Studies in Second Language Learning and Teaching
Department of English Studies, Faculty of Pedagogy and Fine Arts, Adam Mickiewicz University, Kalisz
SSLLT 2 (4). 505-525
http://www.ssllt.amu.edu.pl

Working with corpora in the translation classroom

Ralph Krüger
Cologne University of Applied Sciences, Germany
University of Salford, UK
ralph.krueger@fh-koeln.de

Abstract

This article sets out to illustrate possible applications of electronic corpora in the translation classroom. Starting with a survey of corpus use within corpus-based translation studies, the didactic value of corpora in the translation classroom and their epistemic value in translation teaching and practice will be elaborated. A typology of translation practice-oriented corpora will be presented, and the use of corpora in translation will be positioned within two general models of translation competence. Special consideration will then be given to the design and application of so-called *Do-it-yourself* (DIY) *corpora*, which are compiled ad hoc with the aim of completing a specific translation task. In this context, possible sources for retrieving corpus texts will be presented and evaluated and it will be argued that, owing to time and availability constraints in real-life translation, the Internet should be used as a major source of corpus data. After a brief discussion of possible Internet research techniques for targeted and quality-focused corpus compilation, the possible use of the Internet itself as a *macro-corpus* will be elaborated. The article concludes with a brief presentation of corpus use in translation teaching in the MA in Specialised Translation Programme offered at Cologne University of Applied Sciences, Germany.

Keywords: corpora, translation classroom, do-it-yourself (DIY) corpora, Web as macro-corpus

The Origins of Corpus Use in Translation: Corpus-Based Translation Studies

The start of the systematic use of corpora in translation theory can be dated back to the early 1990s. It was especially propagated by Mona Baker (e.g., 1993, 1995) who, working at that time with John Sinclair at Birmingham University in the context of applied linguistics (Beeby, Rodríguez Inés, & Sánchez-Gijón, 2009, p. 1), laid the groundwork for a research paradigm termed "corpus-based translation studies." Within this field of research, a corpus is generally understood as "a collection of texts held in machine-readable form and capable of being analysed automatically or semi-automatically in a variety of ways" (Baker, 1995, p. 225). In Holmes' (1972) map of translation studies as visualised by Toury (1995; see Figure 1), corpus-based translation studies is subsumed under the descriptive branch of "pure" translation studies and is therefore closely linked to Toury's research paradigm of Descriptive Translation Studies (Laviosa, 2002, p. 5).

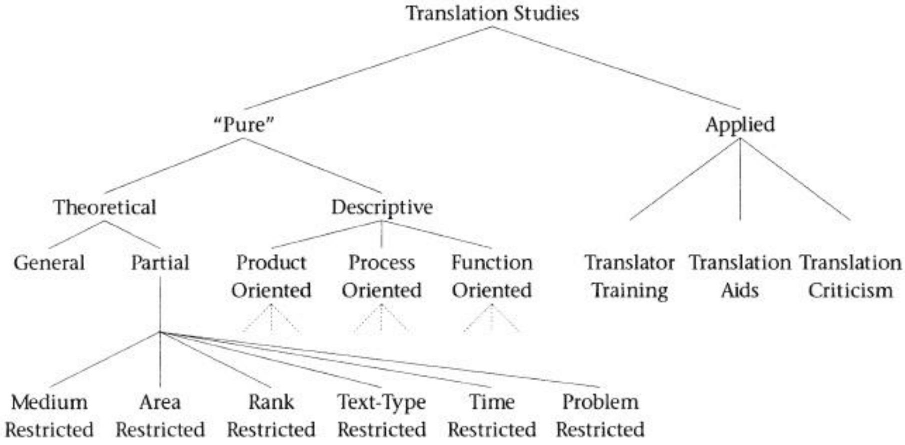

Figure 1 The Holmes-Toury map of translation studies (Toury, 1995, p. 10)

The primary aim of this new field of research was thus to describe characteristics of the translation product and the translation process or the function of a translation in its socio-cultural context and to feed this data to the theoretical branch of "pure" translation studies for testing explanatory hypotheses. The relation between the theoretical, the descriptive and the applied branch is, of course, a dialectical one (see Laviosa, 2002, p. 10).

One of the main advantages of corpus-based translation studies is that it allows for an extension of the traditional and widely practiced comparison between source texts (ST) and target texts (TT). Since corpus-based studies usually involve a rigorous methodology with a set of fine-grained design criteria (see e.g., Krein-Kühle, 2003), they allow for a better contextualisation and control of the texts to be investigated and provide a higher representativeness, generalisability and replicability of the findings. However, the dominant research focus of corpus-based translation studies is not on large-scale, methodologically sound ST-TT comparisons, but rather on the investigation of "the nature of translated text as a mediated communicative event" (Baker, 1993, p. 243). Speaking of the nature of translated text betrays the assumption that translation is somehow "different" from autonomous text production, that it is a "third code," which exhibits "its own standards and structural presuppositions and entailments" (Frawley, 1984, p. 169). Corpus-based translation studies thus set out to uncover the distinctive features of this third code, not by comparing translations to their source texts, but rather to autonomous texts in the target language. The focus here lies on the identification and investigation of so-called "universals of translation," which were defined as "linguistic features which typically occur in translated texts and are thought to be the almost inevitable by-products of the process of mediating between two languages" (Laviosa, 2002, p. 43). Two prominent examples of potential universals of translation would be explicitation and normalization (Olohan, 2004, p. 37), that is, translations are hypothesised to be informationally more explicit and stylistically more conventional than nontranslated texts.

