Arab World English Journal
INTERNATIONAL PEER REVIEWED JOURNAL ISSN: 2229-9327
مجلة اللغة الانكليزية في العالم العربي

Arab World English Journal (AWEJ) Volume.8 Number3 September 2017 Pp. 319 –337
DOI: https://dx.doi.org/10.24093/awej/vol8no3.21

Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching

Hind M. Alotaibi
Collage of Languages & Translation
King Saud University, Saudi Arabia

Abstract

Parallel corpora can be defined as collections of aligned, translated texts of two or more languages. They play a major role in translation and contrastive studies, and are also becoming popular in translation training and language teaching, with the advent of the data-driven learning (DDL) approach. Despite their significance, however, Arabic seems to lack a satisfactory general-use parallel corpus resource. The literature describes few Arabic–English parallel corpora, and these few are usually inaccurate and/or expensive. Some are small in size, while others are restricted in terms of genre, failing to meet the requirements of many academics and researchers. This paper describes an ongoing project at the College of Languages and Translation, King Saud University, to compile a 10-million-word Arabic–English parallel corpus to be used as a resource for translation training and language teaching. The bidirectional corpus can be used to compare translated and source language and identify differences. The corpus has been manually verified at different stages, including translation, text segmentation, alignment, and file preparation; it is available as full-text in XML format and through a user-friendly web interface that provides a concordancer to support bilingual search queries and several filtering options.

Keywords: Arabic, data-driven learning, English, language teaching, parallel corpus, translation training

Cite as: Alotaibi, H. M. (2017). Arabic-English Parallel Corpus: A New Resource for Translation Training and Language Teaching. *Arab World English Journal, 8 (3).* DOI: https://dx.doi.org/10.24093/awej/vol8no3.21

Introduction

According to Baker (1995), a parallel corpus is one that "consists of original, source language texts in language A and their translated versions in language B" (p. 230), while Sinclair (1995) described it as a "collection of texts, each of which is translated into one or more other languages" (p. 19).

Parallel corpora have been used widely in applied linguistics and have gained special attention in language teaching and learning (Botley, Glass, McEnery, & Wilson, 1996; Granger & Lefer, 2016; Wichmann & Fligelstone, 2014), translation studies and translator training (Baker, 1993, 1995; Olohan, 2004), machine translation (Rauf & Schwenk, 2011; Tian et al., 2014), comparative and contrastive linguistics (Johansson, 1999, 2007; Sharoff, Rapp, Zweigenbaum, & Fung, 2013), lexigraphy, and terminology studies (Kilgarriff, 2013; Teubert, 2002).

Many researchers believe that parallel corpora provide new insights into both source and translated languages that cannot be gained from exploring monolingual corpora, enabling researchers and students to compare languages and their semantic and cultural features, among other advantages.
According to Xie (2015) a parallel corpus:

> provides a platform for researchers to conduct crosslinguistic research whereby the linguistic and cultural differences of the two languages and their effects on second language (L2) learning can be compared and analysed systematically. (p. 157)

Thus, parallel corpora enhance our knowledge of languages comparatively, contributing to many fields (aside from the studies cited above, see Aijmer, 2009; Baker, 1998, 1999; Barlow, 2000; Bennett, 2010; Bernardini, 2016; Boulton, 2011; Bowker, 2002; Cobb & Boulton, 2015; Perez et al. ,2014; Sinclair, 1991, 2004; Tribble, 2015; Zanettin, 2014).

Data-driven learning

Data-driven learning, or DDL, is a term first introduced by Tim Johns (1993, 1997) to describe a teaching approach where language learners act as language researchers. This approach involves using corpus linguistics in teaching by exposing students to data and encouraging them to discover linguistic rules and patterns from concordance lines. According to Johns (1991, p. 2) "the language learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data." Odlin (1994) describes DDL as:

> an approach to language teaching that gives central importance to developing the learner's ability to 'puzzle out' how the target language operates from examples of authentic usages. This approach is particularly associated with the use of computer concordances in the classroom but can be extended to other situations where the student has to work inductively from authentic data. (p. 320)

Johns (1997) suggests three steps for a DDL-based lesson:

- *Identification:* Learners are meant to explore from the data what language problem they are to address;

- *Classification:* Learners decide which category or categories of patterns a particular language form represents, and;
- *Generalization:* Learners try to establish patterns and formulate rules on the basis of the data provided to them. (p.101)

One of the greatest advantages of DDL is the exposure to authentic language it ensures (Clifton & Phillips, 2006; Römer, 2008). According to Flowerdew (2015b, 2015d), this approach provides teachers and learners the opportunity to explore naturally occurring language for grammatical patterns, word usage, semantic and pragmatic features, and textual discourse patterns. Römer (2008) suggests that the DDL approach offers authentic examples and encourages "noticing," which helps students enhance their language skills.

Many researchers (Boulton, 2013; Dehghan & Darasawang, 2014; Hunston, 2002) believe that the DDL approach promotes autonomy, as students become less dependent on the instructor, and fosters a learner-centered education environment, in which the emphasis is put on students rather than the instructor or the textbook. In DDL, students are exposed to real language and are expected to discover patterns and rules on their own, while the instructor becomes more of a facilitator, providing assistance when needed. Talai and Fotovatnia (2012) argue that DDL improves learner independence and autonomy, enhances language awareness, and enhance learners' confidence in coping with authentic language.

The literature describes many studies discussing the positive impact of DDL on most areas of language learning (see e.g. Flowerdew, 2015c; Karras, 2016; Lin & Lee, 2015; Luo, 2016; Mehl et al, 2016; Quan, 2016; Tekin, 2015; Vyatkina, 2016).

The following subsections describe the use of parallel corpora in translation training and language teaching and learning. Later, we move on to explore some current projects involving the development of Arabic–English parallel corpora before presenting our own project.

Parallel corpora in language teaching

Interest in using computerized corpora for language teaching and learning started in the mid-1980s and early 1990s, along with growing attention to corpus linguistics in general as a new field. The idea of using corpora in language teaching and learning appeared in many early studies, such as Johns (1986) and Stevens (1991). It has been receiving increasing attention in recent years (see Bennett, 2010; Bernardini, 2016; Boulton, 2010; Cobb & Boulton, 2015; Tribble, 2015).

