International Journal of Arabic-English Studies (IJAES) *Vol.24, No.1, 2024*
DOI: https://doi.org/10.33806/ijaes.v24i1.552

Exploitation and Evaluation of an Arabic-English Composite Learner Translator Corpus

Reem F. Alfuraih and Noha M. El-Jasser
Princess Nourah bint Abdulrahman University, Saudi Arabia

Received: 3.6.2022 Accepted: 23.7.2023 Published: 2.1.2024

Abstract: This paper describes in depth the data collection and exploitation stages in constructing the undergraduate learner translator corpus (ULTC), a 75 million-word sentence-aligned bidirectional parallel corpus of Arabic, English, and French, with Arabic as its central language. We focus on the methodological challenges, and describe the compilation process and problems encountered in the first phase of the project. Our aim is to inform future compilers of similar projects that integrate learner corpus research (LCR) and corpus-based translation studies (CBTS). In the first part, we present design considerations, data collection criteria, and the exploitation of the corpus, and in the second part, we evaluate the systems we used and possible improvements.

Keywords: Arabic multimodal parallel learner corpus, process and product-oriented translation, triangulation

1. Introduction

Few learner translator corpora are available. Therefore, a large-scale member of this corpus family is needed, especially one that is focused on widely used but resource-poor language like Arabic. The undergraduate learner translator corpus (ULTC) is a 75 million-word sentence-aligned parallel corpus of English, Arabic, and French, with Arabic being the central language in the corpus resource. It is the first publicly accessible composite corpus of its kind comprising parallel textual and multimodal translations produced by learners of translation from English or French into Arabic. The ULTC is a combined resource, of which most are learner translator corpora, and some are reference corpora. As of May 2018, the ULTC has been accessible for users via https://arabicparallelultc.com/

The main objective of this paper is to reflect critically on the design of the ULTC, and the methodological challenges, solutions, and implications of the methodological choices during the first stages that were carried out from 2014 to 2018. Moreover, this paper aims to inspire and raise awareness among future, large parallel corpora compliers in learner corpus research (LCR) and corpus-based translation studies (CBTS). This paper also accounts for the theoretical and methodological integration between LCR and CBTS. According to Mikhailov and Cooper, the ULTC "does not take methodological issues for granted, but gives practical advice on how to approach different research problems" (2016).

Alfuraih (2020) also mentioned that the ULTC aims to create a representative authentic and reliable resource that supports contrastive and translation studies research from English and French into Arabic and serves to aid comparisons between texts translated from these two language pairs. The ULTC also attempts to inform translation pedagogy and theories of translation and second language acquisition. It aims to improve translation pedagogical practices such as highlighting what should be presented earlier to a learner to become proficient at translating. Moreover, the ULTC can improve the quality of translation in terms of fluency, idiomatic expressions, and correct term usage in different genres.

The ULTC exemplifies Hareide's (2019) comparable parallel corpora by combining many subcorpora from different language pairs, such as English-Arabic and French-Arabic. The ULTC subcorpora (e.g., English-Arabic learner translator corpus, French-Arabic learner translator corpus, multimodal learner translator corpus, and the multitarget learner translator corpus, inter alia) are also in response to recent calls to adopt the triangulation approach in corpus design and analysis (e.g., Alves 2005; Baker and Egbert 2016; Taylor and Marchi 2018; Malamatidou 2018). The purpose is to enhance the "the completeness ('non-partiality')" representativeness that is evident in the complementarity between CBTS and LCR (i.e., parallel and learner corpora). This allows for the replication of research analysis and findings from second language acquisition research (SLA), CBTS, and LCR to the emerging field of learner translator corpus research (LTCR).

The first sections of this paper present ULTC design considerations, data collection criteria, and the exploitation of the corpus, while the latter sections evaluate the systems used for this corpus and suggest possible improvements. The rest of the paper is structured as follows: Section 2 presents an overview of learner translator corpora; Section 3 discuses design considerations based on a survey of the literature; Section 4 presents the planning phase, data collection criteria, and statistics. The concluding sections evaluate the corpus and such challenges as the lack of standardized corpus metadata, ethical issues in the context of the learner translator corpora, the choice of manual alignment or automatic alignment, and word counts and other statistical data in the case of multiple targets of the same source texts by the same learner.

2. Overview of learner translator corpora

Past decades have seen a great development in LCR (Granger 1993, 1994, 1996) and CBTS (Baker 1993, 1995, 1996). A learner corpus is a "collection of texts – written texts or transcribed spoken language – produced by language learners, and sampled so as to be representative of one or more combinations of situational and learner factors" (Borin and Prütz 2004: 69). The international corpus of learner English (ICLE) (Granger 1993, 1994) and the Cambridge learner corpus (CLC) (Nicholls 2003) are examples of early initiatives of LCR. The compilers of these projects highlighted the need to standardize the metadata that comes with learner corpora, including age, gender, L1, L2, proficiency level, etc. Granger (2004: 126) stated that "there are so many variables that influence learner output that one cannot realistically expect ready-made learner corpora to contain all the variables for which one may want to control."

Learner translator corpora (LTC) emerged at the intersection between LCR and CBTS as an offshoot of corpus linguistics (CL).

> Corpus linguistics is an approach to the study of language that involves collecting large quantities of naturally occurring language and using specialised software that manipulates that language to obtain information about frequencies, co-occurrences and meanings. The language may be spoken, written or signed, in one language variety or more, and one register or more. It consists of language which has occurred in natural contexts, not as the result of elicitation or introspection. The components of the corpus are texts (whole or partial) and thus consist of pieces of connected discourse. The quantity may range from a few hundred thousand words to billions, though the corpus usually contains more texts than could reasonably be read and remembered by an individual (Hunston 2022: 1).