Corpus Typology in Corpus-Based Translation Studies

The canonical corpus typology of corpus-based translation studies was developed by Laviosa (2002, p. 34 ff.) and, due to its complexity, cannot be presented here in full. However, two distinct corpus types can be identified that have traditionally dominated corpus-based translation research along the two research dimensions described above (i.e., ST-TT comparisons and comparisons between translations and nontranslated texts in the target language).

A *translation corpus* (also called *parallel corpus*) consists of one or more texts in the source language and their translation(s) into the target language (Baker, 1995, p. 230). The designation *parallel corpus* is widely established in corpus-based translation studies, but it has sometimes been criticised for its possible terminological confusion (e.g., Johansson, 1998; Krein-Kühle, 2003). As Krein-Kühle (2003, p. 45) points out, the adjective *parallel* is traditionally used in the term *parallel texts*, which refers to original target language texts with a subject matter and communicative function comparable to that of a specific text to be translated (Göpferich, 1998, p. 184). Since the concept of parallel texts will have a prominent role in the discussion of corpora in translation teaching, the designation *translation corpus* will be used in this paper. Translation corpora basically represent an extension of the long-practised ST-TT comparison. As mentioned above, they are usually associated with a more rigorous methodology and provide a better empirical basis than isolated case studies.

A *comparable corpus*, on the other hand, may consist of translations and comparable nontranslated texts in the same language, or it may consist of original (as opposed to translated) texts in one or more languages (Olohan, 2004, p. 35). The first type of a comparable corpus features prominently in translation research, since comparing translations with nontranslated texts in the same language specifically allows for the investigation of translational universals that allegedly constitute the distinctive features of translation. In line with the primary aim of corpus-based translation studies, it is this corpus type which has been investigated most extensively and which has generated the most widely recognised research results (e.g., Olohan & Baker, 2000). The second type is probably more relevant to translation teaching, since students faced with a translation task are normally not interested in distinctive features of translated language. In fact, given that in real life it is generally required that a translation should read like a text originally produced in the target language, these features are indeed what students are normally encouraged to avoid in their translations. Instead, students could use a comparable corpus of original target-language texts to study the idiomatic usage of terms and their collocates or the natural target-language style of specific text types or genres (Bowker & Pearson, 2002, p. 203 ff.) and try to reflect this usage or style in their translations. It is often claimed that studies of translation corpora prioritise the translation process, since a ST-TT comparison allows, at least to some extent, the retracing of the translational decision-making process, while studies of comparable corpora focus on the translation product (Stewart, 2000, p. 210). What can certainly be observed is that the epistemic aims associated with corpora in translation studies are slightly different from those in translation teaching. Therefore, moving from the general role and position of corpora in translation studies, some specific corpus types and their potential applicability in translation teaching will be presented in the following paragraphs.

Corpora in Translation Teaching

The systematic use of corpora in translation teaching[1] started more recently than the theoretical reflection on and the investigation of corpora in translation

[1] In the following discussion, it is generally assumed that the students translate out of a foreign language into their native language, since this is what will usually be required of them in their later professional career. Where corpora can be exploited for a translation into a foreign language, this will be specifically mentioned in the text.

studies. Trying to establish a link between the theoretical work of corpus-based translation studies and the use of corpora as learning aids in the translation classroom, Bernardini, Stewart and Zanettin (2003, p. 1) term this latter enterprise "applied corpus-based translation studies." With reference to the Holmes-Toury map of translation studies, corpora can therefore be used as resources to be employed in translator training and they can be used as translation aids in their own right, having direct relevance to translation practice. One of the main advantages of translation teaching with corpora over traditional translation teaching is generally considered to be the fact that the presence of corpora reduces the role of the teacher's intuition in the translation classroom and at the same time assigns more importance to the students and their documentation skills (Rodríguez Inés, 2009, p. 131). By providing alternative sources of authority as well as a set of authentic data, corpora can also shift the role of the teacher from that of the principal information *provider* to that of an information *facilitator* (Rodríguez Inés, 2009, p. 130, p. 133), who develops the procedural knowledge of the students to enable them to gain declarative knowledge in a more autonomous way.

Approaches to the Use of Corpora in Translation Teaching

There are two complementary approaches to the use of corpora in the translation classroom: *corpus use for learning to translate* and *learning corpus use to translate* (Beeby et al., 2009, p. 1). In the first approach, the compilation and control of the corpus material falls within the responsibility of the translation teacher, who then presents the students with preselected data (which is usually tailored to a specific translation task) and guides the students' analysis of this data. From the students' perspective, this microscopic approach focuses on the immediate relevance of the corpus as a "performance-enhancing tool" (Varantola, 2003, p. 59), which can be queried in order to solve specific translation problems. The second approach represents a more macroscopic perspective in that students themselves have to compile the corpora before they can apply them to solve any translation problems. This approach does not primarily focus on the immediate corpus-use related aspects but instead on the various translation-related issues of corpus compilation, for example, corpus design, search strategies, assessment of potential corpus sources, assessment of the adequacy and relevancy of corpus texts, general software literacy, and so on (cf. Varantola, 2003, p. 69). It should be obvious that these two approaches are highly complementary and should ideally be combined in the translation classroom to provide students with a complete set of corpus skills.

Corpus Typology in Translation Teaching

In translation didactics, there is a general distinction between three major corpus types with different epistemic values (see Bernardini et al., 2003, p. 6): monolingual corpora, comparable bilingual corpora and bilingual translation corpora. This typology overlaps to a considerable extent with the typology established in corpus-based translation studies, but there are still several differences.