According to Römer (2006), approaches to using corpora in language teaching and learning can be divided into two types: direct and indirect. The direct approach involves the use of corpora in classrooms to explore and investigate a language in DDL environment, that is, one in which the learners explore the data themselves. This approach is associated with Tim Johns, who introduced the idea of DDL over 20 years ago (see Johns, 1991, 2002). Davies et al. (2013) summarize the DDL approach as involving the following features:

- A focus on the exploitation of authentic materials even when dealing with tasks such as the acquisition of grammatical structures and lexical items.

- A focus on real, exploratory tasks and activities rather than traditional "drill and kill" exercises.
- A focus on learner-centered activities.
- A focus on the use and exploitation of tools rather than ready-made or off-the-shelf learnware (p. 26).

Several researchers have pointed out the positive impact of "corpus-aided discovery learning" on students' language skills (Bernardini, 2002, 2016; Boulton & Tyne,2013; Römer, 2006), especially its ability to enhance autonomy by allowing students to take charge over their own learning. Students can test a wide range of hypotheses using a concordancer, which is a tool designed to analyze corpus data and display words in an authentic context. According to Lee et al (2015), students can develop strategies on this basis to correctly predict the meaning of new vocabulary and examine syntactic patterns in context in an inductive learning environment.

The indirect approach, on the other hand, uses corpora to help syllabus and course designers make decisions about what to teach and when to teach it. Corpora can be investigated with the aim of providing better descriptions of "used" language; other uses of corpora include exploring language students' own corpora to investigate and compare spoken and written language patterns with those found in native-speaker production. Such an approach has pedagogical implications insofar as it helps identify students' problem areas; this in turn allows teachers and course designers to improve teaching materials and strategies (Römer, 2006).

Many studies have explored the positive impact of using parallel corpora in language teaching and learning. In John (2001), a pilot study was conducted to investigate whether a parallel corpus and a concordancer could be used as tools to supplement a beginner-level German-language teaching program in an unsupervised environment. A beginner student of German was asked to find suitable answers for questions on new vocabulary and formulate appropriate grammar rules by himself, using only the parallel corpus and concordancer as tools. The study concluded that these tools can indeed be of great benefit for beginner-level language students.

In another study, Chujo, Anthony, and Oghigian (2009) used a parallel Japanese–English concordancer to examine specific grammar features in a newspaper corpus. The subsequent analysis of learning outcomes showed a positive impact of this approach on learning basic grammar, for instance the basic structures of noun phrases and verb phrases, in addition to answering more complex grammar questions, such as those found on the TOEIC (Test of English for International Communication). Students also reported positive attitudes toward this approach, finding it both novel and appealing and believing it was useful for grammar and vocabulary learning.

Sangawa (2014) presents data on the use of dictionaries among students of Japanese at the University of Ljubljana. A number of resources and a variety of tasks were developed for the study, aimed at enhancing awareness of various aspects of Japanese vocabulary from both monolingual and contrastive perspectives. The researcher used corpora and lexical profiling systems designed to help intermediate and advanced students to obtain more detailed information

about collocational and stylistic aspects of the learned vocabulary. Students used two parallel corpora: (a) jaSlo, a Japanese–Slovene parallel corpus containing Japanese texts with Slovene translations; and (b) Linguee22, a freely available dictionary combined with a search engine that retrieves translated examples from internet-harvested bilingual texts.

In a similar study (but in a different context), Csató et al. (2010) investigated the use of the data-driven learning approach for teaching Turkish as a foreign language in Sweden. The researchers employed language corpora, concordance programs, and annotation tools developed in collaboration with computational linguists for research in teaching environments. The study used a Swedish–Turkish parallel corpus to help students formulate and test hypotheses concerning lexical, morphological, and syntactic aspects of the Turkish language. These tools were found to aid students in contrasting and translating between Swedish and Turkish.

Reynolds (2015) conducted a study on 25 Taiwanese medical students learning English, who were encouraged to investigate the utility of a web-based English–Chinese parallel corpus and collocational concordancer to self-edit their academic writing. Statistical analysis of students' drafts revealed that the use of the concordancer resulted in increased verb–noun collocation accuracy with each draft for both of two essay types: descriptive and opinion. However, qualitative analysis of student feedback regarding their experience showed varied levels in acceptance and success.

In a more recent study, Wong and Lee (2016) used a parallel corpus to teach Cantonese to Mandarin-speaking undergraduate students at the beginner level. Students explored sentence and word alignment in the parallel corpus and independently looked for sentences to discover their translated equivalents. The study results revealed that this data-driven learning approach helped students enhance their knowledge of Cantonese vocabulary, suggesting the potential for applying parallel corpora at even the beginner level for other L1–L2 pairs of closely related languages.

In another study, Wang, Gao, and Hao (2016) describe the construction and application of a customized medical corpus for Chinese clinicians to aid their research article writing in English: CCUT (Customized Corpus for Urology Team). Their study showed that the parallel corpus was useful in assisting Chinese clinicians to choose words with appropriate semantic relations, finding grammatical patterns different from general English in specialized medical contexts, learning how to use unfamiliar medical terms, and revising unidiomatic expressions.

A review of the literature, however, demonstrates a lack of studies that explore the impact of using parallel corpora on Arabic learners or Arab EFL learners. It seems that this potentially useful tool has not yet been explored enough.

Parallel corpora in translation training

The idea of using parallel corpora in translation training emerged in the early 1990s among researchers including Mona Baker (1993, 1995), John Sinclair (1995), and Guy Aston (1999), alongside the increasing popularity of corpus-based studies. Interest in the topic has been increasing steadily over the last decade among translators and researchers (Bernardini, 2004,

2016; Bernardini & Castagnoli, 2008; Frankenberg-Garcia, 2015; Frérot, 2011, 2016; Kübler, 2011; Marco & van Lawick, 2015; Singer, 2016, among others).

Parallel corpora can be applied to theoretical questions including the study of the translation process and how to express an idea in two (or more) different languages, or to comparison of the characteristics of an original (source) text and a translated text. At the same time, the practical uses of parallel corpora in translation studies include different uses of corpora in the development of machine translation (that is, of computer-assisted translation tools such as translation memory systems and terminology management systems) as well as in translation training; the latter is the focus of this section.

According to Gouadec (2002, p. 32), translator training means "training people to perform clearly identified functions in clearly identified environments where they will be using clearly identified tools and 'systems.'" He claims that professional translators should possess the following skills:

1. A full understanding of the material to be translated;
2. The ability to detect, interpret, and cope with cultural gaps;
3. The ability to transfer information, facts, and concepts;
4. The ability to write and rewrite;
5. The ability to proofread; and
6. The ability to control and assess quality.

