficult to speculate on the reasons for not teaching the use of corpora, the subsequent analysis will be based on the responses of the 72 programmes that do offer corpora in their curricula.

The survey was conducted over quite a long period of time; however, no significant differences between early and late answers were detected.

5 In fact, 23 respondents of the 91 said that they did not have corpus studies in their programmes. However, after checking their answers to other questions, it became clear that four informants obviously either misunderstood the question or pressed the wrong button, because their answers clearly indicate that they do have corpus studies integrated in their curricula.

#### ***4.2. Geographical factors***

The invitation to take part in the survey was sent to university teachers all around the world via different channels: the EMT network (European Master's in Translation), relevant Facebook groups, personal contacts, and finally, directly to university programmes and their administrative and teaching staff. In a number of countries, I did not find any translator training programmes at all, e.g. Laos, Libya, Myanmar, Vietnam. However, even the existence of multiple translator training programmes in a given country did not guarantee that I would obtain any responses. Furthermore, many programmes did not respond despite receiving reminders. A large part of the world therefore is still 'unknown territory': there were no responses from South and Central America, Central Asia, India, Japan or Australia; there were very few responses from North America and Africa, only one answer from Malaysia, and only one answer from New Zealand (see Figure 1).

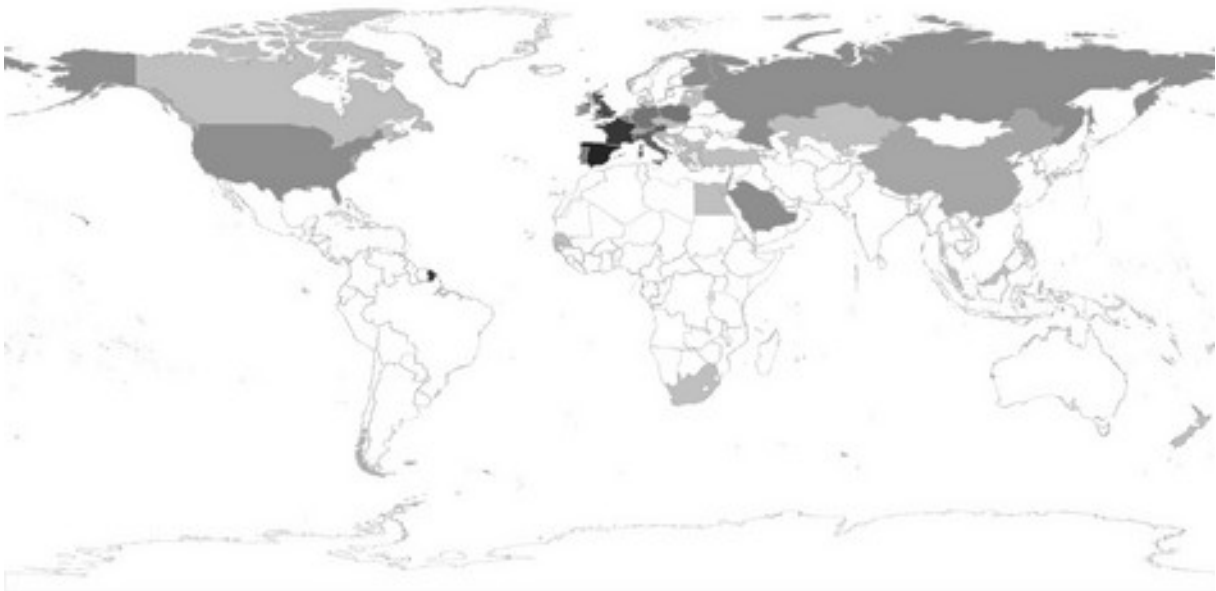


Figure 1. Responses to the online questionnaire, worldwide

Almost 80% of the answers came from universities in Western and Eastern Europe (see Figure 2 and Table 2). The most active respondents to the questionnaire were from Spain, France, Italy, and U.K. Many countries, like Estonia, Serbia and Slovakia, were represented by only one respondent, but this is because smaller countries often have only one or two translator training programmes. Only three answers came from Russia, even though an invitation (in Russian) was sent to more than 30 email addresses. No answers came from Denmark, Norway, Sweden, Hungary, Ukraine or Belarus.



Figure 2. Responses to the online questionnaire, Europe

Within Europe, the smaller countries might have only one program, while larger countries might have more than twenty. The EMT network had 81 programme representatives in 2019–2024 (EMT 2021), but at that time there were many translator training programmes

that had not been admitted to the EMT or had not applied. It would be difficult therefore to speculate even as to the approximate number of university programmes in Europe and an even greater challenge to account for the figures in the table.

Table 2. Number of responses to the questionnaire (Europe)

Country	Number of responses
Spain	9
France	7
Italy	5
United Kingdom	5
Belgium	4
Germany	4
Poland	4
Austria	3
Finland	3
Ireland	3
Russia	3
Switzerland	2
Bulgaria	1
Croatia	1
Czech Rep.	1
Greece	1
Latvia	1
Lithuania	1
Netherlands	1
Portugal	1
Serbia	1
Slovakia	1
Slovenia	1

Whatever the reasons, the overall response rate was quite low. This was possibly due to the passivity of the recipients and a lack of interest in the subject. Programmes that do not

use corpora rarely answered the questionnaire, and there are also reasons to believe that many potential respondents did not even open the invitation.

Another possible problem was that the language of the questionnaire was English. Nowadays of course English is the lingua franca of science and of university education generally, but in many of the countries of Asia, Africa and America there may be difficulties in communicating in English even among academics. Even so, conducting the survey in three or four languages would have only increased the imbalances in the data, thus making a large-scale multilingual survey virtually impossible.