Monolingual corpora, usually containing texts originally produced in the target language, can, for example, provide students with information about the idiomatic use of terms and their collocates, syntactic constructions or genre and domain conventions in the target-language environment. If the corpora are designed as specialised corpora containing texts of a specific subject matter such as engineering or economics, they can also provide students with explanatory contexts for the various concepts of the specialised field (Bowker & Pearson, 2002, p. 207-208). In this context, Sánchez-Gijón (2009, p. 120) claims that factual information is generally obtained from a corpus containing source-language texts, "since that is the language in which cognitive problems will occur." This may be true; however, since the translator produces a target text that is usually geared towards a target-language readership, it is usually the concepts and the field-specific conceptual structuring of the target language and culture that will ultimately be of relevance. Also, it may be less difficult for students who are new to a specialised field to acquire the domain knowledge required for high-quality translation via their native language (which is usually not the source but the target language), because the cognitive load in this language will probably be lower. Armed with a basic knowledge of the field-specific concepts in their native language, it will then be much easier for the students to analyse the source-language concepts. Since these monolingual corpora basically serve the same function as the well-established concept of *parallel texts*, they will be termed *parallel-text corpora*[2] for the purpose of this paper. Offering empirical information on idiomaticity and natural language use, parallel-text corpora can also be a useful resource for students faced with a translation into a foreign language (Bernardini et al., 2003, p. 6), especially considering that their text production competence in the foreign language will usually be much lower than in their native language. Therefore, when translating into a foreign language, students are all the more dependent on authentic examples of natural language use that can be offered by parallel-text corpora.

Comparable bilingual corpora contain original source and target-language texts and allow for a comparative analysis of the same parameters

[2] Which is not to be confused with the discarded designation *parallel corpora* (see above).

that can be studied in a monolingual environment in parallel-text corpora. Working with these corpora, students gain a better understanding not only of original target-language texts but also of original source-language texts and their natural make-up (Bernardini et al., 2003, p. 6). In this case, it would make sense to include the source text to be translated in the comparable corpus and to expand it with texts of a similar subject matter, communicative function, and the like. It is this corpus type that can be used for the conceptual analysis of source-language texts proposed by Sánchez-Gijón (see above).

Finally, *bilingual translation corpora*, containing source texts and their translations, offer insights into the strategies employed by professional translators when dealing with specific translation problems on various levels. For example, students could query a bilingual translation corpus for terminological equivalents and perform a contrastive analysis of the underlying source and target-language terms. They could also analyse how certain stylistic features of the source text (e.g., post-modification or inanimate nouns + action verbs in English) were rendered in the translation. The idiomaticity of the various translation solutions identified could then be checked against a corresponding parallel-text corpus. This leads to another important aspect of corpus use, both in translation theory and in translation practice, namely the quality of the corpus texts (cf. Krein-Kühle, 2003, 2011). When compiling a bilingual translation corpus, it is particularly important to devise a set of quality criteria to ensure that the translations to be included in the corpus do not exhibit any strong signs of “translationese,” that is, unnatural target language patterns or elements or unusual frequencies of specific patterns or elements that can be traced back to source-language interference (see Olohan, 2004, p. 90).[3] As described above, the inclusion of low-quality translations in the corpus could be avoided by comparing the potential corpus texts with similar original target-language texts from a parallel-text corpus to see whether any significant structural or other deviations can be found. In this case, the parallel-text corpus would serve as a “reference corpus” (see Krein-Kühle, 2003, p. 50). An alternative, which may be more feasible if real-life translation constraints (especially time constraints) were to be taken into account, is to implement quality control measures at the corpus design stage and restrict the potential corpus texts to publications by specific authors, companies, organisations, and so on. The issue of corpus quality will be revisited in the discussion of the Internet as a source of corpus texts.

[3] Of course, the question of quality is not restricted to translated texts, but applies just as well to autonomous texts. However, since the issue of “translationese” or poor general translation quality is indeed a central aspect in translation practice (see Krein-Kühle, 2003, p. 3), the discussion will be restricted to translation quality here. The quality criteria proposed later in this paper can certainly be applied to nontranslated texts as well.

It has now been generally recognised in translation didactics that corpus use should not be regarded as a mere additional qualification to be acquired independently of “pure” translation competence, but that it rather forms part of wider translation competence itself (Rodríguez Inés, 2009). In the field of translation process research, considerable work has gone into developing models of translation competence, the best-known of which is perhaps the model devised by the PACTE[4] Group based at the Universitat Autònoma de Barcelona (e.g., PACTE 2003).

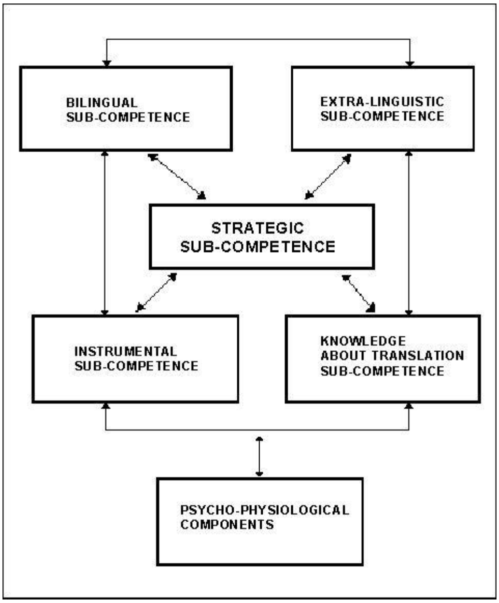

Figure 2 Translation competence model of the PACTE Group (PACTE, 2003, p. 60)