In order to master these skills, translators should know how to obtain the required information and knowledge as well as the appropriate terminology and phraseology.

Parallel corpora are a valuable tool that can be incorporated into translation training programs so that students can employ them in translation projects (Singer, 2016). It has recently been observed that many translation instructors encourage their students to compile their own specialized corpora according to the types of translation projects they are working on (legal, medical, scientific, etc.). Instructors are also becoming keener to provide their students with necessary skills related to corpus design issues, corpus types, corpus analysis, and search tools. According to Molés-Cases and Oster (2015), "corpora are of crucial importance in translator education, because they promote autonomy, motivation and authenticity" (p. 3). Parallel corpora enhance students' awareness of differences between original texts and their translations and of how to transfer an idea or expression across two (or more) languages. Students can also explore parallel corpora to investigate collocations, fixed expressions, and idiomatic or common structures in both source and translated languages. Examples obtained from parallel corpora can help with the verification of translation decisions derived from other sources (e.g., dictionaries, glossaries, and lexicons) or from intuition and can assist in the selection of appropriate translations according to the context.

According to Pearson (2003), parallel corpora also provide the opportunity for students to compare their work with the work of professional translators and identify the various approaches and strategies those translators have adopted and implemented. Parallel corpora also increase students' awareness of the distinctive nature of different specialized texts (legal, medical, etc.) as

well as their ability to examine the special terminology, syntax, and semantic features associated with those texts (see Kübler, 2011).

The literature includes many studies showing positive impacts of parallel corpora on translators' skills and supporting the integration of such resources into students' training programs (Pearson, 2003; Bernardini, 2016). In a study conducted by Frankenberg-Garcia and Santos (2003), a Portuguese–English parallel corpus (Compara) was used to design exercises to assist translation students in identifying and handling the contrastive features of the two languages, features which can lead to translation problems. The researchers highlighted the positive impact of this approach on the training of translators; they concluded their study by encouraging translation curriculum designers to integrate parallel corpora as a teaching resource.

Parallel corpora are often used to investigate the nature of translated texts by comparing them with the source texts. However, many researchers have also recently become interested in using parallel corpora to study the original (source) texts themselves. For instance, in a recent study by Salhi (2013), the English-Arabic Parallel Corpus of United Nations Texts (EAPCOUNT) was used to study complementary polysemy and Arabic translations of the English noun *destruction*. Salhi emphasized the benefits of using parallel corpora to help students detect the semantic and syntactic features of an original text as well as a translated text.

In another study, Rodríguez-Inés (2014) presented a number of exercises for use in the training of Spanish–English translators, based on the exploration of a Spanish–English parallel corpus. Rodríguez-Inés stressed the significance of this source in raising students' awareness of collocations, contrastive features, fixed expressions, and specialized terminology. An end-of-study survey indicated an increase in students' confidence in identifying collocations and specialized terminology after consulting the parallel corpus.

In a similar study, Li and Dai (2014) used an English–Chinese parallel corpus in the training of 90 translation students at a Chinese university. Students were divided into two groups: The first group was taught in a traditional way, while the second group used training exercises that required the exploration of the corpus. A pre- and post-test were conducted that showed a significant improvement in the second group's translation skills, which the researcher attributed to the use of the parallel corpus.

Heylen and Verplaetse (2015) recently investigated the use of parallel corpora to train students in a medical translation course and analyzed its impact on their performance. Students were divided into two groups and asked to translate medical bulletins from English to German; only one group used the parallel corpus. The results showed that that group performed better than the other group.

The literature discussed above highlights the value of integrating parallel corpora into translator training and their positive impact on translators' performance. However, there seems to be a lack of studies that investigate this matter in the Arab world and among Arabic translators. This may be attributable largely to a lack of such language resources (see Al-Sulaiti & Atwell, 2006; Al-Ajmi, 2011).

The next section attempts to shed light on some of the current projects involving the development of Arabic–English parallel corpora.

Arabic–English parallel corpora

The past few years have borne witness to a growing interest in parallel corpora among Arab researchers and to increased awareness of their significance. However, the literature describes relatively few Arabic–English parallel corpora, indicating a corresponding lack of such resources. According to Al-Ajmi (2004, p. 327), "this is partly due to the lack of the necessary programs to compile such resources and the funding authorities' doubts and uncertainty regarding the effectiveness of parallel corpora."

The English–Arabic Parallel Corpus of the United Nations Texts (EAPCOUNT) is one of the most well-known English–Arabic corpora projects, containing 341 texts aligned on a paragraph level. The 5,392,491-word corpus was compiled using two sub-corpora: The first contains the English originals, and the second contains their Arabic translations. The texts mainly include resolutions and annual reports issued by different UN organizations and institutions, along with some texts taken from publications of other international institutions (Salhi, 2013).

In a similar project, the European Commission funded the development of a multilingual parallel corpus, MultiUN, at the Language Technology Lab in Germany. The corpus includes 300 million words extracted from UN documents published between 2000 and 2009 on the official UN website (see Eisele & Chen, 2010).

Tiedemann (2012) developed the Open Parallel Corpus (OPUS): a free, multilingual parallel corpus containing translated texts collected from the web. OPUS also provides open-source tools for processing parallel and monolingual data as well as several interfaces for searching the data to help with various research activities. According to its website, all pre-processing was done automatically, suggesting that no manual corrections were made.

EuroMatrix is another multilingual parallel corpus, funded by the European Union. The corpus contains the proceedings of the European Parliament translated into Arabic and many other languages. The corpus includes 51 million words, 1.5 million of which are Arabic. The project's aims involve developing and promoting machine translation systems.

A project carried out in the Arab world, at Kuwait University, has developed a parallel corpus that includes Arabic translations taken from the *World of Knowledge* book series published by the National Council for Culture, Arts and Letters (NCCAL) in Kuwait. The corpus contains 3 million words and is only available to the university's staff and students, mainly for lexicography and translation courses (Al-Ajmi, 2011).

The Linguistic Data Consortium (LDC) has been involved in many parallel corpora projects, including in Arabic. One such project is the GALE Phase 2 Arabic Broadcast News Parallel Text. This corpus contains modern standard Arabic source texts and corresponding English translations selected from broadcast news data collected by the LDC between 2005 and 2007 and transcribed by the LDC or under its direction. The corpus consists of 60 source–

translation document pairs, amounting to 42,089 words of Arabic source text and their English translations. (See LDC, 2013.)