Finally, email and social networks do not function equally well in all parts of the world. In some countries regular mail and telephone remain the main means of communication, and email is checked only occasionally, if at all.

To sum up, I cannot claim that the responses to my questionnaire are truly representative of translator training programmes generally. The results are Europe-centred with a bias towards Western Europe. However, translator training and corpus studies are most popular in this region and the data collected serves the objective of the survey by showing the general trends.

#### ***4.3. The respondents***

The answers in the general information section of the survey demonstrated once again that there exist many different ways of organising TT: as a small specialization inside a language department; as a separate M.A. programme with three or four language pairs; or as a large autonomous institution with more than a hundred teachers and more than a thousand students.

Most of large and very large programmes are in Austria, Germany, Italy and Spain. But in some cases it is difficult to decide whether a particular programme is large or small. In such cases, there might be just two or three permanent staff members doing lecture courses and thesis supervision, and many part-time instructors teaching language, translation and interpreting courses.

Most of the TT programmes are M.A. programmes, which means that the duration of teaching is 2–3 years with the training centred on translating and interpreting rather than on teaching language skills. Nevertheless, there also exist a certain number of B.A. programmes (e.g. in Spain and Italy) and even B.A. + M.A. programmes (e.g. in Germany and Ireland). In some cases, it is difficult to decide whether the programme in question is at the M.A. or B.A. level, and some programmes are basically programmes in foreign languages with only a supplementary course in translation. The level of the programme and its duration obviously influence the way in which corpora will be incorporated, but the answers of the respondents did not provide enough information to determine this.

A TT programme can train translators, interpreters and specialists in translation technologies and localization. Among the 72 programmes that use corpora in their curricula, 70 programmes specifically train translators, with only two not offering translator training: one of these is a programme for training interpreters and the other a programme in translation technologies. There are 23 programmes providing translator training only, 36 programmes offering translator training and translation technologies, and 14 programmes offering all three modules: translator training, interpreter training, and translation technologies. This means that only half of the programmes in the survey have a separate module in translation technologies, which to a certain extent might indicate that the traditional methods of training translators are

still common in universities (input: paper – output: Word document; or input: Word document – output: Word document).

The number of language combinations taught varies from one to fifteen. The most popular languages are English, French, German, Spanish, Italian, Russian, Chinese and Japanese. Usually the combinations are with the official language(s) of the country, e.g. in Germany it is German ↔ other languages; in Belgium it is Dutch and French ↔ other languages.

#### ***4.4. Research or practical use?***

The purpose of corpus studies in university curricula can vary. Some programmes teach theoretical considerations, but do not use corpora in the actual teaching of translation. To check the rate of integration of corpus studies in the teaching process, the respondents were asked to specify the different ways in which they use corpora in their programmes. The respondents were offered various options to choose from (Table 3).

It is not very likely that any programme uses corpora extensively, because many TT programmes are short M.A. courses with a high level of specialization. Few of them have room for many theoretical courses, courses in linguistics or practical language courses. If the topic is only introduced briefly in a lecture course, it is unlikely to be followed by any practical applications.

Table 3. Integrating corpora in TT programmes

<b>Place in curriculum</b>	<b>Number of responses</b>
Corpora in M.A. thesis	48
Integrated into translation courses	42
Integrated into Translation Studies	37

<b>Place in curriculum</b>	<b>Number of responses</b>
Integrated into Language Technologies	37
Integrated into terminology courses	34
Separate course in corpus studies	17
Integrated into language courses	11
Integrated into linguistics course	9

As was mentioned above, corpora can be used either as a source of data for research or as a practical tool. In Table 3 the eight different ways of using corpora are listed together with the number of programmes incorporating them in their curricula. In most cases corpora are present as a research instrument for M.A. theses (48); often they are briefly introduced in theoretical courses on translation studies (37) or language technologies (37). As a rule, the programmes reported on here do not have separate courses on corpora (because there is no room for such courses in the curriculum), and they seldom use corpora in the teaching of foreign languages (which is likely to take place at B.A. level). More than half of the programmes use corpora in translation courses (42) or in courses in terminology (34).

The answers to the question about the purpose of using corpora in these programmes show that corpora are seen as being equally important in research, translation and terminological work (see Table 4). However, the figures in Table 4 do not correspond directly with those in Table 3: for example, the use of corpora to investigate special terms may often accompany a theoretical course in terminology. Probably, though, corpus-based terminological studies will more often be integrated into practical courses in translation or language technology than into theoretical courses on terminology. Some programmes, of course, may have no course on terminology in their curricula.

Table 4. Use of corpora in the teaching process

<b>Purpose of using corpora</b>	<b>Number of responses</b>
For translating	59
For terminology work	55
As a research instrument	54
For language learning	19

#### ***4.5. Types of corpus data used in translator training***

According to the responses in the survey, almost all of the 72 programmes that have corpus studies in their curricula use monolingual and/or parallel corpora (Table 5). Comparable corpora are less common (48). Only 28 programmes use translation corpora (translations vs. non-translations); these are used for research rather than in translating.

Table 5. Types of corpora used in the teaching process

<b>Type of corpora</b>	<b>Number of responses</b>
Monolingual corpora	63
Parallel corpora	62
Comparable corpora	48
Translation corpora	28

In the previous sections, I pointed out the usefulness of Do-It-Yourself (DIY) corpora for translating specialist texts. The responses in the survey show that compiling one's own corpus is less common than using ready-made text collections. Altogether, 33 programmes provide a brief introduction to compiling DIY corpora. In 40 programmes students compile their own corpora in practical courses. As for the different activities that involve DIY corpora, many programmes cover only collection and alignment of parallel texts (55 and 45

programmes respectively). Other activities, like scanning and OCR, parsing, lemmatization, etc, are taught only occasionally (Table 6).