As can be seen in Figure 2, this model divides translation competence into four individual sub-competences (bilingual, extra-linguistic, instrumental and knowledge about translation), which are controlled by a strategic sub-competence. In this model, corpus use would form part of the instrumental sub-competence, which involves “procedural knowledge related to the use of documentation sources and information and communication technology applied to translation” (PACTE, 2003, p. 59). It should be pointed out that this description of the instrumental sub-competence only covers the technical side of corpus use, that is, the compilation of corpora and the application of specific corpus analysis software. The actual linguistic

[4] Process in the Acquisition of Translation Competence and Evaluation

or conceptual analysis of a corpus and the interpretation of the analysis results fall outside the description of this sub-competence. Within the instrumental sub-competence of the PACTE model, Rodríguez Inés (2009, p. 136) proposes a further sub-competence which refers to "the ability to use electronic corpora adequately to solve translation problems in an adequate manner." This specific sub-competence consists of four elements, namely the assimilation of basic principles involved in working with corpora, the building of corpora, the handling of corpus-related software and the use of corpora to solve translation problems (Rodríguez Inés, 2009, p. 136). The last element of this sub-competence covers the actual corpus analysis with regard to specific translation problems and thus fills the gap identified in the general description of PACTE's instrumental sub-competence.

Another prominent translation competence model that recognises corpus use as an integral part of overall translation competence is the model developed within the European Master's in Translation (EMT) network of the European Union. The EMT network is a partnership project between the European Commission and higher-education institutions in the Member States and was established in order to provide a quality label for translation programmes at university level that meet specific educational standards. The reference framework for translation competences shown in Figure 3 is specifically geared towards such university programmes and is intended to serve as a basis for developing the content of individual training modules (EMT Expert Group, 2009, p. 3).

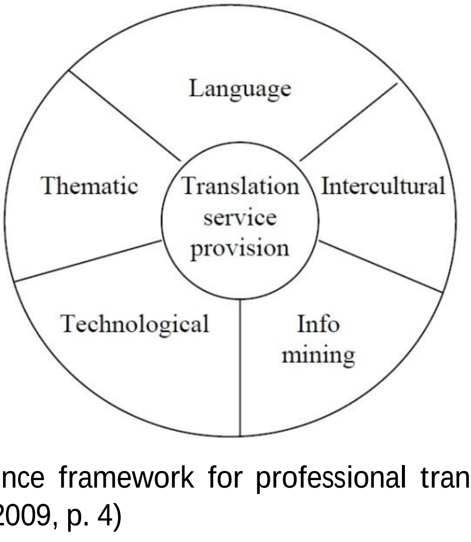

Figure 3 EMT reference framework for professional translation competence (EMT Expert Group, 2009, p. 4)

In this reference framework, corpus use is included as a component of the information mining competence and refers to the procedural competence of "knowing how to use tools and search engines effectively (e.g. terminology software, electronic corpora, electronic dictionaries)" (EMT Expert Group, 2009, p. 6). As with the PACTE model, this description only covers the technical side of corpus use, the linguistic side in this case being covered mainly by the language competence.

The inclusion of corpus use in these two major translation competence models is certainly evidence of the fact that the multiple advantages of these resources for translation teaching and practice have by now been widely recognised. Ideally then, a corpus approach should be adopted in the translation classroom that tries to reconcile the requirements of both translation teaching and practice, to familiarise students with the various aspects of corpus compilation and analysis while taking the constraints of translation practice into account.

Do-It-Yourself Corpora

In the following, a corpus type will be introduced that may prove especially useful to in vivo translation courses that try to model, at least to some extent, the professional working environment and the working conditions of future translators. Such a course would include, for example, access to IT software (especially text processing, translation memory and terminology management software) and the Internet and the translation of texts with high professional relevance but also the assessment of the various tools and techniques with regard to their applicability in professional contexts, which are often characterised by time and financial constraints. In this regard, Aston (2009, p. IX-X) points out the discrepancy between the increasing use of corpora in translation teaching and the rather low acceptance of these resources among professional translators. The reason, according to Aston, is that the construction and consultation of corpora is very time-consuming, making corpus-use "anti-economic in the short term" (p. IX-X). If we want to bridge this gap between corpus use in translation teaching and in translation practice, a corpus type is required that retains the epistemic advantages of corpora in general yet is tailored to the specific constraints and requirements of translation practice.

Such corpora are known as *do-it-yourself* (DIY) *corpora*, which are compiled "for the sole purpose of providing information – either factual, linguistic or field-specific – for the purposes of completing a translation task" (Sánchez-Gijón, 2009, p. 115). Other designations for this corpus type are *ad hoc corpora*, *disposable corpora*, or *virtual* or *ephemeral* corpora (see Corpas Pastor & Seghiri, 2009, p. 78). The designations *disposable* or *ephemeral* imply, however, that these corpora may not become part of a more permanent corpus or may not be retained by the translator as reference materials for future translations, which is probably not the case. Especially when establishing permanent relationships with clients requesting similar translations on an ongoing basis, texts which were initially intended as ad hoc resources for a specific translation task are often consulted again for subsequent translations and thus acquire a more permanent status as proven reference materials, which could over time lead to the compilation of a more comprehensive and stable corpus. Seen in this light, the designations *DIY corpora* or *ad hoc corpora* seem more suitable. According to Zanettin (2012, p. 64), one of the main advantages of DIY corpora is precisely being able to create them ad hoc in response to specific translation problems or information needs. Therefore, these corpora tend to be very precise and can be expanded anytime as required. The design criteria of these corpora are primarily determined by the source text, which basically guides the material to be included in the DIY corpus (see Varantola, 2003, p. 56). Depending on the epistemic requirements, DIY corpora can be constructed as monolingual parallel-text corpora, comparable bilingual corpora or bilingual translation corpora (as detailed above).