The LDC also developed Arabic–English Automatically Extracted Parallel Text. These texts were extracted automatically from two monolingual corpora: Arabic Gigaword Second Edition (LDC2006T02) and English Gigaword Second Edition (LDC2005T12). The data consist of news articles published by the Xinhua News Agency (in Chinese) and Agence France-Presse (in French). The corpus consists of 1,124,609 sentence pairs; the word count on the English side is approximately 31 million words. (See LDC, 2013.)

In another project, Izwaini (2003) developed a multilingual corpus at UMIST. This specialized corpus involves information technology texts in English and two translational corpora: Arabic (1 million tokens) and Swedish (2.7 million tokens). The texts mainly consist of manuals and online help text for computer systems, hardware, and software, as well as material from multilingual IT-specialized websites. This corpus is not available for public use, and copyright permission is obtainable only for research investigations.

AMARA is a recent project implemented by the Qatar Computing Research Institute (Abdelali et al., 2014; Guzman et al., 2013). The corpus contains community-generated video subtitles from well-known educational platforms, such as Technology, Entertainment, and Design (TED) and the Khan Academy. It consists of 2.6 million Arabic words and 3.9 million English words. The researchers' aim was to prepare data for machine translation tasks. The project also offers an editor tool for subtitling and captioning (see www.amara.org).

Also recently, Alkahtani and Teahan (2015) compiled a parallel corpus consisting of 27.8 million Arabic words and 30.8 million English words collected from two sources: the *Al-Hayat* newspaper and the OPUS corpus. The project aims to promote research in the field of machine translation.

Hassan and Atwell (2016) have recently designed a multilingual special corpus for the Hadith, the words of the Prophet Mohammed (Peace Be Upon Him). The corpus consists of around 2 million words of Hadith in Arabic and their English, French, and Russian translations.

Table 1 provides a summary of Arabic–English parallel corpora.

Table 1
A Summary of Arabic–English Parallel Corpora

Corpus	Organization	Size	Purpose	Content	Availability
The English–Arabic Parallel Corpus of United Nations Texts (EAPCOUNT)	Carthage University	5,392,491 words	A research tool for applied and theoretical linguistic research	Resolutions and annual reports issued by different UN organizations and institutions	Available through: http://conferences.unite.un.org/UNCorpus
The Open Parallel Corpus	Uppsala University		To support research in	Web-collected texts	Available through:

(OPUS)			machine translation		http://opus.lingfil.uu.se/index.php
MultiUN	Language Technology Lab at DFKI, Germany	300 million words	To support research in machine translation	United Nations' websites	Available through: http://www.euromatrixplus.net/multi-un/
European Union corpus (EuroMatrix)	European Union	51 million words in total, with 1.5 million Arabic words	To support research in machine translation	Proceedings of the European Parliament	Available through: http://www.euromatrix.net/
Kuwait University English–Arabic Parallel Corpus	Kuwait University	3 million words	Teaching, translation, and lexicography	Publications from Kuwait National Council	Only available for Kuwait University staff and students
AMARA	Qatar Computing Research Institute	2.6 million Arabic words; 3.9 million English words	To support research in machine translation	Subtitles from educational videos	Available through: https://amara.org/en/
Arabic–English parallel corpus	Bangor University	27.8 million Arabic words; 30.8 million English words	To support and test machine translation systems	Translated texts from *Al-Hayat* newspaper and OPUS	Artificial Intelligence and Intelligent Agents research group, w.j.teahan@bangor.ac.uk
Hadith Corpus	Leeds University	2 million words	Religious studies	Hadith of Sahih Albukhari	Not yet available

Research Objective

Many of the Arabic–English parallel corpora discussed in the previous section are inaccurate and/or expensive. Some are small in size, while others are restricted in terms of genre, failing to meet the requirements of academics and researchers. Therefore, there is a great need for a large Arabic–English parallel corpus that takes into account the quality of source- and target-text materials and covers a wide range of text types.

The ongoing project described in this paper is underway at the College of Languages & Translation, King Saud University; it aims to compile a 10-million-word Arabic–English parallel corpus to be used as a resource for translation training and language teaching. The corpus, a work in progress, is bidirectional and can be used to compare the features of translated and source language and identify differences. To enhance its quality, the corpus has been manually verified at different stages, including translation, text segmentation, alignment, and file preparation. The corpus is available as full-text in XML format through a user-friendly web

interface, which includes a concordancer that supports bilingual search queries and several filtering options.

The following section describes the development of this bilingual Arabic–English parallel corpus (AEPC) and discusses the design criteria and design stages.

The Design of the AEPC

Reviewing the literature revealed the great need for an Arabic–English parallel corpus with high-quality source- and target-text materials and high accuracy of alignment. The aim of this project is to create a useful resource for language teaching and translation training. Therefore, the corpus needs to take into account the following factors (Biber, 2003).

Size

There seems to be some debate about how large a corpus should be (see Krishnamurthy, 2001; Leech, 1991). However, most or all agree that the size of the corpus depends on the purpose for which it is intended and on several practical factors, such as copyright permissions and availability.

Many of the previously discussed Arabic–English corpora are limited in size, several to around 1–3 million words, which restricts their effectiveness. Therefore, the first phase of this project involved collecting 10 million words; this number will increase in later stages of the project.

Representativeness

According to Biber (1993), *representativeness* indicates "the extent to which a sample includes the full range of variability in a population" (p. 243). Two main factors affect the representativeness of a corpus: *balance*, which refers to the variety of genres included in a corpus, and *sampling*, which refers to how the chunks of text for each genre are selected. Most of the Arabic–English corpora discussed in the literature tend to be restricted in terms of genre and text types, which negatively affects their usefulness as language teaching and/or translation training resources. They also tend to be static in nature, which decreases their representativeness increasingly over time; as Hunston (2002, p. 30) argues, "any corpus that is not regularly updated rapidly becomes unrepresentative." Hence, the proposed corpus aims to cover a wide range of genres and text types and to remain open and conduct regular updates to maintain its dynamic and representative nature.

Quality

Since this corpus is intended for instructional purposes, quality is of central importance. Here, human-translated text samples have been compiled, cleaned, and aligned manually to ensure high levels of accuracy.

Availability

Many of the Arabic–English parallel corpora described in the previous section are expensive (e.g., LDC products) or have a non-user-friendly interface (e.g., OPUS), which makes them hard to use for non-experts. Our project aims to attract a wide range of users by providing a free, user-friendly website and easy-to-use search tools.