Table 6. DIY corpora in the teaching process

Activities	Number of responses
Collecting texts	55
Aligning parallel texts	45
Parsing	26
Lemmatization	22
Metadata	20
Scanning and OCR	15
Copyright issues	8

#### *4.6. Querying corpora*

The results of the survey revealed that concordancing is still the commonest type of corpus query, but other types of query are also quite actively used in many courses (Table 7).

Table 7. Types of corpus queries covered in the programmes

Corpus query type	Number of responses
Concordances	65
Collocations	61
Frequency lists	58
Keywords	50
Ngrams	31

Just a decade ago corpus queries other than concordancing were only available to those using DIY corpora via software like WordSmith Tools, while users of online corpora were restricted almost exclusively to concordancing. Our survey showed that one positive

development resulting from the popularity of SketchEngine is that users can now run different kinds of corpus queries both with ready-made corpora and with their own data. Collocations and frequency lists have become almost as popular as concordances, and indeed they are very useful in situations where the translator wants a large range of alternatives.

#### ***4.7. Corpus software used***

Among the software applications used by the survey respondents, old-fashioned desktop programs like WordSmith Tools or AntConc are still popular. These are easy to install and easy to use, and are commonly employed to query DIY corpora collected by translators themselves. The main weakness of these products is that they are designed to work with raw text and do not support lemmatized and grammatically annotated corpora (except with the help of rather complicated technical modifications). Sharing data might also be difficult.

For this reason, in recent years SketchEngine has begun to attract many users: it is web-based, it has built-in parsers for many languages, and it has numerous querying functionalities, including concordances, collocations, frequency lists, keywords, Ngrams, and the word sketch and thesaurus facilities. SketchEngine users can work with ready-made corpora or upload their own collections of monolingual or aligned parallel texts, or even translation memories that can be used as a kind of parallel corpora. It is also possible to collect corpora from the web with the help of a built-in crawling WebBootCat application. At present, other web-based corpus portals (OPUS, English.corpora.org, etc) can only be used for querying ready-made corpora. Although some of these software packages can be installed on the user's server in order to work with one's own data, this is quite complicated and an extensive background in IT is needed.

Not surprisingly therefore, the most popular software among our respondents was SketchEngine, which is mentioned 35 times. Next came AntConc (27) and WordSmith Tools (14). Occasionally other software is used, e.g. IntelliText, TextStat, TXM and CQPweb. Additional tools that are used include aligning software like LF Aligner, taggers like TreeTagger, and terminology management tools like TermoStat or Logiterm.

At some universities getting familiar with corpus tools is considered to be extremely important. One of the respondents remarked: ‘We start with very simple ones like AntConc. We do more specific and specialized operations with Hyperbase, TXM or Iramuteq. They also need <to know> how to write in the xml environment and how the TreeTagger works.’ In contrast, at other universities, students use only ready-made online corpora.

#### ***4.8. Overall trends***

To check if the survey revealed any general trends, I next grouped the questions asked into five main topics: Querying corpora, Language checking, Translating, Compiling corpora, and Research.

For example, for the topic ‘Querying corpora’, I used the answers given to the following question: ‘What types of corpus queries are introduced in your programme?’ The question had five options: frequency lists, concordances, collocations, Ngrams, keywords.

For each programme, scores were calculated per topic (1 point for each activity), and histograms created for the purpose of visualization (Figure 3). For example, if a particular programme only covered concordancing and frequency lists, this received 2 points. The scores for other activities were then calculated in the same manner, by counting the number of

positive answers. The statistics were based on the answers received from the 72 programmes that do use corpora.

**Querying corpora** is a basic skill that is needed for all kinds of corpus-based activity. Indeed, the majority of the respondents considered querying to be highly important. We can see in Figure 3 that most programmes that took part in the survey use three or more kinds of search routines in their courses. Strangely, 7 programmes do not include any corpus querying at all in their curricula. No explanation could be found for this, suggesting that the students are probably expected to be able to have mastered these basic skills without the teacher’s help.