In the corpus-compilation stage, Sánchez-Gijón (2009, p. 115 ff.) identified three possible sources for retrieving potential corpus texts: the client, specialist centres and the Internet. In the translation classroom, the teacher can obviously act as client and provide the students with pre-selected texts for their DIY corpora. This option is characterised by a strong bias towards the *corpus-use-for-learning-to-translate* approach and does not develop the corpus compilation skills that students need to become fully autonomous corpus users. Providing students with access to specialist documentation sources (e.g., databases, academic journals, specialised libraries, etc.) would ensure a high quality of the corpus texts, but as Sánchez-Gijón (2009, p. 116) points out, these sources are not always readily available for the different subject matters of the texts to translate in the classroom or in later translation practice. Moreover, consulting these sources would again be quite time-consuming and run counter to the efforts of making corpora a feasible resource for translation teaching *and* practice.

The Internet as a Source of Corpus Texts

The Internet, in contrast, does not suffer from any of these constraints. It can be used as a source for autonomous corpus compilation by students, it is highly accessible and it provides a vast, albeit unstructured, body of information. The Internet therefore seems to be the most viable source for compiling DIY corpora (cf. Sánchez-Gijón, 2009, p. 116). However, the undisputed advantages of the internet as a source of corpora are accompanied by several drawbacks, the most prominent being the lack of structure of the content provided and its varying quality. Therefore, if a targeted and high-quality DIY Web corpus is to be compiled from the Internet, this lack of structure as well as potential quality concerns have to be compensated by a rigorous corpus compilation approach, which can be roughly divided into the following three phases (for the first two phases, cf. Sánchez-Gijón, 2009, p. 116-117):

1. Determining the characteristics of the resource that will provide the corpus texts.
2. Devising specific search strategies to carry out more precise searches.
3. Establishing and applying quality criteria.

These three phases will now be discussed in detail.

Determining the characteristics of the resource that will provide the corpus texts. Normally, students will access the Internet using a conventional search engine like Google, Yahoo! or Bing. Before they start compiling a DIY Web corpus, students should be made aware of the characteristics and functioning principles of these search engines in order to make more informed searches. The most fundamental principle to be pointed out in this context is the difference between the *surface Web*, that is, that part of the Internet which can be accessed via conventional search engines, and the much bigger *deep Web*, which is inaccessible by these engines (e.g., password-protected websites or sites that are dynamically generated using local database content; cf. Griesbaum, Bekavac, & Rittberger, 2009). Therefore, students should be aware that the information they obtain using conventional search engines is by no means all the information that exists on a particular subject, but that more differentiated searches may be necessary. Students should also be introduced to the difference between universal search engines (Google et al.) and vertical search engines that focus on a specific field or discipline (Sánchez-Gijón, 2009, p. 117). If, for example, students are tasked with a scientific or technical translation, they could use the vertical search engine scirus.com to obtain more targeted search results. Other relevant characteristics of search engines that can be made transparent to students are the ranking criteria which determine the sorting order of the search results (e.g., Google's PageRank algorithm, which assigns a numerical weight to different Web pages according to the number of hyperlinks to these pages and the PageRank of the pages hosting these hyperlinks; see Dopichaj, 2009) and the distinction between natural listings (those results that are listed according to objective ranking algorithms) and paid listings (those results that the search engine provider is paid for to present regardless of objective ranking criteria) (see also Lewandowski & Höchstötter, 2009). Understanding these principles of result presentation by conventional search engines may prompt students to look harder for the information required instead of relying on the first two or three results that are presented at the top of the page.

Devising specific search strategies to carry out more precise searches. After reviewing these basic characteristics of common search engines, the next step would be to introduce students to specific search strategies to narrow down the searches to yield only results with direct relevance to the DIY corpus being compiled (cf. Sánchez-Gijón, 2009, p. 117). A common search operator[5] offered by sites such as Google or Yahoo! is the operator *site:*, which restricts the search to a specific website or domain. For example, the search string *site www.deutsche-bank.de* would only yield results from the website of Deutsche Bank, whereas the search string *site:.edu* would restrict the search to websites with the top-level domain *.edu*, that is, to sites of educational institutions. These search strategies may be useful if the students are interested in the terminology or the style employed by a specific client (e.g., the corporate language or terminology of Deutsche Bank) or if they are looking for high-quality explanatory texts on a specific subject matter (which can reasonably be expected to be provided by educational institutions). Another helpful search operator is *filetype:*, which restricts the search to documents with a specific file format. The search string *filetype: pdf*, for example, would only yield PDF files, which are usually claimed to have a more stable content compared to files in other text formats (Zanettin, 2012, p. 58). A last strategy to be presented here is excluding a word or a complete website from the search by placing a dash (-) before the corresponding site or word. If, for example, students are looking for texts about computer mice and they want to exclude any search results referring to the identically named animals (cf. Zanettin, 2012, p. 57), they could use the search string *mouse -animal* or *mouse -rodent*. Likewise, if the students know that a specific site does not offer texts that meet their quality requirements, they can exclude that site from their search by using the search string *-name of the site to exclude*. There are many more options available for specifying Web searches (e.g., setting the language and region of the website, setting a date range, etc.), and most search engines provide a support page with corresponding information.[6]

It is important to mention at this point that these search strategies are primarily geared towards the compilation of monolingual corpora (see Zanettin, 2012, p. 62). The compilation of a translation corpus via the Internet is more complicated (if only because there may not be a translation of a particular text in the first place) and generally requires additional search strategies. For example, the websites of major international companies or organisations are often available

