Lang Resources & Evaluation (2020) 54:801–830
https://doi.org/10.1007/s10579-019-09472-6

PROJECT NOTES

# The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics

**Reem F. Alfuraih**[1]

Published online: 24 July 2019


**Abstract** Around the world, a growing interest has been seen in learner translator corpora, which are invaluable resources for teaching and research. This paper introduces a new resource to support researchers from different interdisciplinary areas such as computational linguistics, descriptive translation studies, computer-aided translation technology, Arabic machine translation applications, cognitive science, and translation pedagogy. Motivated by the lack of learner translator resources that provide data about learners of translation from and into Arabic, the undergraduate learner translator corpus (ULTC) is an ongoing, error-tagged sentence-aligned parallel corpus of English, Arabic, and French, with Arabic as its main language. The present corpus, consisting of parallel texts of female learners of translation from English or French into Arabic, is the first of its kind in terms of the languages represented, tasks covered, and number of students involved. It is also unique in terms of combining many complementary corpora of cross-lingual data, each of which has its own web-based query interface and corpus analysis tools. This paper describes the ULTC compilation process, preliminary findings, and planned future expansion and research.

**Keywords** Translation pedagogy · Arabic parallel corpus · Multilingual corpus · Multimodal corpus · Interpreting corpus · Triangulation

✉ Reem F. Alfuraih
Rfalfuraih@pnu.edu.sa

[1] College of Languages, Princess Nourah bint Abdulrahman University, PO Box 7455, Riyadh 14215, Saudi Arabia

## 1 Introduction

Parallel corpora are textual databases that contain "two or more versions of the same texts; first as original texts and then their translations in the other language(s)" (Mikhailov and Cooper 2016). Learner corpora, on the other hand, are computerized textual databases produced by non-native speakers (Granger 2002). This project aims to build an error-tagged learner translator corpus that is focused during its initial stage on graduation projects of female students majoring in translation at Princess Nourah Bint Abdurahman University (PNU), a public women's university located in Riyadh, Saudi Arabia. The corpus has been extended to represent different proficiency levels of translation learners. The learners' mother tongue is Arabic. The current corpus size is more than 55 million-word tokens and it contains many subcorpora that are cross-lingual in nature. The corpus has been annotated with metadata searchable via a web-based query interface on factors that can affect the students' performance such as information about the translation task, information about the learner, and information about the instructor.

The aim of the present paper is to introduce the undergraduate learner translator corpus (ULTC) and its complementary subcorpora and to briefly describe its design, annotation, and web-based search interface. The remaining paper is structured as follows: Sect. 2 lists the project objectives and potentials. Then, a comprehensive overview of related work to ULTC is stated in Sect. 3. The project design, compilation, annotation, analysis tools, and search interface are discussed in Sect. 4. Preliminary findings are presented in Sect. 5. Finally, the concluding sections are dedicated to ULTC future work and expansion.

## 2 ULTC rationale and objectives

The present project is intended to create a representative and reliable learner translator resource of naturalistic data to support various studies of translation from and into the underrepresented language—Arabic. It also aims to develop a standardized corpus-driven error taxonomy that can provide teachers, students, and researchers with resources on errors that are common among undergraduate learners in translations from and into Arabic. In addition, one of the main objectives of the ULTC is to develop a corpus-driven quality-assessment framework that measures competence, creativity, and positive practices in the translations of undergraduate learners from and into Arabic.

Another primary focus of the project is to provide a representative amount of undergraduate learner interpreting data to examine interpreted texts, interpreting errors, and other corpus-based interpreting studies. Unlike existing learner translator corpora, this project involves the development of consistent comprehensive standardized metadata, searchable via a web-based query interface. The project will explore undergraduate learner translators' drafting and revision patterns. Moreover, this resource will provide data to shed light on studies of the features of translation as a product, such as universals; simplification, normalization,

explicitation, standardization and leveling out, and contribute objectively to the new studies of the process of translation. Thus, the ULTC will present an invaluable resource for corpus triangulation and comparability where multiple corpora can be integrated in one study of a single phenomenon.