The project went through two phases: The first phase involved compiling, cleaning, and aligning high-quality human-translated text samples of various text types, as noted above. The second phase involved designing a web interface with a bilingual concordancer, where users can explore the content of the AEPC in both English and Arabic. Both phases will be detailed in the next sections.

- **Implementing the AEPC: Phase 1**

The project began in 2015, by compiling, cleaning, and aligning the samples. The first phase involved preparing the files and segmenting, aligning, and verifying the texts.

Preparing the files

Texts were collected from several sources: printed material, such as books, magazines, and newspapers; websites; and translation graduation projects. OCR software (ABBYY FineReader and Readiris 15) was used to convert printed documents to machine-readable texts, as both programs support both Arabic and English. The texts were categorized into the following eight genres: Social, Biographical, Literary, Administrative, Medical, Legal, Religious, and Scientific. The texts were segmented on a sentence level, aligned manually, and then stored as MS Excel files. Each file includes the following metadata in both languages: Title, Author, Publisher, Year of Publication, Country, Author's Gender, Medium, Domain, and Topic. The metadata will enable corpus users to select texts according to their specific requirements.

Segmentation

The next step in building the parallel corpus was segmentation. The texts were divided into short segments on a sentence level. There are applications available today that can carry out this process automatically, such as translation memory software (e.g., WinAlign, Memsource). Such applications are considered to be language independent, and can manage various types of text files. However, the results are not error-free; this is because, most of the time, it is difficult for such applications to determine what a sentence is. Punctuation signs, such as periods, exclamation marks, and question marks, are commonly used to indicate the end of a sentence, but might be challenging for applications to identify. The period, for example, might not necessarily indicate the end of a sentence but instead a decimal or an abbreviation such as Dr. or Ms. To address this issue, segmentation was done manually in order to ensure the accuracy and quality of the processed texts.

Alignment

Quah (2006, p. 100) defined alignment as "the process of binding a source-language segment to its corresponding target-language segment." This process can be performed at various levels: word, sentence, paragraph, or text; however, most parallel corpora align texts at the sentence or paragraph level.

Aligning source texts and translations was a challenging process, as translators do not necessarily translate texts in a predictable or linear manner. Frankenberg-Garcia and Santos (2003) noticed that translators often split source-text sentences into two or more translated sentences; join two or more source-text sentences together, rendering them as a single translated sentence; leave things out; reorder sentences in different ways; and/or insert elements that were not present in the source text.

For this project, it was decided to carry out the alignment process manually. Manual alignment entails going through the text sequentially and linking the first sentence in the source text with the first sentence in the target text, and so forth. Automatic alignment manages completed translations by splitting source and target texts into segments and linking segments that belong together. There are some alignment tools available on the market that allow users to verify the accuracy of automatic alignment and manually edit and realign mismatched segments. Unfortunately, many of these tools cannot support languages (such as Arabic) that are not based on the Roman alphabet, and fail to produce accurate results in particular when the source and target language have different structures and text directions, like Arabic and English.

- **Implementing the AEPC: Phase 2**

The second phase involved allowing translators and language learners and instructors to freely explore the content of the corpus via a web interface designed to be user friendly. The web interface also offers a bilingual concordancer, a search tool for use with a parallel corpus (Bowker, 2002) that retrieves all occurrences of a specific search word within a given context in both source and target languages; users can also refine their search according to criteria like, in the present case, domain, year of publication, country, topic, medium, or author's gender. The beta version of the website is currently available at http://aeparallelcorpus.net/.

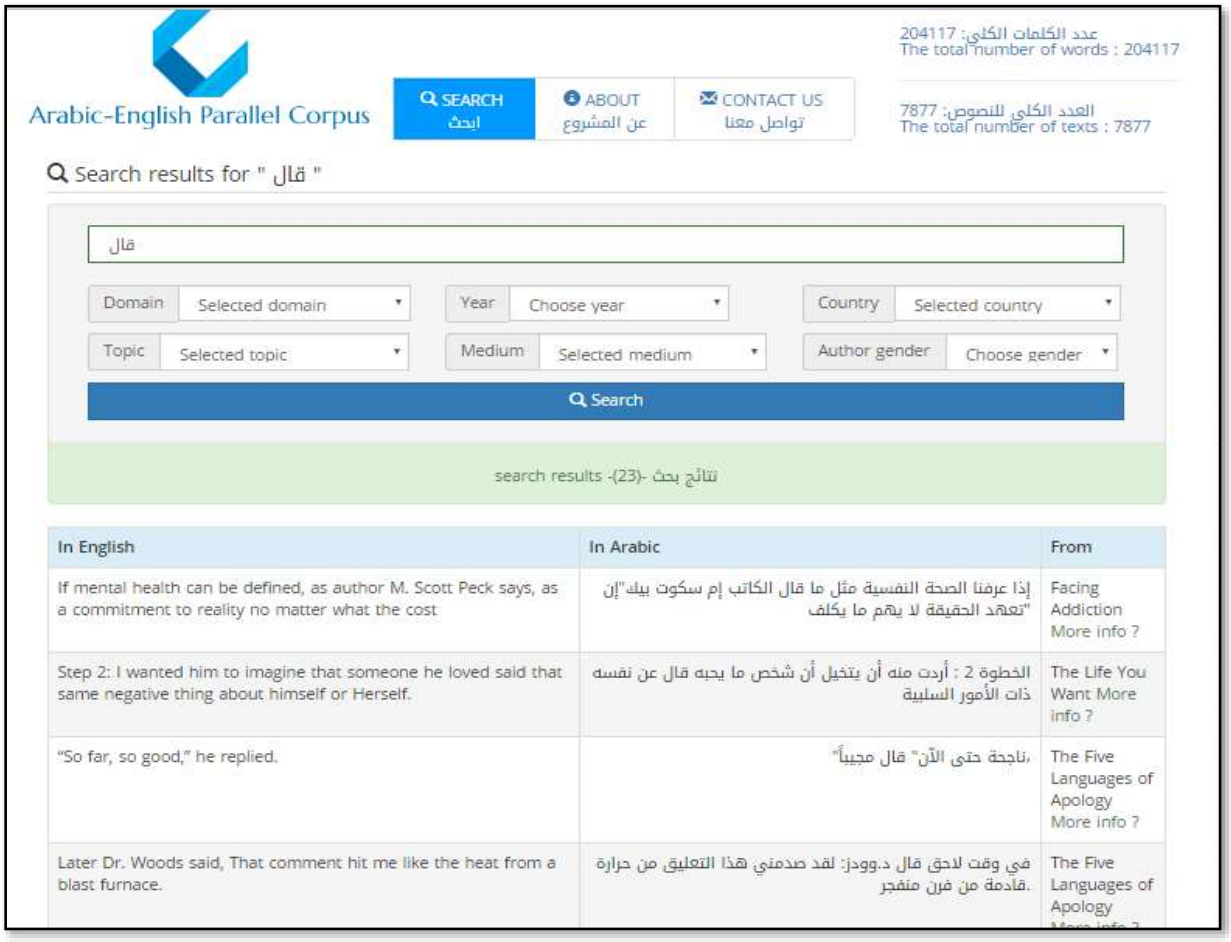

Figure 1. Screenshot of AEPC interface (beta version).