[5] For a list of common search operators offered by Google see http://support.google.com/websearch/bin/answer.py?hl=en&answer=136861 (last accessed: 18/01/2013)

[6] The corresponding Google support page can be found under support.google.com/websearch/ (last accessed: 21/01/2012).

in a variety of languages, and these languages are usually shown at the top of the page. If the required language is available on the website, the sitemap of the original version of the website can be used to identify the category structure of the website and to locate the source text in this structure (e.g., a report in the "Reporting and Events" subcategory of the category "Investor Relations" of the website of a financial institution). If the target-language version of the website has the same category structure (this is not always the case), it should be fairly easy to locate the corresponding target text. If the link to the source text in the URL bar of the browser contains an ISO language code (e.g., */en/*), it may also be possible to substitute this code with the ISO code of the target language (e.g., */de/*) to locate the translation. However, as already mentioned, a translation may not be available in the first place, and even if one is available, creative solutions beyond the strategies just described may be necessary to locate it. Nonetheless, using the aforementioned as well as further search strategies can provide a structured approach to corpus compilation from an inherently unstructured source of information. It is, however, important to incorporate a qualitative dimension in such an approach.

Establishing and applying quality criteria. While the Internet offers a vast amount of potential corpus texts, these texts will be characterised by a very uneven quality (Zanettin, 2012, p. 56). If students want to avoid carrying over erroneous or inappropriate solutions into their translations, they must be aware of the potential quality issues involved in corpus compilation. It is certainly advisable for students to carry out a rough quality assessment of the potential corpus texts. With this procedure, seriously flawed texts like machine translations without any post-editing will almost certainly be spotted and excluded. However, less obvious quality defects such as source-language interferences in translations may not be as easy to detect by students, since their textual competence will usually not be fully developed yet. This will probably be the case with their native language, and almost certainly with their foreign language(s). It may thus be advisable to develop a list of extra-textual criteria (like authorship or publishing organisation) and to reflect these criteria in the search strategies previously described to identify the corpus texts. The operator *site:*, for example, could be used to restrict the search a priori to specific websites or top-level domains which are likely to fulfil the established quality criteria, or a dash (-) could be used to exclude websites or top-level domains that may not meet the quality requirements.

Using the Internet as a Macro-Corpus

As well as being used as a source for corpus compilation, the Internet can itself be used as a *macro-corpus* (Zanettin, 2012, p. 56) to be queried directly for linguistic or conceptual information related to a specific translation task. Since this is again a primarily monolingual approach, this macro-corpus will probably be used as a parallel-text corpus.

Using conventional search engines. In order to query the Internet as a macro-corpus, students could principally resort to conventional search engines again, using some of the above mentioned strategies in combination with several more linguistically-oriented search functions. For example, most conventional search engines allow for the use of wildcard characters (Zanettin, 2012, p. 57). If students were faced with a source-text element like "to issue shares" and they know the German equivalent of *shares* (which would be *Aktien* in this case) but are unsure about the proper German collocate, they could devise a search string that contains a wildcard for the unknown verb (e.g., *die von dem Unternehmen * Aktien*) and check the results for a verb that might fit. In this case, a potential candidate would be *ausgeben*, and the word group *Aktien ausgeben* could be further verified using one of the general search strategies explained above (e.g., by using the operator *site:* it could be established whether the phrase *Aktien ausgeben* is used on the websites of major publicly traded German companies, which would be strong evidence that *ausgeben* is indeed the required collocate). The Internet may also be queried directly for explanatory contexts for specific concepts. Conventional search engines like Google offer the search operator *define:*, which yields a list of definitions of the search term along with hyperlinks to the corresponding Web pages. A more sophisticated strategy would be to formulate a search string in the form of the classical Aristotelian definition, leaving the *definiens* unspecified (cf. Bowker & Pearson, 2002, p. 206 ff.). If, for example, students were looking for a definition of the term *frequency converter*, they could query the Internet using the search string *A frequency converter is*. The linking element can also be varied to yield hypernymic information (e.g., *A frequency converter is a kind of*), meronymic information (e.g., *A frequency converter consists of/contains/is a part of*) or functional information (e.g., *A frequency converter is used to*).

Using special Web concordancers. While these strategies may provide a viable approach to using the Internet itself as a macro-corpus, conventional search engines generally do not present the results in a format that invites straightforward linguistic analysis. The selection and ordering of results does not follow any linguistic criteria (but is rather determined by ranking algorithms and commercial aspects, as detailed above) and the results as presented by a search engine are static and do not allow for any manipulation such as sorting the concordance lines, generating collocations, and the like (see Zanettin, 2012,

p. 58 ff.). Considering these shortcomings, special applications for using the Internet as a macro-corpus have been developed which offer specific functions tailored to the linguistic analysis of the results. One of the best-known of these Web concordancers is the programme WebCorp Live (2013), which was developed by the Research Development Unit for English Studies in the School of English at Birmingham City University (see also Zanettin, 2012, p. 59). This concordancer involves a search phase and a postprocessing phase. In the search phase, options such as the search engine to use (e.g., Google or Bing), the language of the Web pages to search and the number of concordance lines per Web page can be specified. Several options which are provided by conventional search engines are also available in WebCorp Live, for example the specification of a certain site or domain to search (this is the equivalent of Google's search operator *site*), the use of word filters (e.g., words that must or must not appear on the same page as the search term) and the use of wildcards in the search string. The search strategy previously described to retrieve explanatory contexts for individual concepts can also be applied in WebCorp Live. In the post-processing phase, the number of words or characters to display to the left and right of the search term can be specified, and the concordances can be sorted by date or alphabetically (e.g., sort by the words to the left or right of the search term). Also, a table with the most frequent collocates can be generated and stopwords to exclude from the list (e.g., high-frequency words like *a* and *the*) can be specified. *WebCorp Live* also caches the search results for seven days. The results can be saved on a local computer (Zanettin, 2012, p. 60) and the hyperlink to the cached results can be shared with other researchers, students or translators. Other tools with a range of functions similar to that of WebCorp Live are for example WebAsCorpus (2013) and KWiCFinder (2013) (Zanettin, 2012, p. 59). These tools offer some powerful functions to conduct a linguistically-oriented analysis using the Internet as a macro-corpus. These functions should ideally be combined with the various research strategies described previously in order to obtain high-quality results with direct relevance to the translation task at hand.