The main interest of CL is on native data, and LCR, CBTS, and LTCR shifted the focus toward nonnative data. Learner translator corpora (LTC) contain authentic parallel data produced by second or foreign language learners or translation students. The student translation archive (STA) and student translation tracking system was an early LTC (Bowker and Peter 2003). This was followed by projects like the multiple Italian student translation corpus (MISTiC; Castagnoli 2009), the multilingual MeLLANGE learner translator corpus (Kübler 2008; Castagnoli, Ciobanu, Kübler, Kunz and Volanschi 2011), the Norwegian-English student translation corpus (NEST; Graedler 2013), the Universitat Pompeu Fabra learner translation corpus (LTC-UPF; Espunya 2014), the Russian learner translator corpus (RusLTC; Kutuzov and Kunilovskaya 2014), and the Czech-English learner translation corpus (CELTraC; Fictumova, Obrusnik, and Stepankova 2017). None of these projects, however, included Arabic data produced by foreign language learners or translation students. Thus, the ULTC is the first corpus that attempts to represent systematically authentic parallel Arabic learners' data produced by English and French foreign language learners and students of translation.

3. ULTC design considerations

The ULTC project began by reviewing many publicly available corpora to look at some practical issues and identify the best practices in translation and learner corpora. The main issues that were investigated were small versus large corpora, monolingual versus parallel corpora, product versus process corpora, native speaker versus learner or non-native corpora, and a sample corpus versus a whole-texts corpus.

In the debate over the sufficient size of a corpus resource (e.g., Hunston 2002; Granger, Gilquin and Meunier 2015; Timmis 2015), no clear conclusion was made. According to Fiona Farr and Anne O'Keeffe "for spoken corpora anything over one million words is considered to be moving into the 'larger' range, for written anything below five million is quite small" (2003). With regards to multimodal corpora that contain subtitled videos aligned with their parallel transcripts, no statement has been made about the representative size of the reviewed literature, especially considering the limitations of online and offline storage systems.

Word count and other statistical data for multiple targets of the same source texts by the same learner is another consideration for corpus compilation and statistical tool designs. How should words be counted in a multitarget parallel corpus? Do we count the same source text many times for each phase of the translation process? To the best of our knowledge, such questions have not been addressed since most of the available parallel corpora focus only on the product. Moreover, methodological issues are rarely discussed in detail.

With regards to genre, the labelling of text types has also been discussed. The concept of genre is difficult to define and the categories of genre are inconsistent and fluid in the literature. Distinguishing between concepts like genre, subgenre, domain, medium, and mode (Paltridge 2012) are problematic. The compilers of this ULTC decided to use genre and subgenre for the classifying of texts. Another important decision is on the number of texts to be included from each genre in a balanced corpus with different categories that still reflects the reality of PNU courses and tasks.

Using samples or whole-text corpora is another important consideration for translation corpora. As stated by Mikhailov and Cooper, "compiling a samples corpus involves more manual work than compiling whole-texts corpus. Because they are shorter, and there are more of them" (2016). In any case, the tasks of a naturalistic learner in the ULTC would determine the type of texts to be used.

The implementation of process versus product data is a key feature in the design of the ULTC project (see Section 4.2). In the literature, three approaches are used for monolingual corpus data. Two online approaches are: a) think-aloud protocols (TAPs) where translators are asked to reveal their mental processes in real-time while carrying out a translation task (Bernardini 1999: 181), which has been criticized for being a subjective approach, and b) a translog system (Jakobsen 1999), which has been promoted for being an objective documentation of the user's behavior. Translog is a software program that uses special characters in a linear representation of the keystrokes and pauses to indicate the translator's behavior while carrying out a translation task. The third approach, which is offline, involves analyzing multiple draft versions (Utka 2004).

Broadly speaking, the process-oriented approach is concerned with cognitive behavior that occurs in the mind of a translator or interpreter. The product-oriented approach, in contrast, focuses on the output product of a translated item. Basically, the ULTC functions as a raw data resource for researchers to serve their research objectives. Representing the compiled data as a pre-edited target text and post-edited target text enables researchers to use various methodologies in their translation studies, involving product-oriented, process-oriented, participant-oriented, or context-oriented research, for example (Saldanha and O'Brien 2013). The process- and product-oriented data can be explored with different parameters for descriptive and/or evaluative purposes, as driven-based and/or corpus-based samples, for comparable and/or parallel analyses, for quantitative versus qualitative sampling, as whole-texts or limited texts, or for longitudinal versus cross-sectional studies. The ULTC provides external criteria for users to be incorporated in their analyses along with the internal criteria (i.e., the texts). This design of the ULTC, that combines the process- and product-oriented methods can address the recent call of scholars to include a triangulation of methods when carrying out research

(Saldanha and O’Brien 2013). The design empowers researchers to use different perspectives, with the ultimate objective of finding innovative, precise, and reliable results.

Parallel corpus reusability is one of the main interests of the project, served by the design that promotes reusability of the data in the corpus. The product of the main parallel corpus can be compared to another parallel corpus, which thus allows triangulation, replication, and representativeness. Hareide’s “comparable parallel corpora” are taken into account with the design of the ULTC project, which is inspired by the reusability of the Norwegian-Spanish parallel corpus and comparable to P-ACTRES 2.0 (González and Izquierdo 2019).

The design of a resource must address a conglomerate of corpora in response to recent calls to adopt the triangulation approach (e.g., Alves 2005; Baker and Egbert 2016; Malamatidou 2018; Taylor and Marchi 2018) to enhance the completeness (non-partiality) and representativeness and to allow for replication of the research analyses. According to Sylvia Jaworska and Karen Kinloch:

> Triangulation is not a new territory in corpus linguistics; some researchers have adopted forms of triangulation, specifically investigator triangulation (Marchi and Taylor 2009) and method triangulation (Baker and Egbert 2016) demonstrating their benefits as well as limitations for CADS research. Yet, little attention has been paid to multiple data sets and data triangulation (2018).

Likewise, the choice of manual or automatic alignment, alignment software, amount of manual checking, cleaning files from noise, and storing the multitarget and multimodal data must be considered for such a resource. The definition of noisy text, however, is unclear in the literature on learner corpora, and parallel multitarget and multimodal corpora.