- **Next Steps: Phase 3**

The last phase of the project will begin in 2017 in collaboration with the College of Computer and Information at King Saud University. It will involve the development of more sophisticated tools for use with the data, such as a collocation search tool, graphical statistical analysis functions, and more filtering options to display more targeted results. This stage will also include examination of part-of-speech taggers to annotate corpus texts, which will contribute to improved corpus search results and enhance the corpus's efficiency. For details about some well-known software taggers available today, see Khoja (2001) and Pasha et al. (2014).

Conclusion

This paper has described an ongoing project at the College of Languages & Translation, King Saud University, to compile a 10-million-word Arabic–English parallel corpus to be used as a resource for translation training and language teaching. The bidirectional corpus can be used to compare and study the differences between translated and source languages. The corpus has been manually verified at different stages, including translations, text segmentation, alignment, and file preparation, to enhance its quality. The corpus is available in XML format and through a user-friendly web interface. The web interface includes a concordancer which supports bilingual search queries and several filtering options. This parallel corpus is intended to be used as a teaching resource in language teaching and translator training classrooms.

Acknowledgements

The author would like to thank the Research Center for the Humanities, Deanship of Scientific Research, King Saud University, for funding this project.

About the Author

Dr. Alotaibi is the former Vice-Dean of College of Languages & Translation, King Saud University. She has a PhD in Education in Computer-Assisted Language Learning from The University of Manchester. Her research interests include ICT in Education, distance learning, Mobile- Assisted Learning and Computer Applications in Translation. She is a member of iWAN and BCI in L2 research groups at KSU and currently working on developing several educational applications for language and translation students.

References

Abdelali, A., Guzman, F., Sajjad, H., & Vogel, S. (2014, May). The AMARA Corpus: Building Parallel Language Resources for the Educational Domain. In *LREC* (Vol. 14, pp. 1044-1054).

Aijmer, K. (Ed.). (2009). *Corpora and language teaching* (Vol. 33). Amsterdam and Philadelphia: John Benjamins Publishing.

Al-Ajmi, H. (2003). Compiling an English–Arabic parallel text corpus. In M. Murata, S. Yamada, & Y. Tono (Eds.), *Proceedings of Asian Association for Lexicography* (pp. 51–54). Tokyo: Asialex.

Al-Ajmi, H. (2004). A new English–Arabic parallel text corpus for lexicographic applications. *Lexikos, 14*(1), 326–330.

Al-Ajmi, H. (2011). A New English Arabic Parallel Text Corpus for Lexicographic Applications. *Lexikos, 14*.

Alkahtani, S., & Teahan, W. J. (2015, December). A new parallel corpus of Arabic/English. In *Proceedings of the 8th Saudi Students Conference in the UK* (p. 279). Singapore: World Scientific.

Al-Sulaiti, L., & Atwell, E. S. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, *11*(2), 135-171.

Aston, G. (1999). Corpus use and learning to translate. *Textus*, *12*(2), 289–314.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). Amsterdam and Philadelphia: John Benjamins.

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

Baker, M. (1998). *Routledge encyclopedia of translation studies*. London: Routledge.

Baker, M. (1999). The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics*, *4*(2), 281–298.

Barlow, M. (1996). Parallel texts in language teaching. In S. Botley, J. Glass, T. McEnery, & A. Wilson (Eds.), *Proceedings of Teaching and Language Corpora 96* (UCREL Technical Papers, Vol. 9) (pp. 45–56). Lancaster: UCREL.

Bennett, G. (2010). *Using corpora in the language learning classroom: Corpus linguistics for teachers*. Ann Arbor: University of Michigan Press.

Bernardini, S. (2002). Exploring new directions for discovery learning. *Language and Computers*, *42*(1), 165-182.

Bernardini, S. (2004). Corpus-aided language pedagogy for translation education. In K. Malmkjær (Ed.), *Translation in undergraduate degree programmes* (pp. 97–111). Amsterdam and Philadelphia: John Benjamins Publishing.

Bernardini, S. (2016). Discovery learning in the language-for-translation classroom: corpora as learning aids. *Cadernos de Tradução*, *36*(1), 14–35.

Bernardini, S., & Castagnoli, S. (2008). Corpora for translator education and translation practice. In E. Yuste-Rodrigo (Ed.), *Topics in language resources for translation and localization* (pp. 39–55). Amsterdam and Philadelphia: Benjamins Publishing.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Biber, D. (2003).Representativeness in corpus design in: Literary and Linguistic Computing, 8/4: 243-57.

Boulton, A. (2011).Data-driven learning: the perpetual enigma.. S. Goźdź-Roszkowski. *Explorations across Languages and Corpora*, Peter Lang, pp.563-580.

Boulton, A. (2013). Separating fact from fiction: The real story of corpus use in language teaching. In L. Bradley & S. Thouesny (Eds.), *20 years of EUROCALL: Learning from the past, looking to the future: Proceedings of the 2013 EUROCALL Conference* (pp. 51–56). Dublin: Research-publishing.net

Boulton, A., & Tyne, H. (2013). Corpus linguistics and data-driven learning: A critical overview. *Bulletin suisse de Linguistique appliquée*, *97*, 97–118.

Botley, S., Glass, J., McEnery, A., & Wilson, A. (1996). *Proceedings of Teaching and Language Corpora 96* (UCREL Technical Papers, Vol. 9). Lancaster: UCREL.

Bowker, L. (2002). *Computer-aided translation technology: A practical introduction*. University of Ottawa Press.

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

Chujo, K., Anthony, L., & Oghigian, K. (2009). DDL for the EFL classroom: Effective uses of a Japanese–English parallel corpus and the development of a learner-friendly, online parallel concordancer. In *Proceedings of the Corpus Linguistics Conference, CL2009.* Liverpool: University of Liverpool.