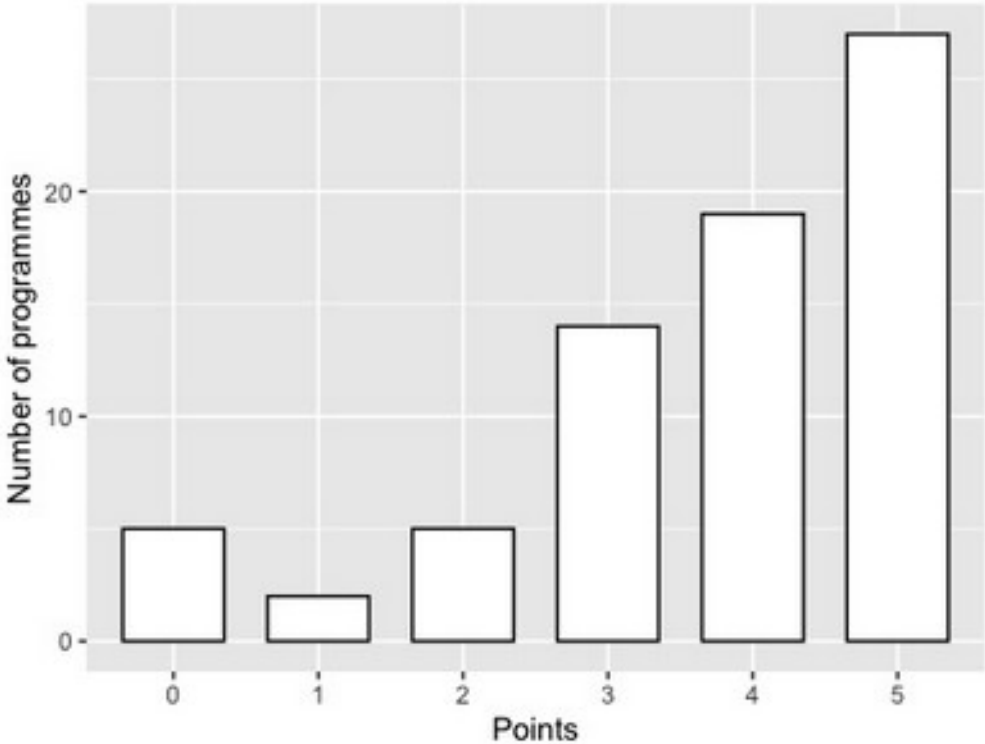


Figure 3. Querying corpora

**Checking the language.** This corpus functionality does not seem to have the attention it deserves in the training of translators. It is clear from Figure 4, that almost half of the programmes only use corpora to check grammar and vocabulary at a rather minimal level. The probable reason is that most of the programmes in the survey were M.A. programmes with few or no courses in language and linguistics.

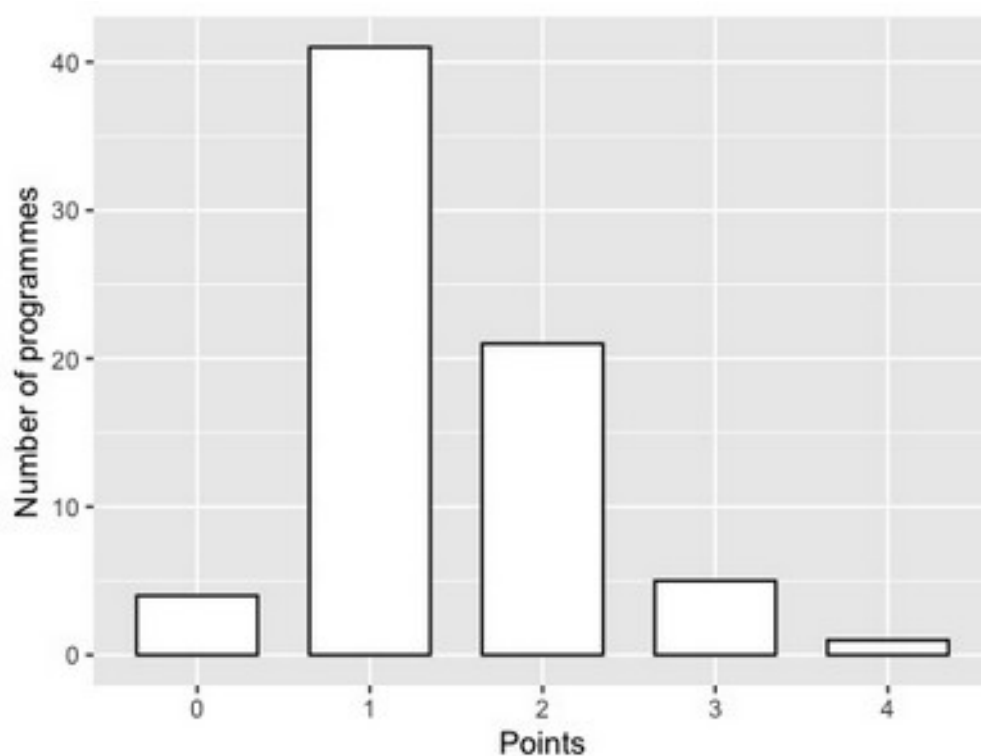


Figure 4. Using corpora for checking language

**Translating.** One can assume that the main purpose of using corpora in translator training would be as a tool to facilitate translation and improve its quality. And there were no surprises here: the overwhelming majority of respondents mention four or more activities connected with translating (see Figure 5).