Finally, ethical issues are crucial in learner corpora, though little has been discussed on such topics in the literature. The researcher will obtain consent from the learners and from the institution to use their task data, but the learners’ identities will be anonymized. The source text copyrights, however, are not manageable with the massive number of sample texts that are translated and the tasks since most source books are not fully translated.

Data leakage in learner translator corpora is a sensitive issue for consideration. The exfiltration of such naturalistic data collected in a methodological framework (i.e., source, draft, final) is an inevitable possibility, and the large naturalistic project with its massive number of texts, learners, and instructors may not be able to prevent the leaking of data. The researcher should decide what must be done in case of any leaked data. Although learners’ naturalistic tasks can be used in any research, the learners’ data stemming from the drafting pattern: (i.e., source, draft, and final) have not been documented in any course before the ULTC compilers determined the data collection criteria. Cory Doctorow (2016) ironically stated that “the best way to secure data is never to collect it in the first place. Data that is collected is likely to leak."

4. ULTC design and data collection criteria

In the ULTC project, the data collection criteria were developed from a review of many translator learner corpora that identify best practices and consider the main objective of launching the corpus. The overall aim is to enrich the research in learner translation corpus studies from a number of approaches, described in the following sub-sections.

4.1 ULTC planning background

The current version of the ULTC was captured in November and December 2014. Due to the scarcity of translated materials serving academic research, especially in the Arabic language, that may be easily accessed by researchers and is feasible and rich in content, the idea arose to build a corpus housing graduation projects by undergraduates majoring in translation. The project targets copies of previous assignments and projects produced by undergraduates in the translation program. In 2014, a pilot project began collecting graduation projects by students majoring in translation at Princess Nourah bint Abdulrahman University (PNU), a large, public women's university located in Riyadh, the capital of Saudi Arabia. Unfortunately, graduation projects were only documented as hardcopies in the college library, and thus, we began collecting softcopies of the graduation projects each semester.

4.2 Data collection criteria and course overview

The ULTC corpus focuses on the content of graduation projects (i.e., source texts (STs) and target texts (TTs)) that can easily be compiled based on the last version of the students' projects, as long as the corpus is a leaner one. After all, it is important to trace how students produced their final work. The compilers suggested adopting the product- and process-oriented approaches as the design framework for the project. Hence, participants were asked to save their drafts. The English-Arabic and French-Arabic graduation projects that were submitted to the ULTC were based on the following corpus methodology: a) for the source text, the sample text was usually a book excerpt; b) for the pre-edited text (or draft), the first immediate draft produced by a learner of translation was used; and c) for the final version, the revised and final version of the draft was used after corrective feedback was given by the instructor.

The first stage of data collection occurred from 2015 to 2018, where graduation projects from eight successive academic semesters were collected. The graduation project course focuses on translator-to-be translational skills with the aim to enable students to practically translate based on what they learnt in previous semesters.

With the advice of supervisors, students selected the material to be translated, and the module and direction of translation. Two modules were available for students, namely: written translation and audiovisual translation. The average word count of a source text (ST) was 5,000 words, which each student translated during one semester under the guidance of the course supervisor. Students submitted their translations each week to their supervisors who gave corrective feedback on their work. The students then edited their translations before finally submitting their projects by the end of the semester. After final project submission, the students were evaluated by their supervisors. Since PNU is a women's university, all participants were female learners, though some of the supervisors are male.

Data was collected at the end of each academic semester, and the course supervisors submitted their students' files via the corpus e-mail. The compilers checked the e-mail submissions and then stored the files according to the corpus design criteria. The files were classified according to the academic calendar. The preprocessing stage involved keeping the data (the ST, pre-edited texts, and post-edited texts), removing any unnecessary data, and resolving unexpected or technical problems that arose. Based on the preprocessing analysis, the data was checked for how it accounted for the corpus. While checking the individual project files, metadata was created for each student's project. The metadata included the book title, genre, domain, word count, name of student, name of professor, and any remarks about the document satisfying the collection criteria.

After the projects were compiled for a semester, they were re-examined to preprocess data for alignment. Only the texts were saved, and the supervisors' comments, headers, names, assignment numbers, highlighted words (or phrases), glossaries, dedication and acknowledgements, cover page, table of contents, images, and any extra text boxes were removed manually by the corpus compilers.

5. ULTC exploitation

The French-Arabic graduation projects were uploaded in the *French-Arabic learner translator corpus* as a sub-corpus due to the significant differences of the English-Arabic projects from the main *English-Arabic learner translator corpus*. We included information that was related to the task, learner, and instructor as it may have some use as part of the metadata for the collected texts, in regards to the gender of the learner and the instructor, text genre, year of publication and translation, the learner's native language, level, and grade, which are connected to the learner parents' level of education and mother tongue. Other bits of information that may have been used in the metadata included the learners' scores in a standardized test and their length of exposure to English- or French-speaking communities. The students were also asked to submit a learner profile with their submission of the three documents mentioned above.

After reviewing some of the projects, we decided to add a translator's preface. Students were instructed to write the translator's introduction (i.e., preface) in the TL, which was a reflective essay directed to readers that summarized the learner's experience. It allowed students to gain an understanding of the practices in real-life working conditions. The length of the introduction was about 200-300 words, and it typically covered the chosen translation approaches, strategies, and technical and translation challenges that the learner needed to overcome. Students summarized their experience of translating, the kinds of problems (e.g., translation techniques, cultural gaps, degree of formality, etc.) they resolved, the knowledge they gained from the course, and any other points that wanted to mention (like introducing a new term or their reason for using one method over another). Because of their potential value for future research, the prefaces were uploaded in a subcorpus.

Most of the pre-edited texts included the comments of supervisors, though the corpus compilers decided to remove them at this stage for a number of reasons. The comments were unsystematic, unclear, idiosyncratic, or merely highlighted phrases that would not likely contribute to the objectives of the project. The removal of the comments could also lead to other potential research endeavors that would be untainted by the supervisors' comments. In any case, the comments were saved in the raw corpus for possible future consideration.