Clifton, J., & Phillips, D. (2006). Ensuring high surrender value for corporate clients and increasing the authority of the language instructor: The dividends of a data-driven lexical approach to ESP. *The Journal of Language for International Business*, *17*(2), 72–81.

Cobb, T., & Boulton, A. (2015). Classroom applications of corpus analysis. In D. Biber & R. Reppen (Eds.), *Cambridge handbook of corpus linguistics* (pp. 478–497). Cambridge: Cambridge University Press.

Csató, É. Á., Kilimci, S., & Megyesi, B. (2010). Using Parallel Corpora in Data-Driven Teaching of Turkish in Sweden. *Psychological Issues*, *385*, 330.

Davies, G., Otto, S., & Rüschoff, B. (2013). Historical perspectives on CALL. In M. Thomas, H. Reinders, & M. Warschauer (Eds.), *Contemporary computer-assisted language learning* (pp. 19–38). London: Bloomsbury.

Dehghan, A. & Darasawang, P. (2014) Independent learning through the use of data driven learning. In *Proceedings of the International Conference: DRAL 2*. Bangkok: Independent Learning Association.

Eisele, A., & Chen, Y. (2010). MultiUN: A multilingual corpus from United Nation documents. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)* (pp. 2868–2872). Luxembourg: European Language Resources Association.

Flowerdew, L. (2015a). Data-driven learning and language learning theories: Whither the twain shall meet. In A. Leńko-Szymańska, & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 15–36). Amsterdam: John Benjamins.

Flowerdew, L. (2015b). Corpus-based research and pedagogy in EAP: From lexis to genre. Lang*uage Teaching*, *48*(1), 99–116.

Flowerdew, L. (2015c). Using corpus-based research and online academic corpora to inform writing of the discussion section of a thesis. *Journal of English for Academic Purposes*, *20*, 58–68.

Flowerdew, L. (2015 d). Data-Driven learning and language learning theories. *Mult Afford Lang Corpora Data-Driven Learn*, *69*, 15-36.

Frankenberg-Garcia, A. (2015). Training translators to use corpora hands-on: Challenges and reactions by a group of thirteen students at a UK university. *Corpora*, *10*(3), 351–380.

Frankenberg-Garcia, A., & Santos, D. (2003). Introducing Compara: The Portuguese–English parallel corpus. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 71–87). Manchester: St. Jerome.

Frérot, C. (2011). Parallel corpora for translation teaching and translator training purposes. In Góźdź-Roszkowski (Ed.), *Explorations across languages and corpora* (pp. 433–450). Frankfurt: Peter Lang.

Frérot, C. (2016). Corpora and corpus technology for translation purposes in professional and academic environments. Major achievements and new perspectives. *Cadernos de Tradução*, *36*(1), 36–61.

Gouadec, D. (2002). Training translators: Certainties, uncertainties, dilemmas. *Training the language services provider for the new millennium* (pp. 31–41). Oporto: University of Porto.

Granger, S., & Lefer, M. A. (2016). From general to learners' bilingual dictionaries: Towards a more effective fulfillment of advanced learners' phraseological needs. *International Journal of Lexicography*, *1*(10), 1–17.

Guzman, F., Sajjad, H., Abdelali, A., & Vogel, S. (2013). The AMARA corpus: Building resources for translating the web's educational content. In *Proceedings of the International Workshop on Spoken Language Translation, IWSLT 2013*. Heidelberg: IWSLT.

Hassan, S., & Atwell, E. S. (2016). Design and implementing of multilingual Hadith corpus. *International Journal of Recent Research in Social Sciences and Humanities*, *3*(2), 100–104.

Heylen, K., & Verplaetse, H. (2015). Parallel corpora for medical translation training: An analysis of impact on student performance. In *Proceedings of the 4th International Conference on Corpus Use and Learning to Translate (CULT)*. Alicante, Spain: Universitat d'Alacant.

Hu, K. (2016). Corpus-based study of translation teaching. In *Introducing Corpus-based Translation Studies* (pp. 177–191). Berlin/Heidelberg: Springer.

Izwaini, S. (2003). A corpus-based study of metaphor in information technology. In *Proceedings of the workshop on corpus-based approaches to figurative language, corpus linguistics* (pp. 1-8).

Johansson, S. (1999). Corpora and contrastive studies. *Multiple Languages–Multiple Perspectives. AF in LA Yearbook, 57*, 116–125.

Johansson, S. (2007). *Seeing through multilingual corpora: On the use of corpora in contrastive studies*. Amsterdam and Philadelphia: John Benjamins.

John, E. S. (2001). A case for using a parallel corpus and concordancer for beginners of a foreign language. *Language Learning & Technology*, *5*(3), 185–203.

Johns, T. (1986). Micro-concord: A language learner's research tool. *System*, *14*(2), 151–162.

Johns, T. (1993). Data-driven learning: An update. *TELL & CALL*, *3*, 23–32.

Johns, T. (1997). Contexts: the background, development and trialling of a concordance-based CALL programme. In A. Wichmann, S. Fligelstone, T. McEnery, & G. Knowles (Eds.), *Teaching and learning corpora* (pp. 100–115). Harlow, UK: Addison-Wesley Longman.

Johns, T. (2002). Data-driven learning: The perpetual challenge. In B. Kettermann & G. Marko (Eds.), *Teaching and learning by doing corpus linguistics* (pp. 107–117). Amsterdam: Rodopi.

Karras, J. N. (2016). The effects of data-driven learning upon vocabulary acquisition for secondary international school students in Vietnam. *ReCALL*, *28*(02), 166–186.

Khoja, S. (2001, June). APT: Arabic part-of-speech tagger. In *Proceedings of the Student Workshop at NAACL* (pp. 20–25). San Diego: NAACL.

Kilgarriff, A. (2013). Terminology finding, parallel corpora and bilingual word sketches in the Sketch Engine. In *Proceedings of ASLIB 35th Translating and the Computer Conference*. London: ASLIB.

Krishnamurthy, R. (2000). Size matters: Creating dictionaries from the world's largest corpus. In *Proceedings of KOTESOL 2000 Casting the Net: Diversity in Language Learning*. Daegu: KOTESOL.

Kübler, N. (2011). Working with different corpora in translation teaching. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston, *New trends in corpora and language learning* (pp. 62–80). London: Continuum.

Lee, J. H., Lee, H., & Sert, C. (2015). A corpus approach for autonomous teachers and learners: Implementing an on-line concordancer on teachers' laptops. *Language Learning & Technology*, *19*(2), 1–15.