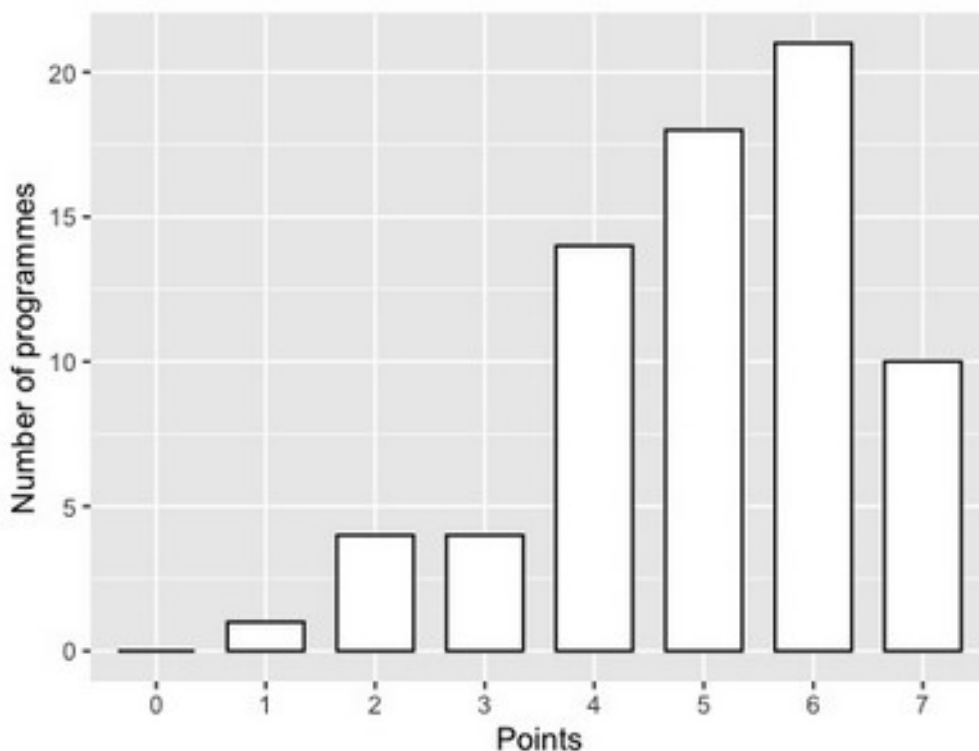


Figure 5. Using corpora in translation courses

**Compiling corpora.** The importance of compiling corpora in translating has been mentioned several times already. Nevertheless, the results of the survey show that the process of integrating DIY-corpora into translator training is still in its infancy. It can be seen from Figure 6 that a large number of programmes do not teach the compiling of DIY corpora at all (0 points). Alternatively, they sometimes include a brief introduction in a theoretical course (1 point) and text collecting in practical courses (2 points). The number of programmes that pay a good deal of attention to compiling corpora (5–7 points) accounts for only about 10% of the responses, and the mean (3–4 points) does not exceed 20%.

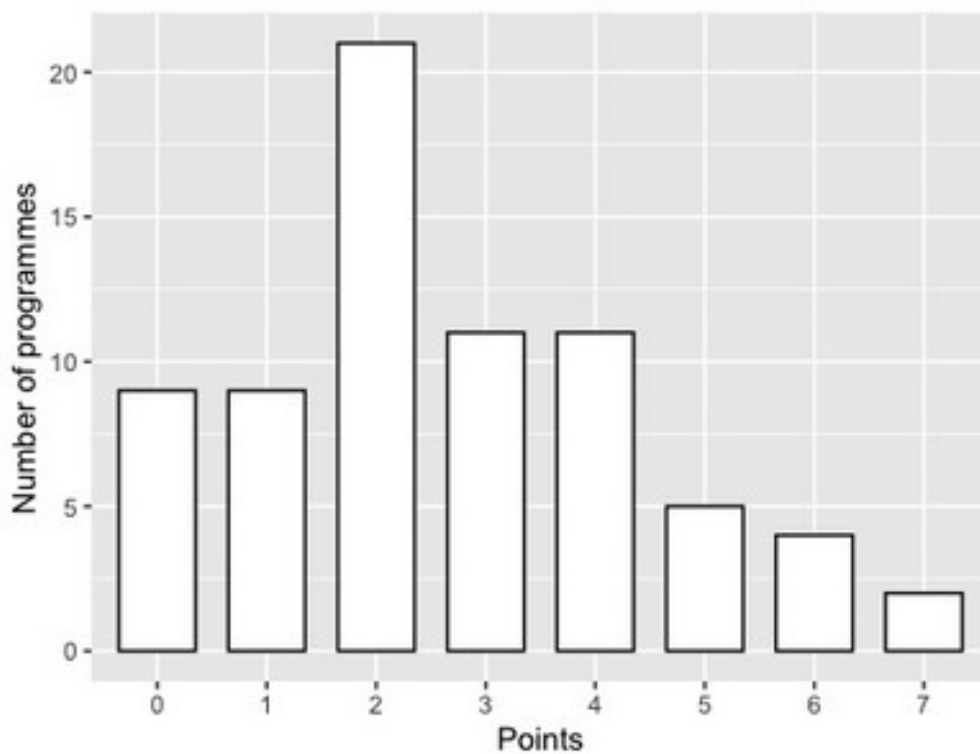


Figure 6. Compiling corpora in the curricula

**Research.** Last but not least, the theoretical background on corpora and their use as a research instrument is considered to be quite important. The overall tendency in the histogram is towards 4–5 points, which indicates that in most programmes the use of corpora for research purposes is integrated in different modules (Figure 7).

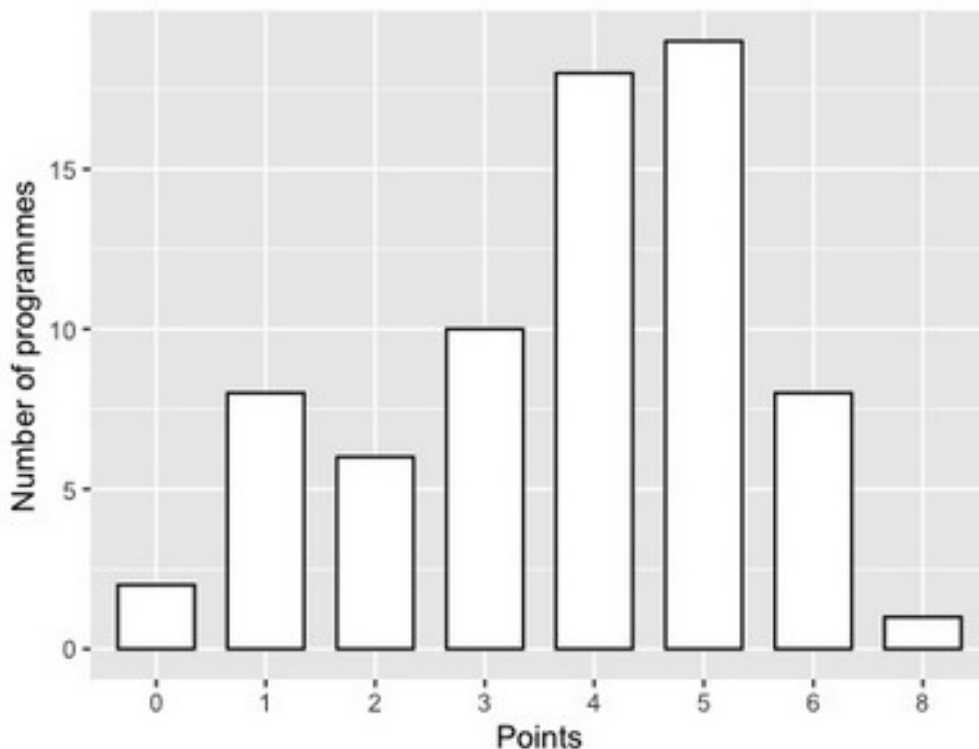


Figure 7. Using corpora as research instrument

#### ***4.9 Overall trends summarized***

An analysis of all the responses using a conditional inference tree (R, *party* package, *ctree* function) reveals that the use of corpora for research is indeed the most important factor in university curricula. The analysis checked the connections between the total number of teaching activities that include corpora (0 to 8 points) and the scores for each of the above-mentioned groups of activities (research, translating, compiling corpora, etc). The results can be seen in Figure 8.