The corpus is not concerned with course requirements such as project formatting. Therefore, the cover page, dedication, acknowledgement, glossaries, table of contents, photos, graphs, student name, assignment number, and the formatting for headers, footers, page numbers, and page border are removed. Most pre-edited texts include the SL and the draft translation and are displayed in tables or paragraphs to make the reviewing process easier for the supervisor. SL parts are also removed in the pre-edited texts to avoid ST duplication. In some projects, drafts were submitted in separate files, and after being examined, they were reorganized by compiling them into one folder for each project.

The ULTC projects are generally sample texts. Some of the projects are whole texts translated by different learners, however, they are uploaded in the corpus resource as sample texts that can be retrieved as whole texts by searching for the title of the text. They are presented diachronically in the corpus, and reflect the temporal sequence of graduation projects based on the academic semesters.

5.1 ULTC multimodal corpus

By the second semester of the academic year 2018, the translation department dedicated most of the graduation project sections for the translation of videos. Still, many sections were dedicated to book translations. Audiovisual graduation projects were considered under the *ULTC multimodal corpus,* which was comprised of almost 341 video projects in this phase. The number of word tokens for each source script was 2000 to 2500. Because most of the projects were documentaries, the scientific genre prevailed. Other genres included health, medicine, geography, history, linguistics, technology, psychology, and biography. Written transcriptions of the videos are available in the ST and TT files using the same collection criteria (i.e., source, pre-edited or draft, final or post-edited transcriptions). All videos were subtitled in the Arabic language. For long videos, a group of students performed the translation, where the source script word count was about 2,500 words per student.

We can infer that using a "multimodal" as a subcorpus shows promise regardless of the representative size. A subcorpus could serve as the raw content for researchers to describe, understand, analyze, and/or evaluate student practices in audiovisual translation. The results from future research studies would be expected to contribute by consolidating many of the issues being addressed in this field, and help clarify methods, methodologies, and pedagogies, etc. The benefits from new research could be extended to compare, for example, audiovisual versus written translations.

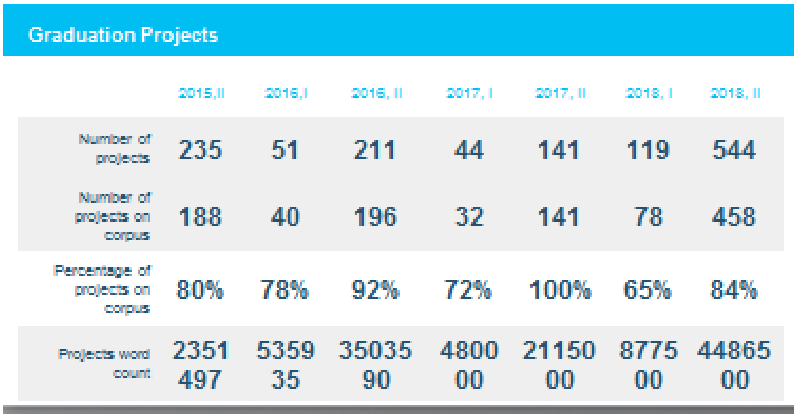
5.2 ULTC multi-target corpus

Learner tasks and assignments from other translation courses were implemented in the academic year 2018 as a new family member in the ULTC project (*the multi-target corpus*) to consider different proficiency levels of the translation learners. This subcorpus began by housing assignments and tasks from the *Specialized Translation En-Ar I* course, taught to undergraduates in Level three. It is the first course for practicing translation from English into Arabic, and the students use the same texts from three fields: science and technology, medicine, and literature. The total number of assignments was 13, with 5 from science, 4 from medicine, and 3 from literature. For each assignment, from 200 to 300 multiple translations of the same task were submitted to the corpus. The length of the source text for each assignment was from 250 to 450 words.

6. ULTC statistics

The total number of projects in 2015, semester II was 235, of which 181 were submitted to the corpus. In 2016, semester I, the total number of projects was 51, of which 40 were submitted. In 2016, semester II, 221 of 224 projects were submitted. In 2017, semester I, 32 of 44 projects were uploaded to the corpus; and in semester II, all of the 144 projects were submitted to the corpus. In 2018, semester I, the total number of projects was 119, of which 117 were submitted to the corpus. In semester II, 458 of 544 projects were submitted to the corpus. The current size of the corpus is almost 75 million word tokens (Table 1).

Table 1. ULTC statistics



Graduation Projects

	2015,II	2016,I	2016, II	2017, I	2017, II	2018, I	2018, II
Number of projects	235	51	211	44	141	119	544
Number of projects on corpus	188	40	196	32	141	78	458
Percentage of projects on corpus	80%	78%	92%	72%	100%	65%	84%
Projects word count	2351497	535935	3503590	480000	2115000	877500	4486500

Most of the projects submitted to the corpus meet the corpus design criteria, and only a few of the submitted projects were incomplete or excluded. Most of the problems encountered are technical, such as unreadable files, scanned drafts or STs, or files that cannot not be accessed or are empty. The most problematic semester, in terms to identifying and correcting the issues, occurred in semester I in 2015.

Four scanned STs were submitted and one ST file could not be opened. In addition, 20 drafts were missing and 5 projects had incomplete drafts, and 12 TTs were unreadable due to file conversion errors. In semester II of the same year, clear instructions and a better awareness reduced the number of technical problems to just one missing draft and one missing TT. In the subsequent semesters, while technical problems still arose, the issues tended to be minor and controllable.

Figure 1 shows the approximate word counts of the STs, drafts, and TTs. For instance, in 2015 the ST word count was 952,203, the draft word count was 698,879, and the TT word count was 700,415. In 2016 the ST word count was 188,847, the draft word count was 158,241, and the TT word count was 188,847. In 2016, semester II, the ST word count was 1,130,324, the draft word count was 1,364,670, and the TT word count was 1,008,596. Usually, the second semester has a higher number of students, because of the academic calendar.