Leech, G. N. (2010). Corpus linguistics. In K. Malmkjaer (Ed.), *The Routledge linguistics encyclopedia* (3rd ed.) (pp. 103–113). London: Routledge.

Li, H., & Dai, Z. (2014). Effectiveness of self-built Chinese–English corpus on assisting translation teaching. *International Journal of Humanities and Social Science*, *4*(7), 96–98.

Lin, M. H., & Lee, J.-Y. (2015) Data-driven learning: Changing the teaching of grammar in EFL classes. *ELT Journal*, 69(3), 264–274.

Linguistic Data Consortium (LDC). (2013). LDC catalog. [online]. Retrieved from: http://catalog.ldc.upenn.edu

Luo, Q. (2016). The effects of data-driven learning activities on EFL learners' writing development. *SpringerPlus*, *5*(1), 12–55.

Marco, J., & van Lawick, H. (2015). Enhancing translator trainees' awareness of source text interference through use of comparable corpora. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 225–244). Amsterdam: John Benjamins

Mehl, S., Wallis, S., & Aarts, B. (2016). Language learning at your fingertips: Deploying corpora in mobile teaching apps. In *Creating and digitizing language corpora* (pp. 211–239). London: Palgrave Macmillan UK.

Molés-Cases, T. & Oster, U. (2015). Webquests in translator training: Introducing corpus- based tasks. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 201–224). Amsterdam: John Benjamins.

Odlin, T. (1994). *Perspectives on pedagogical grammar*. Cambridge: Cambridge University Press

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

Pasha, A., Al-Badrashiny, M., Kholy, A. E., Eskander, R., Diab, M., Habash, N., … Roth, R. (2014). Madamira: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic. In *Proceedings of the 9th International Conference on Language Resources and Evaluation*. Reykjavik: European Language Research Association.

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

Perez, M. M., Paulussen, H., Macken, L., & Desmet, P. (2014). From input to output: The potential of parallel corpora for CALL. *Language Resources and Evaluation*, *48*(1), 165–189.

Quan, Z. (2016). Introducing "mobile DDL (data-driven learning)" for vocabulary learning: an experiment for academic English. *Journal of Computers in Education*, *3*(3), 273–287.

Rauf, S. A., & Schwenk, H. (2011). Parallel sentence generation from comparable corpora for improved SMT. *Machine translation*, *25*(4), 341-375.

Reynolds, B. L. (2015). Action research: Applying a bilingual parallel corpus collocational concordancer to Taiwanese medical school EFL academic writing. *RELC Journal*, *47*(2), 213–227.

Rodríguez-Inés, P. (2014). Using corpora for awareness-raising purposes in translation, especially into a foreign language (Spanish–English). *Perspectives*, *22*(2), 222–241.

Römer, U. (2006). Pedagogical applications of corpora: Some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik*, *54*(2), 121–134.

Römer, U. (2008). Corpora and language teaching. In A. Ludeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook* (Vol. 1) (pp. 112–130). Berlin: Mouton de Gruyter.

Salhi, H. (2013). Investigating the complementary polysemy and the Arabic translations of the noun Destruction in EAPCOUNT. *Meta: Translators' Journal*, *58*(1), 227–246.

Sangawa, K. H. (2014). The learner as lexicographer: Using monolingual and bilingual corpora to deepen vocabulary knowledge. *Acta Linguistica Asiatica*, *4*(2), 53–65.

Sharoff, S., Rapp, R., Zweigenbaum, P., & Fung, P. (Eds.). (2013). *Building and using comparable corpora*. Berlin: Springer.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford: Oxford University Press.

Sinclair, J. (1995). Corpus typology: A framework for classification. *Stockholm Studies in English*, *85*, 17–33.

Sinclair, J. M. (Ed.). (2004). *How to use corpora in language teaching*. Amsterdam: John Benjamins Publishing.

Singer, N. (2016). A proposal for language teaching in translator training programmes using data-driven learning in a task-based approach. *International Journal of English Language & Translation Studies*, *4*(2), 155–167.

Stevens, V. (1991). Classroom concordancing: Vocabulary materials derived from relevant, authentic text. *English for Special Purposes Journal*, *10*, 35–46.

Talai, T., & Fotovatnia, Z. (2012). Data-driven learning: A student-centered technique for language learning. *Theory and Practice in Language Studies*, *2*(7), 15–26.

Tekin, B. (2015). Data-driven vocabulary learning vs. traditional instruction at a high school in Uganda. *Journal of Education*, *4*(1), 79–85.

Teubert, W. (2002). The role of parallel corpora in translation and multilingual lexicography. In B. Altenberg & S. Granger (Eds.), *Lexis in contrast* (pp. 189–214). Amsterdam: John Benjamins.

Tian, L., Wong, D. F., Chao, L. S., Quaresma, P., Oliveira, F., & Yi, L. (2014). UM-Corpus: A Large English-Chinese Parallel Corpus for Statistical Machine Translation. In *LREC* (pp. 1837-1842).

Tiedemann, J. (2012). Parallel data, tools and interfaces in OPUS. In *Proceedings of the 8th International Conference on Language Resources and Evaluation (LREC '12)* (pp. 2214–2218). Istanbul: European Language Research Association.

Tribble, C. (2015). Teaching and language corpora: Perspectives from a personal journey. In A. Leńko-Szymańska & A. Boulton (Eds.), *Multiple affordances of language corpora for data-driven learning* (pp. 37–62). Amsterdam: John Benjamins.

Vyatkina, N. (2016). Data-driven learning for beginners: The case of German verb-preposition collocations. *ReCALL*, *28*(2), 207–226.

Wang, X., Gao, Y., & Hao, T. (2016). The construction of a customized medical corpus for assisting Chinese clinicians in English research article writing. In *China National Conference on Chinese Computational Linguistics* (pp. 241–252). Berlin: Springer.

Wichmann, A., & Fligelstone, S. (2014). *Teaching and language corpora*. London: Routledge.

Wong, T. S., & Lee, J. S. (2016). Corpus-based learning of Cantonese for Mandarin speakers. *ReCALL*, *28*(2), 187–206.

Xie, Q. (2015). Recent developments in corpus linguistics and corpus-based research/Department of Linguistics and Modern Language Studies at the Hong Kong Institute of Education. *Language Teaching*, *48*(1), 156–160.

Zanettin, F., Bernardini, S., & Stewart, D. (2014). *Corpora in translator education*. Routledge.