Corpus Use in Translation Teaching at Cologne University of Applied Sciences

In order to exemplify some of the issues discussed above, I will briefly illustrate how corpora are used in the 2-year MA programme in Specialised Translation[7] offered by the Institute of Translation and Multilingual Communication at Cologne University of Applied Sciences. The focus will be on the course called Translation

[7] For more information on the programme see http://www.international-office.fh-koeln.de/english/faculties/overview/f03/courses/u/01550.php (last accessed: 17/01/2013).

Project using Translation Tools, which tries to project into the translation classroom as closely as possible the professional environment that students will encounter in their later careers as translators. In this in vivo translation course, the students are introduced to key aspects of real-life translation projects (client communication, document management, handling client instructions, research strategies for the Internet, etc.) as well as to relevant computer software (e.g., translation memory software, terminology software, quality assurance software, Web concordance software). After this introductory phase, the students are asked to work on various small translation projects (usually involving specialised texts of medium difficulty in the fields of economics or engineering) in which these competences are brought together and developed further. The course thus has a strong, although by no means exclusive, focus on the instrumental sub-competence in the PACTE competence model or the information mining competence and the technical competence in the EMT reference framework. Corpora are introduced in the course within the context of linguistic and conceptual research for the translation projects, and the necessary competences are taught in two different teaching units, which roughly correspond to the two didactic approaches introduced previously, that is, *corpus use for learning to translate* and *learning corpus use to translate*. In the first unit, “Internet research strategies for translators,“ the students are introduced to the basic characteristics of conventional search engines and to the various techniques for compiling DIY corpora using the Internet. In this unit, the students are also familiarised with the tools and strategies for querying the Internet itself as a macro-corpus. This unit thus focuses on the corpus compilation skills of the students. The second unit, “Text analysis in translation,” is loosely based on the translation-oriented text analysis model developed by Nord (2009) and introduces the students to various linguistic concepts that are relevant to a well-founded corpus analysis (e.g., genre conventions, register, textual micro- and macro-structure, lexical and syntactic analysis, etc.). This unit, therefore, is more concerned with the competences required for the actual corpus analysis with regard to specific translation problems.

In one exercise, the students were asked to translate various excerpts of the annual report of an international bank from English into German (e.g., the letter from the chairman of the management board). Using the research competences acquired during the course, the students built a parallel-text corpus consisting of German annual reports of various German banks as well as a translation corpus containing original English annual reports of various international banks and their translations into German. The students were then asked to investigate the genre conventions, register and lexis of the parallel-text corpus and to reflect the results of their analysis in the translation. The parallel-text corpus was also used to provide explanatory contexts for specialised concepts. The translation corpus was mainly used to identify terminological equivalents or to see how specific source-text patterns were rendered in the translation. The students were also encouraged to use WebCorp Live to query the Internet as a macro-corpus. The analysis of the parallel text and translation corpora was conducted manually, without resorting to special concordance software for monolingual corpora (e.g., WordSmith) or bilingual corpora (e.g., ParaConc), since the preprocessing of the texts was found to be too time-consuming to be reconcilable with professional constraints. This was especially the case for copy-protected PDF files and for bilingual files, which had to be prealigned for their use in a bilingual concordancer or to be preprocessed for the automatic alignment function offered by these programmes. Thus, the only true concordance functions available to the students were those offered by WebCorp Live.

The students' feedback on the use of corpora was largely positive. They particularly appreciated the availability of a high-quality translation corpus which provided immediate solutions to various translation problems. The parallel-text corpus was, for the most part, not used as an independent resource. The students mainly used it as a "back-up" corpus to check whether the terminology and structural patterns found in the target texts of the translation corpus were also present in original target-language texts. They also made extensive use of the Internet as a macro-corpus, especially when the parallel-text and translation corpora did not yield any ready-made solutions to their translation problems. At first, the students were reluctant to work with WebCorp Live and mostly used Google for their searches, but once they became more familiar with WebCorp, they used the programme more readily. In this context, the function for generating a list of frequent collocates was seen as particularly helpful. Altogether, corpus use made the students feel more confident with their own translation solutions, especially if these were justified by "independent" sources of natural language data such as parallel-text corpora.

Concluding Remarks

This paper has hopefully made a strong case for the use of corpora in the translation classroom by highlighting the multiple advantages of these resources as teaching aids but also by demonstrating how the application of corpora can be reconciled with the constraints of translation practice. The competent use of corpora requires various competences, both linguistic and technological, and therefore working with corpora in the translation classroom provides an ideal test case for bringing these diverse competences together. The technological dimension of corpus use also ties in perfectly with the general call for computer literacy in today's professional world, and the research and documentation skills the students acquire by compiling corpora (especially from the Internet) will certainly be valuable beyond the immediate field of translation practice.