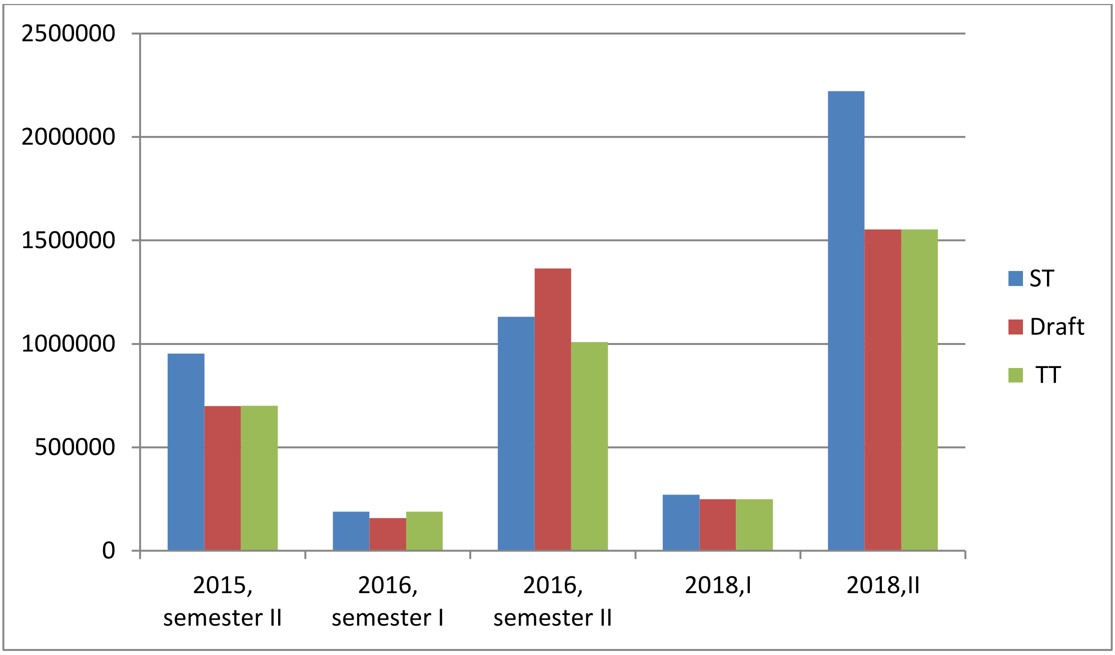


Figure 1. Word counts of projects submitted to the ULTC.

With regards to the genres in the corpus (Figure 2), the ULTC contains a variety, including: education, medicine and health, translation, science, self-help, management, communication skills, culture, religion, sport, politics, psychology, sociology, linguistics, and literature. For instance, in 2015, the projects were classified as self-help books (87), medicine and health (36), education (24), management (20), communication skills (8), culture (4), religion (1), sport (1), translation (1), science (1), psychology (1), and sociology (1). In semester I of 2016, the projects were classified as education (20), medicine and health (15), translation (1), and self-help (1). Nevertheless, in semester II of the same year, the projects were classified as health (67), education (55), medicine (20), management (17), literature (10), linguistics (10), social studies (6), and translation (1). In 2017 and 2018, the projects were classified as health, education, translation, linguistics, psychology, and literature.

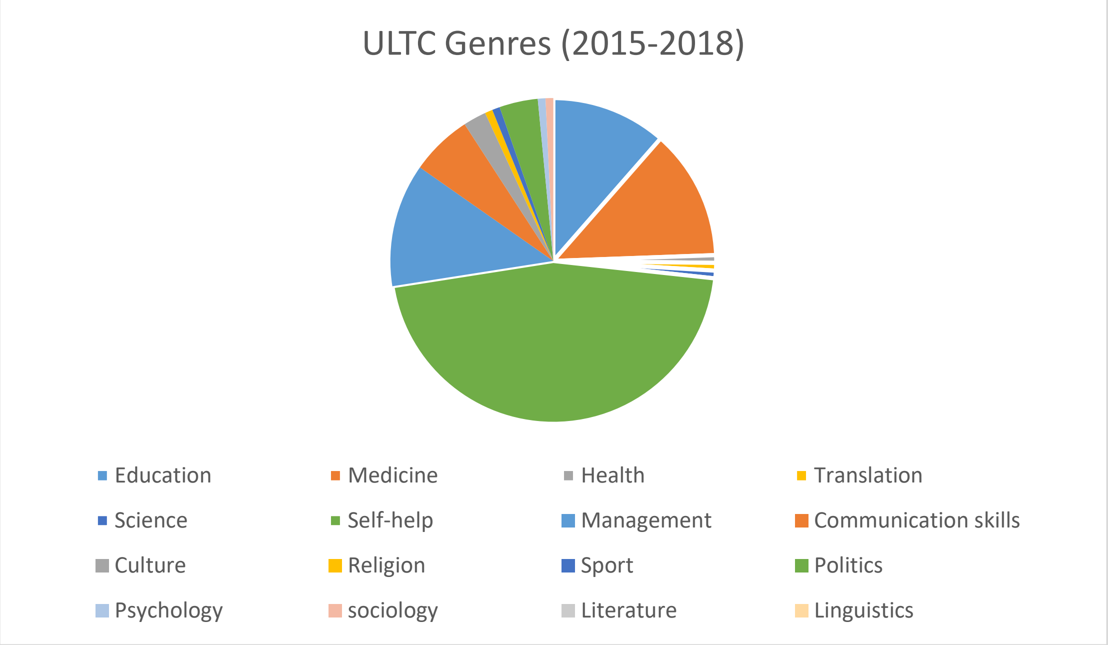


Figure 2. ULTC genres (2015-2018).

7. ULTC multiple alignment

For the English-Arabic learner translator corpus, French-Arabic learner translator corpus, and the multimodal learner translator corpus, multiple alignment involves the following steps: 1) source and pre-edited data submitted to the corpus are aligned automatically using +Align and Trados alignment tools; 2) source and final or post-edited data submitted to the corpus are aligned automatically using +Align and Trados alignment tools; 3) the three versions are manually aligned in an MS-Excel file for each project; 4) projects are checked and re-aligned after spotting errors, mis-aligned segments, and empty links; 5) files are converted to the XML format and then uploaded to the ULTC website.