The main parameter that emerged was use in research: the programmes that use corpora for research have highest scores. The use of corpora in practical translation courses comes second. The highest scores are found with the 29 programmes which use corpora for

research and which also get 5 or more points for the use of corpora in translating. The remaining activities (compiling and querying corpora, language checking) are not visible at all.

The results of this test make me feel somewhat pessimistic: it seems that corpora are currently seen mainly as a research instrument, or simply as a tool for looking up translation equivalents, rather than as a ubiquitous tool for multiple translation activities (see Table 1). This is also seen in the reluctance to compile DIY corpora (see Figure 5).

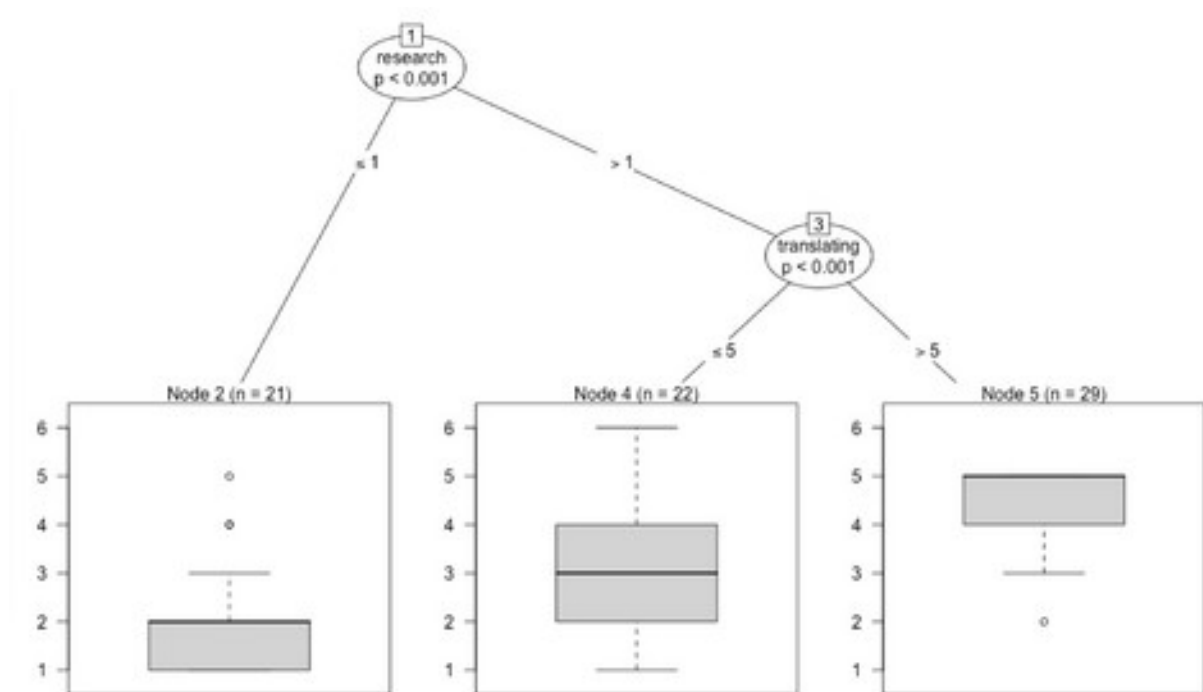


Figure 8. The purposes of using corpora in the curricula vs. overall score of using corpora.

#### 4.9. Comments of respondents

It became clear from the comments of the respondents that corpora have been introduced in some programmes only recently; in general, however, the resource is considered to be useful and worth expanding.

‘Corpora as a language / translation resource is introduced briefly in a couple of lecture courses, but the students are not able to learn these in detail (e.g. through hands-on)’

‘Last year our faculty offered a small course on the use of Sketch Engine. We are looking into the possibilities of including even more corpora work and training.’

Some respondents point out that practice varies even within the programme: some teachers are eager to use corpora and integrate them in the courses they teach, while others are less enthusiastic. In addition, larger programmes may consist of separate sections for different language pairs or different specialisations, some of which are using corpora while others are not.

‘The use of corpora in other modules besides the Corpus Linguistics one very much depends on individual lecturers, so it is impossible to achieve consistency across modules. It is encouraging nevertheless to see a few members of staff using the technology in Specialised Translation classes, as well as in the Writing for Special Purposes module.’

Some respondents mention that the use of corpora in their programmes is skewed towards certain activities:

‘We use corpora almost exclusively for terminological purposes. We have an optional module in Terminology Management, but all students are exposed to Sketch Engine and encouraged to use it as part of their compulsory group CAT tools project, to discover and compile terminology.’

Some programmes would be interested in expanding the corpus studies element, but they are not able to do so for lack of funding or for administrative reasons. Other programmes are about to introduce new curricula with more courses on corpora or are in the process of developing such courses.

Respondents often mention that their students are enthusiastic about compiling and using corpora:

‘We had initial reservations about the students’ willingness to build and explore corpora. Students, however, turned out to have a very positive attitude towards using corpora.’