References

Aston, G. (2009). Foreword. In A. Beeby, P. Rodríguez Inés, & P. Sánchez-Gijón (Eds.), *Corpus use and translating. Corpus use for learning to translate and learning corpus use to translate* (pp. IX-X). Amsterdam: John Benjamins.

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. 17-45). Amsterdam: John Benjamins.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, 7(2), 223-243.

Beeby, A., Rodríguez Inés, P., & Sánchez-Gijón, P. (2009). Introduction. In A. Beeby, P. Rodríguez Inés, & P. Sánchez-Gijón (Eds.), *Corpus use and translating. Corpus use for learning to translate and learning corpus use to translate* (pp. 1-8). Amsterdam: John Benjamins.

Bernardini, S., Stewart, D., & Zanettin, F. (2003). *Corpora in translator education: An introduction*. In S. Bernardini, D. Stewart, & F. Zanettin (Eds.), Corpora in translator education (pp. 1-13). Manchester: St. Jerome.

Bowker, L., & Pearson, J. (2002). *Working with specialised language. A practical guide to using corpora*. London: Routledge.

Corpas Pastor, G., & Seghiri, M. (2009). Virtual corpora as documentation resources: Translating travel insurance documents (English-Spanish). In A. Beeby, P. Rodríguez Inés, & P. Sánchez-Gijón (Eds.), *Corpus use and translating. Corpus use for learning to translate and learning corpus use to translate* (pp. 75-107). Amsterdam: John Benjamins.

Dopichaj, P. (2009). Ranking-Verfahren für Web-Suchmaschinen. In D. Lewandowski (Ed.), *Handbuch Internet-Suchmaschinen* (pp. 101-115). Heidelberg: AKA.

EMT Expert Group (2009). *Competences for professional translators, experts in multilingual and multimedia communication*. Retrieved from the European Master's in Translation website of the DG Translation of the European Commission: http://ec.europa.eu/dgs/translation/programmes/emt/key_documents/emt_competences_translators_en.pdf

Frawley, W. (1984). Prolegomenon to a theory of translation. In W. Frawley (Ed.), *Translation: literary, linguistic and philosophical perspectives* (pp. 159-175). London: Associated University Press.

Göpferich, S. (1998). *Paralleltexte*. In M. Snell-Hornby, H. G. Hönig, P. Kußmaul, & P. A. Schmitt (Eds.), *Handbuch Translation* (pp. 184-185). Tübingen: Stauffenburg.

Griesbaum, J., Bekavac, B., & Rittberger, M. (2009). Typologie der Suchdienste im Internet. In D. Lewandowski (Ed.), *Handbuch Internet-Suchmaschinen* (pp. 18-52). Heidelberg: AKA.

Holmes, J. S. (1972). *The name and nature of translation studies. 3rd international congress of applied linguistics: Abstracts*. Copenhagen.

Johansson, S. (1998). On the role of corpora in cross-linguistic research. In S. Johansson & S. Oksefjell (Eds.), *Corpora and cross-linguistic research* (pp. 3-24). Amsterdam: Rodopi.

Krein-Kühle, M. (2003). *Equivalence in scientific and technical translation. A text-in-context-based study* (Unpublished doctoral dissertation). University of Salford, UK.

Krein-Kühle, M. (2011). Register shifts in scientific and technical translation. A corpus-in-context study. In M. Olohan, M. Salama-Carr (Eds.), *Science in translation. Special issue of The Translator*, *17*(2), 391-413.

KWiCKFinder [Computer software]. (2013). Retrieved from http://www.kwicfinder .com

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

Lewandowski, D., & Höchstötter, N. (2009). Standards der Ergebnispräsentation. In D. Lewandowski (Ed.), *Handbuch Internet-Suchmaschinen* (pp. 204-219). Heidelberg: AKA.

Nord, C. (2009). *Textanalyse und Übersetzen. Theoretische Grundlagen, Methode und didaktische Anwendung einer übersetzungsrelevanten Textanalyse*. Tübingen: Groos.

Olohan, M. (2004). *Introducing corpora in translation studies*. London: Routledge.

Olohan, M., & Baker, M. (2000). Reporting *that* in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, *1*(2), 141-158.

PACTE Group (2003). Building a translation competence model. In F. Alves (Ed.), *Triangulating translation: Perspectives in process oriented research* (pp. 43-66). Amsterdam: John Benjamins.

Rodríguez Inés, P. (2009). Evaluating the process and not just the product when using corpora in translator education. In A. Beeby, P. Rodríguez Inés, & P. Sánchez-Gijón (Eds.), *Corpus use and translating. Corpus use for learning to translate and learning corpus use to translate* (pp. 129-149). Amsterdam: John Benjamins.

Sánchez-Gijón, P. (2009). Developing documentation skills to build do-it-yourself corpora in the specialised translation course. In A. Beeby, P. Rodríguez Inés, & P. Sánchez-Gijón (Eds.), *Corpus use and translating. Corpus use for learning to translate and learning corpus use to translate* (pp. 109-127). Amsterdam: John Benjamins.

Stewart, D. (2000). Poor relations and black sheep in translation studies. *Target*, *12*(2), 205-228.

Toury, G. (1995). *Descriptive translation studies and beyond*. Amsterdam: John Benjamins.

Varantola, K. (2003). Translators and disposable corpora. In S. Bernardini, D. Stewart, & F. Zanettin (Eds.), *Corpora in translator education* (pp. 55-70). Manchester: St. Jerome.

WebAsCorpus [Computer software]. (2013). Retrieved from http://webascorpus.org

WebCorp Live [Computer software]. (2013). Retrieved from http://www.webcorp.org.uk/live/index.jsp

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