The *multi-target learner translator corpus* data is only aligned manually as it contains multiple targets of the same source text that cannot be handled automatically. After conversion to the XML format, it is uploaded to the corpus website as a subcorpus. Project metadata and annotations can be added manually to the MS-Excel file (as a template) before it is uploaded to the corpus admin website https://arabicparallelultc.com. The corpus is available in the XML format through the user-friendly web interface, which has a concordancer that supports bilingual search queries and several filtering options (Figure 3).

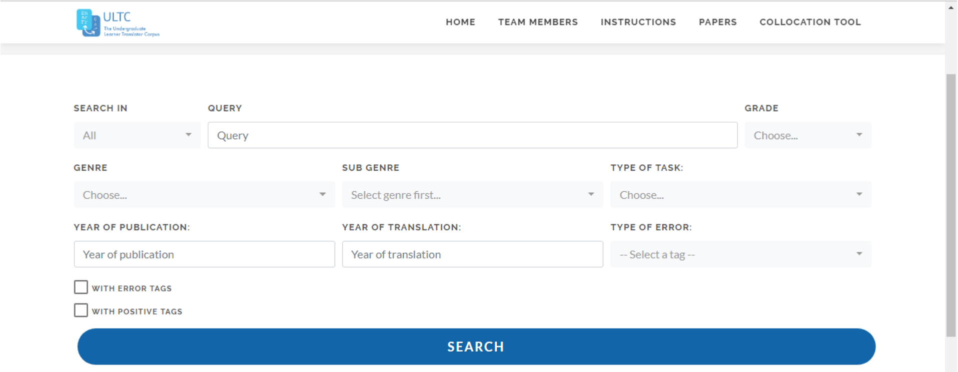


Figure 3. ULTC advanced search.

8. ULTC evaluation

This section evaluates the ULTC corpus and highlights its potential benefits and limitations. The ULTC is a promising resource for teaching and research, as it provides interesting data with standardized metadata and comprehensive information about the translation learner, the task, the instructor, and the course. One of the main benefits of the resource is the synergies that exist between multiple corpora with different language pairs (i.e., English-Arabic and French-Arabic), different modules (i.e., written and multimodal), and multitargets of a source text translated by a single learner (i.e., draft and final versions of a graduation project), and multitargets of the same source text translated by different learners (i.e., multitarget translation of assignments and tasks at other proficiency levels). Hence, the ULTC has comparable parallel corpora that allow users to triangulate their teaching and research.

The ULTC project is continually expanding. The issue of representativeness is one of the main concerns of the project, and the ULTC can be seen as a representative corpus in terms of its size in its early stage (>70 million tokens). Nevertheless, the collection of naturalistic data from PNU undergraduate courses presents some limitations. First, some genres (i.e., self-help, science, medicine, and management) dominate, because of the nature of the courses. Researchers will need to make a tough decision about whether to eliminate some projects, so as to end up with a balanced corpus or keep projects from the dominant genre for their potential teaching and research benefits. Second, the ULTC is limited in terms of representing male learners of translation since PNU is a female university. Translations produced by male learners of translation from other universities will be considered as the project expands in the future. In any case, the nature of projects and tasks tends to vary across different institutions, which may lead to different data collection criteria and texts types.

One of the challenges encountered when constructing the resource was to consider how to count the words and tokens in the multitarget parallel resource. To the best of our knowledge, this issue has not been discussed in prior research dealing with the methodological framework that was used in this project. The question remains: should we count the tokens of a source text translated by different learners for each submission of the same text? If yes, this would lead to the duplication of source segments when they are submitted to the corpus website. The system used at the ULTC website was designed to consider the wordcounts of source text, and not the target segment wordcount. A decision has been made to update the system and add the ability to count target tokens, however, the task is not an easy one since the project compliers had to remove all files from the website and upload them again to be counted.

The ULTC templates have been set up to be compatible with the system used at the corpus website. This creates a sensitive issue that then restricts the flexibility in considering future adjustments to the corpus metadata or design as the resource expands in the future. Moreover, sustainability and future maintenance challenges are inevitable, when dealing with a self-funded project that contains a large amount of data from different modules (i.e., text, video, and audio files) that require a large, high-cost server.

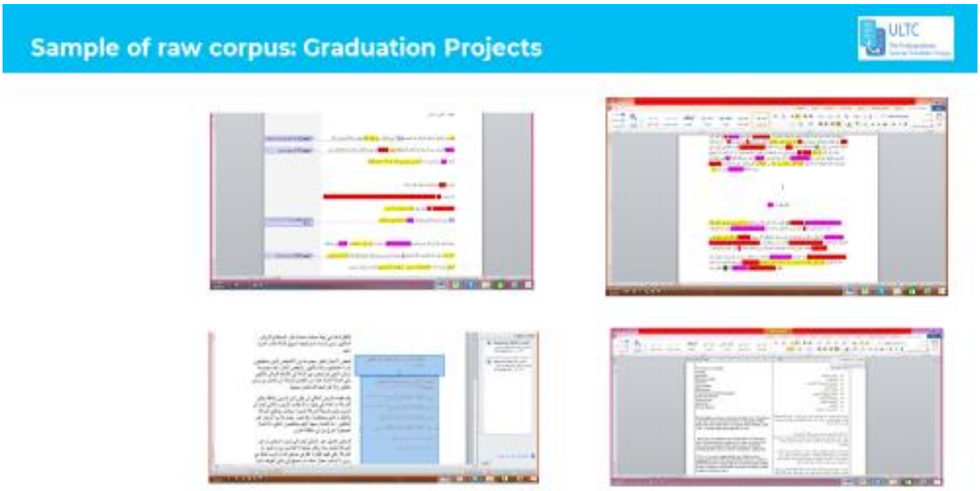
As a learner corpus resource, the corpus will be fully annotated and error-tagged in the near future. Researchers will be able to add their tags online at the ULTC interactive interface designed for error-tagging and annotations from the admin website. As most researchers do not have access to the admin website, they will need to add tags to the files offline for later uploading to the corpus once the system is completed. Nevertheless, full text access cannot be provided since source text copyrights will create some ethical concerns during the error-tagging phase.