## **5. Conclusions**

After surveying the state of the art both in the translation industry and in the technologies associated with translation, it has become obvious that corpora deserve much more attention than they currently attract. The more active use of corpora in a translator’s work in addition to the use of CAT tools would certainly improve the quality of translation and add creativity to the translation process (see e.g. Frankenberg-Garcia 2015; Kübler 2010). The use of CAT tools alone makes translating more mechanical (see e.g. LeBlanc 2013).

Change is unlikely to happen by itself, however, because the translators tends to favour traditional methods. Obviously, therefore, it is translator training that should play a central role in initiating change. It should actively introduce new technologies to further facilitate the translation process and help students to develop the relevant IT skills. It is now widely accepted that the training of translators can no longer be based solely on language teaching. The curricula are being complemented by courses supporting other skills that are important for the translators of the present day . The EMT, for example, recommends the following areas of competence: language and culture, translation skills, technology, personal and interpersonal skills, and service provision (EMT 2017). All of these are interrelated: e.g. skills in using computer technologies (electronic dictionaries, language utilities in office tools), help to improve language skills; they are also extremely important in the translation

process (TM, MT, corpora); and they play a role in developing personal and interpersonal skills (in social media) and in service provision (CAT tools workflow applications). It is obvious, that in teaching IT-skills, the variety of technology and information mining tools should be introduced. Using corpora is one of the skills that are not taught enough.

However, it is clear that changes in university curricula cannot happen overnight; they take place gradually, especially when it comes to improving technological competence. The challenges are obvious: to develop these skills, training programmes need the relevant technical equipment (computer labs, software); staff capable of providing instruction in the use of different applications; and the integration of these applications in translation courses. At the moment, corpora are more often used as a research instrument rather than as a tool for practical translation work; however, the popularity of SketchEngine may boost the non-academic use of corpora of all kinds: monolingual, parallel, ready-made and Do-It-Yourself corpora.

Currently, according to the results of the present study, corpora are being integrated in translator training mostly in the EU countries. (Universities in America, Africa, Asia, and Oceania did not take a very active part.) The survey also demonstrated that the pace of development varies from country to country and from university to university, and that we are clearly only at the beginning of a very long journey.

## References

- Austermühl, Frank. 2013. "Future (and not-so-future) trends in the teaching of translation technology." *Revista Tradumàtica*, 11: 326-337.  
<https://revistes.uab.cat/tradumatica/article/view/n11-austermuehl/pdf>

- Baker, Mona. 1995. "Corpora in translation studies: an overview and some suggestions for future research." *Target*, 7(2): 223-243.
- Baker, Mona. 1999. "The role of corpora in investigating the linguistic behaviour of professional translators." *International Journal of Corpus Linguistics*, 4(2): 281-298.
- Baroni, Marco, and Bernardini, Silvia, eds. 2006. *Wacky! Working papers on the web as corpus*. Bologna: Gedit.
- Bernardini, Silvia, Dominic Stuart, and Federico Zanettin, eds. 2003. *Corpora in Translator Education*. Manchester: St. Jerome.
- Beeby, Allison, Patricia Rodríguez Inés and Pilar Sánchez Gijón, eds. 2009. *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Amsterdam: John Benjamins.
- Biel, Łucja. 2017. "Enhancing the communicative dimension of legal translation: comparable corpora in the research-informed classroom." *The Interpreter and Translator Trainer*, 11(4): 316-336. DOI: 10.1080/1750399X.2017.1359761
- Bowker, Lynne, and Jennifer Pearson. 2002. *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.
- Cerutti, Giorgina. 2017. "Evaluating tools for legal translation research needs. The case of fourth-generation concordancers." In *Legal Translation and Court Interpreting: Ethical Values, Quality, Competence Training*, edited by Annikki Liimatainen, Arja Nurmi, Marja Kivilehto, Leena Salmi, Anu Viljanmaa, and Melissa Wallace, 355-391. Berlin: Frank & Timme.
- Corpas Pastor, Gloria. 2001. "Compilación de un corpus ad hoc para la enseñanza de la traducción inversa especializada." *Trans*, 5: 155-184.
- Corpas Pastor, Gloria, and Miriam Seghiri. 2009. "Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish)." In *Corpus use and Translating: Corpus use for Learning to Translate and Learning Corpus Use to Translate*, edited by Allison Beeby, Patricia Rodríguez Inés, and Pilar Sánchez-Gijón, 76-107. Amsterdam: John Benjamins.
- EMT 2009 = Gambier, Yves *et al.* 2009. *Competences for professional translators, experts in multilingual and multimedia communication*. Brussels, European Commission.