Limitations are also present in this promising research area, with regards to the alignment and part-of-speech tagging tools. The question remains: should existing alignment tools be used and then any misalignments be fixed manually, or should the files be aligned manually from the start? Another concern is in regards to the part-of-speech tagging for different language pairs as they require different tools for each language. The noise part-of-speech tags could add further errors, and the use of positive tags in the context of learner data is another important challenge.

Empty links, crossing lines, and the correspondence case are common issues for a learner resource to represent the desired case. Such issues are challenging as they can be confused with any misalignments or noisy texts. According to Čulo, Hansen-Schirra, Maksymski and Neumann:

> Units in the target text may not have matches in the source text and vice versa; thus, no connection can be drawn and we speak of empty links. Units which do have a counterpart with which they are aligned may be embedded in higher units which are not aligned, resulting in crossing lines. This is, for instance, the case when a word is embedded in a chunk with the subject function in one language, and its counterpart in a chunk with the object function (2017).

Dataset creation projects can be difficult and laborious and can require even more scientific rigor than standard research projects. Most of the problems with the ULTC are technical, such as the main problem with STs that make the scanned texts difficult to align. Noisy file problems also occur when files are converted from PDF to Word documents (Figure 4). Missing or incomplete drafts is a common problem with pre-edited texts. As for the TTs, some errors can be missed, and some files may have a loss in accuracy with regards to data formatting, including indentation, paragraphing, spacing, or use of Italics or bold type. Another challenging problem can arise when dealing with footnotes, which are important since they often reflect the translator’s decisions. Although not common, a problem can occur when marking or distinguishing footnotes from the main text when a file is converted to XML.



Sample of raw corpus: Graduation Projects

Figure 4. Noisy raw files.

9. Conclusion

The ULTC is a parallel corpus that is housed within a collection of undergraduate tasks and projects in the translation department at PNU. The main aim of the project is to make data accessible for teaching and research. This paper critically reflected and evaluated the project’s potentials and limitations, while considering the design criteria and compilation process of data gathering. The first phase of the project has targeted graduation projects from 2015 to 2018 and a collection of assignments from a written translation course in 2018. The raw data has been processed to be aligned and become available to ULTC users. The estimated word count of the corpus is about 75 million word tokens.

We tried to cast light on the unique aspects of the resource that may serve potential users with regards to the detailed metadata, languages represented, tasks covered, and the large population of learners that were involved. Multiple corpora can allow for multidimensional studies such as examining the change of frequencies in types between a source and the targets. We hope that new studies will address translation issues that have not been previously examined, and that future studies will further examine topics like word sense disambiguation, terminology extraction, descriptive translations, comparative stylistics, contrastive discourse analyses, lexicography, the translation process, collocation, colligation, translation assessment and evaluation, error analysis, cross-cultural pragmatics, statistical and psycholinguistic modelling of the translation process, transfer, translators’ L1 or

source language interference, comparisons of the frequency of a node word against the target, and comparisons between learner and professional corpora.

Reem F. Alfuraih (Lecturer in Applied Linguistics) – Corresponding Author
Department of Applied Linguistics, College of Languages
Princess Norah bint Abdul Rahman University, Saudi Arabia
ORCID Number: 0000-0003-3199-1989
Email: Rfalfuraih@pnu.edu.sa

Noha M. El-Jasser (Lecturer in Translation)
Department of Translation, College of Languages
Princess Noura bint Abdulrahman University, Saudi Arabia
ORCID Number: 0009-0008-4350-3958
Email: nmeljaseer@pnu.edu.sa

References

Alfuraih, Reem. (2019). 'The undergraduate learner translator corpus: A new resource for translation studies and computational linguistics'. *Language Resources and Evaluation,* doi:10.1007/s10579-019-09472-6

Alves, Fabio. (2005). 'Triangulation in process-oriented research in translation'. In Fabio Alves (ed.), *Triangulating Translation: Perspectives in Process Oriented Research*, 3-24. Amsterdam: John Benjamins. doi:10.1075/btl.45.04alv

Baker, Mona. (1993). 'Corpus linguistics and translation studies: Implications and applications'. In Mona Baker, Gill Francis and Elena Tognini-Bonelli (eds.), *Text and Technology: In Honour of John Sinclair*, 233-50. Amsterdam: John Benjamins.

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

Baker, Mona. (1996). 'Corpus-based translation studies: The challenges that lie ahead'. In Harold Somers (ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, 175-86. Amsterdam: John Benjamins.

Baker, Paul and Jesse Egbert. (2016). *Triangulating Methodological Approaches in Corpus Linguistic Research*. London: Routledge.

Bernardini, Silvia. (1999). 'Using think-aloud protocols to investigate the translation process: Methodological aspects'. In N.J. Williams (ed.), *RCEAL: Working Papers in English and Applied Linguistics 6*, 179-199. Cambridge: University of Cambridge.

Biber, Douglas. (2014). 'Using multi-dimensional analysis to explore cross-linguistic universals of register variation'. *Languages in Contrast*, 14(1): 7-34. doi:10.1075/lic.14.1.02bib

Bowker, Lynne and Peter Bennison. (2003). 'Student translation archive: Design, development and application'. In Federico Zanettin, Silvia Bernardini and Dominic Stewart (eds.), *Corpora in Translator Education*, 103-117. London and New York: Routledge.

Borin, Lars and Klas Prütz. (2004). 'New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language'. In Guy Aston, Silvia Bernardini and Dominic Stewart (eds.), *Corpora and Language Learners*, 67-87. Amsterdam: John Benjamins.

Castagnoli, Sara. (2009). Regularities and variations in learner translations: A corpus-based study of conjunctive explicitation. PhD Dissertation, University of Pisa.

Castagnoli, Sara, Dragoș Ciobanu, Natalie Kübler, Kerstin Kunz and Alexandra Volanschi. (2011). 'Designing a learner translator corpus for training purposes'. In Natalie Kübler (ed.), *Corpora, Language, Teaching, and Resources: From Theory to Practice,* 221-248. Bern: Peter Lang.

Čulo, Oliver, Silvia Hansen-Schirra, Karin Maksymski and Stella Neumann. (2017). 'Empty links and crossing lines: Querying multi-layer annotation and alignment in parallel corpora'. In Silvia Hansen-Schirra, Stella Neumann, and Oliver Čulo (eds.), *Annotation, Exploitation and Evaluation of Parallel Corpora*, 47-80. Berlin: Language Science Press. doi:10.5281/zenodo.283498

Doctorow, Cory. (2016). The privacy wars are about to get a whole lot worse. https://locusmag.com/2016/09/cory-doctorowthe-privacy-wars-are-about-to-get-a-whole-lot-worse/

Espunya, Anna. (2014). 'The UPF learner translation corpus as a resource for translator training'. *Language Resources and Evaluation*, 48: 33-43.

Fictumova, Jarmila, Adam Obrusnik and Kristyna Stepankova. (2017). 'Teaching specialized translation error tagged translation learner corpora'. *Sendebar*, 28: 209-241.

Graedler, Anne-Line. (2013). *NEST*—a corpus in the brooding box. *Studies in Variation, Contacts and Change in English: Corpus Linguistics and Variation in English: Focus on Non-Native Englishes*, 13.

Granger, Sylviane. (1993). 'The international corpus of learner English'. In Jan Aarts, Pieter de Haan and Nelleke Oostdijk (eds.), *English Language Corpora: Design, Analysis and Exploitation*, 57-69. Amsterdam: Rodopi.

Granger, Sylviane. (1994). 'The learner corpus: A revolution in applied linguistics'. *English Today*, 39 (10/3): 25-29.

Granger, Sylviane. (1996). 'From CA to CIA and back: An integrated contrastive approach to computerized bilingual and learner corpora'. In Karin Aijmer, Bengt Altenberg and Mats Johansson (eds.), *Languages in Contrast. Text-Based Cross-Linguistic Studies*, 37-51. Lund: Lund University Press.

Granger, Sylviane. (2004). 'Computer learner corpus research: Current status and future prospects'. In Ulla Connor and Thomas Upton (eds.), *Applied Corpus Linguistics: A Multidimensional Perspective*, 123-145. Leiden: Brill.

Granger, Sylviane, Gaëtanelle Gilquin and Fanny Meunier. (2015). *The Cambridge Handbook of Learner Corpus Research.* 10.1017/CBO9781139649414

Hareide, Lidun. (2019). 'Comparable parallel corpora: A critical review of current practices in corpus-based translation studies'. In Irene Doval and M. Teresa Sánchez Nieto (eds.), *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*, 19-38. Amsterdam: John Benjamins.

Hareide, Lidun and Knut Hofland. (2012). 'Compiling a Norwegian–Spanish parallel corpus: Methods and challenges'. In Michael Oakes and Meng Ji (eds.), *Quantitative Methods in Corpus Based Translation Studies*, 75-114. Amsterdam: John Benjamins.

Hunston, Susan. (2002). *Corpora in Applied Linguistics.* Cambridge: Cambridge University Press.

Jakobsen, Arnt Lykke. (1999). 'Translog documentation'. In Gyde Hansen (ed.), *Probing the Process in Translation: Methods and Results*, 151-186. Copenhagen: Samfundslitteratur.

Kübler, Natalie. (2008). 'A comparable learner translator corpus: Creation and use'. In *Proceedings of the Comparable Corpora Workshop of the LREC Conference*, 73, 78, Marrakech, 28-30 May 2008. http://www.lrecconf.org/proceedings/lrec2008/workshops/W12_Proceedings.pdf

Kutuzov, Andrei and Maria Kunilovskaya. (2014). 'Russian learner translator corpus: Design, research potential and applications'. In Petr Sojka, Ales Horak, Ivan Kopecek and Karel Palak (eds.), *Text, Speech and Dialogue. Lecture Notes in Computer Science*, 315-323. Berlin: Springer.

Marchi, Anna and Charlotte Taylor. (2018). *Corpus Approaches to Discourse: A Critical Review*. New York: Routledge.

Malamatidou, Sofia. (2018). *Corpus Triangulation: Combining Data and Methods in Corpus-Based Translation Studies*. New York: Routledge

Mikhailov, Mikhail and Robert Cooper. (2016). *Corpus Linguistics for Translation and Contrastive Studies: A Guide for Research.* London: Routledge.

Nicholls, Diane. (2003). 'The Cambridge learner corpus – Error coding and analysis for lexicography and ELT', In Dawn Archer, Paul Rayson, Andrew Wilson and Tony McEnery (eds.), *Proceedings of the Corpus Linguistics 2003 Conference*, 572-581. Lancaster University.

Paltridge, Brian. (2012). *Discourse Analysis.* 2nd edition. London: Bloomsbury.

Sanjurjo-González, Hugo and Marlén Izquierdo. (2019). 'P-ACTRES 2.0: A parallel corpus for cross-linguistic research'. In Irene Doval and M. Teresa Sánchez Nieto (eds.), *Parallel Corpora for Contrastive and Translation Studies: New Resources and Applications*. John Benjamins: Amsterdam.

Saldanha, Gabriela and Sharon O'Brien. (2013). *Research Methodologies in Translation Studies*. New York: Routledge.

Timmis, Ivor. (2015). *Corpus Linguistics for ELT: Research and Practice*. New York: Routledge.

Utka, Andrius. (2004). 'Phases of translation corpus: Compilation and analysis'. *International Journal of Corpus Linguistics*, 9: 195-224. 10.1075/ijcl.9.2.03utk