- EMT 2017 = Toudic, Daniel *et al.* 2017. *EMT Competence Framework 2017*. Brussels, European Commission.
- EMT 2021. List of EMT members 2019-2024.  
[https://ec.europa.eu/info/resources-partners/european-masters-translation-emt/list-emt-members-2019-2024\\_en](https://ec.europa.eu/info/resources-partners/european-masters-translation-emt/list-emt-members-2019-2024_en)
- Francis, William. 1992. "Language Corpora B.C." In *Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82, Stockholm 4-8 August 1991*, edited by Jan Svartvik, 17-35. Berlin: Mouton de Gruyter.
- Frankenberg-Garcia, Ana. 2015. "Training translators to use corpora hands-on: challenges and reactions by a group of 13 students at a UK university." *Corpora*, 210 (3): 351-380. DOI: 10.3366/cor.2015.0081
- Gallego-Hernández, Daniel. 2012. *Traducción económica y corpus: del concepto a la concordancia. Aplicación al francés y al español*. Alicante: Universidad de Alicante.
- Gallego-Hernández, Daniel. 2015. "The use of corpora as translation resources: A study based on a survey of Spanish professional translators." *Perspectives*, 23(3): 375-391, DOI: 10.1080/0907676X.2014.964269
- García-Izquierdo, Isabel, and Tomás Conde. 2012. "Investigating specialized translators: Corpus and documentary sources." *Ibérica*, 23: 131-156.  
[www.aelfe.org/documents/07\\_23\\_Garcia.pdf](http://www.aelfe.org/documents/07_23_Garcia.pdf)
- Kilgarriff, Adam. 2007. "Googleology is bad science." *Computational Linguistics*, 33(1): 147-151. [www.kilgarriff.co.uk/Publications/2007-K-CL-Googleology.pdf](http://www.kilgarriff.co.uk/Publications/2007-K-CL-Googleology.pdf).
- Kübler, Natalie. 2003. "Corpora and LSP translation." In *Corpora in Translator Education*, edited by Silvia Bernardini, Dominic Stuart, and Federico Zanettin, 25-42. Manchester: St. Jerome.
- Kübler, Natalie. 2010. "Working with corpora for translation teaching in a French-speaking setting." In *New Trends in Corpora and Language Learning*. edited by Guy Aston, Lynne Flowerdew, and Ana Frankenberg-Garcia, 62-80. New York, N.Y.: Continuum.
- LeBlanc, Matthieu. 2013. "Translators on translation memory (TM). Results of an ethnographic study in three translation services and agencies." *Translation and Interpreting*, 5(2): 1-13. <http://www.trans-int.org/index.php/transint/article/view/228>.

- Maia, Belinda. 2003. "Some languages are more equal than others. Training Translators in Terminology and Information Retrieval using Comparable and Parallel Corpora." In *Corpora in Translator Education*, edited by Silvia Bernardini, Dominic Stuart, and Federico Zanettin, 43-53. Manchester: St. Jerome.
- Makowska, Aleksandra. 2016. "Creating multilingual corpora to teach scientific translation." In *New insights into Corpora and Translation*, edited by Daniel Gallego-Hernández, 57-78. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Mauranen, Anna and Riitta Jääskeläinen. 2005. "Translators at work: a case study of electronic tools used by translators in industry." In *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*, edited by Geoff Barnbrook, Pernilla Danielsson, and Michaels Mahlberg, 48-53. London: Continuum.
- Mikhailov, Mikhail. 2015. Minor language, major challenges: the results of a survey into the IT competences of Finnish translators. *The Journal of Specialised Translation*, 24: 89-111. [http://www.jostrans.org/issue24/art\\_mikhailov.pdf](http://www.jostrans.org/issue24/art_mikhailov.pdf)
- Mikhailov, Mikhail, and Robert Cooper. 2016. *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research*. London: Routledge.
- Olohan, Maeve. 2004. *Introducing Corpora in Translation Studies*. London: Routledge.
- PACTE 2010 = Hurtado Albir *et al.* 2010. "Results of the Validation of the PACTE Translation Competence Model: Translation Project and Dynamic Translation Index." In *IATIS Yearbook 2010*, edited by Sharon O'Brien. London: Continuum.
- Rodríguez-Inés, Patricia. 2008. *Uso de corpus electrónicos en la formación de traductores (inglés-español-inglés)*, Unpublished PhD thesis, Barcelona: Departament de Traducció i d'Interpretació, Universitat Autònoma de Barcelona.  
<https://ddd.uab.cat/record/129868?ln=en>
- Rodríguez Inés, Patricia. 2009. "Evaluating the process and not just the product when using corpora in translator education." In *Corpus use and Translating: Corpus use for Learning to Translate and Learning Corpus Use to Translate*, edited by Allison Beeby, Patricia Rodríguez Inés, and Pilar Sánchez-Gijón, 129-149. Amsterdam: John Benjamins.

- Rodríguez-Inés, Patricia. 2010. "Electronic corpora and other ICT (Information and communication technologies) tools: an integrated approach to translation teaching". *The Interpreter and Translator Trainer*, 4(2): 251-282.  
<https://www.tandfonline.com/doi/abs/10.1080/13556509.2010.10798806>
- Rothwell, Andrew and Tomáš Svoboda 2019. "Tracking translator training in tools and technologies: findings of the EMT survey 2017." *The Journal of Specialised Translation*, 32: 26-60. [https://www.jostrans.org/issue32/art\\_rothwell.pdf](https://www.jostrans.org/issue32/art_rothwell.pdf).
- Sánchez-Gijón, Pilar. 2009. "Developing documentation skills to build do-it-yourself corpora in the specialized translation course." In *Corpus use and Translating: Corpus use for Learning to Translate and Learning Corpus Use to Translate*, edited by Allison Beeby, Patricia Rodríguez Inés, and Pilar Sánchez-Gijón, 109-127. Amsterdam: John Benjamins.
- Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Varantola, Krista. 2003. "Translators and disposable Corpora." In *Corpora in Translation Education*, edited by Federico Zanettin, Silvia Bernardini, and Dominic Stewart, 43-54. Manchester: St. Jerome Publishing.
- Veiga Díaz, Maria Teresa. 2016. "Compilación y explotación de un corpus ad hoc como herramienta para la adquisición de competencias específicas y transversales en el aula de traducción científica y técnica." In *New insights into Corpora and Translation*, edited by Daniel Gallego-Hernández, 41-56. Newcastle upon Tyne: Cambridge Scholars Publishing.
- Zanettin, Federico. 2002. "Corpora in translation practice." In *Language Resources for Translation Work and Research. LREC 2002 Workshop Proceedings*, edited by Elia Yuste Rodrigo, 10-14. Las Palmas de Gran Canaria, 10-14.
- Zanettin, Federico. 2012. *Translation Driven Corpora: Corpus Resources for Descriptive and Applied Translation Studies*. Manchester: St. Jerome Publishing.