## 3 Corpora in translation studies

Corpora have been widely used in applied linguistics and translation studies. The focus of corpus-based translation studies in the early 1990s were on the study of professional translations to investigate translation universals, translator style, and features of translated language. A composite approach that combines balanced parallel and comparable corpora was introduced due to the lack of available parallel resources in many languages (Xiao and McEnery 2002). Many studies attempted to extract parallel data from multilingual non-parallel resources, mainly comparable corpora of data from different languages that were linked according to the same sampling method (Fung and Cheung 2004; Hewavitharana and Vogel 2011; Smith et al. 2010; Stefanescu et al. 2012). In the case of Arabic, most corpus-based studies have been on multilingual corpora where Arabic is one of the languages available in them, such as *The Translational English Corpus* (Baker 1993, 1999) and the *Information Technology Translational Arabic Corpus* (Izwaini 2003).

Few English–Arabic/French–Arabic parallel corpora are available that support research in contrastive linguistics, machine translation, or lexicography. One of the available Arabic parallel corpora is the *English–Arabic Parallel Corpus of United Nations Texts* (EAPCOUNT), a 5,392,490-word parallel corpus comprised of United Nations (UN) annual reports for supporting linguistic research (Rafalovitch and Dale 2009; Salhi 2013). Another resource is the *Open Parallel Corpus* (OPUS), which presents different parallel texts collected from the web to support machine translation research (Tiedemann 2012). Moreover, the *Arabic–English Parallel Corpus* (AEPC) was designed to enhance translation training and language teaching (Alotaibi 2017).

Other English–Arabic and French–Arabic parallel corpora are either very limited in their availability or size, such as the *Kuwait University English–Arabic Parallel Corpus*, which is only available to the university's staff and students (Al-Ajmi 2004) and the *Linguistic Data Consortium GALE Corpus* that consists of only 42,089 words (Li et al. 2013). Further, *An-Nakel El-Arabi1 Translate-Net*, *800-Translate*, and *TRAD* are examples of restricted Arabic–French corpora that were supported by different institutions and companies: CIMOS, Responsive Translation, and the French Ministry of Defense, respectively. A larger, more available Arabic parallel corpora is needed.

Some available Arabic parallel corpora are multimodal, combining textual parallel data with other modalities. *AMARA* is a multimodal English–Arabic corpus that focuses on subtitles from educational videos from Technology, Entertainment, and Design (TED) and the Khan Academy by the Qatar Computing Research Institute (Guzman et al. 2013). *An Arabic–Hebrew parallel corpus* of TED talks is another resource comprised of 2000 talks subtitled in Arabic and Hebrew for a total

of about 7 M tokens (Cettolo 2016). Furthermore, *MultiNews* is a corpus of comparable documents and their images in nine languages including Arabic from web news articles of the Euronews website (Afli et al. 2017).

On the other hand, corpora that contain audio recordings and their transcriptions are known as interpreting corpora. A few interpreting corpora are available (Shlesinger 2008; Russo et al. 2012; Hu and Tao 2013). *Arabic Speech Corpora* include only translated speech transcripts (Kumar et al. 2014; Zaidan and Callison-Burch 2014). The *WAW Corpus* is the only publicly available English–Arabic interpreting resource with interpretations of lectures and speeches from international conferences aligned with their transcripts of "the original speeches and of their interpretations, as well as human translations of both kinds of transcripts into the opposite language of the language pair" (Temnikova et al. 2017). Accordingly, Arabic interpreting corpora are very rare compared to those that are textual parallel.

Some parallel corpora are based on electronic texts produced by learners of translation. Research in building learner translator corpora has expanded over the last few years, with the learner corpora being "systematic computerized collections of texts produced by language learners" (Granger 2002). In learner translator corpora, electronic texts produced by learners of translation are collected. Bowker and Peter (2003) proposed that the "Student Translation Archive (STA) and student translation tracking system" be used for storing and retrieving student translations from the tracking system. The *Polish and English Language Corpora for Research and Applications* (PELCRA) is another early learner translation corpus (Uzar and Walinski 2001).

Most learner translator corpora focus on the representation of the performance of translation trainees, such as the *ENTRAD* (Florén 2006), the *Multilingual e-Learning in LANGuage Engineering* (MeLLANGE; Castagnoli et al. 2011), *Korpusprojekt zur Translatinsevaluation* (KOPTE; Wurm 2013), and the *Russian Learner Translator Corpus* (RusLTC; Kutuzov and Kunilovskaya 2014). Other corpora represent the performance of the translation learners across different proficiency levels such as the *Russian Translation Learner Corpus* (RuTLC; Sosnina 2006), the *Norwegian–English Student Translation Corpus* (NEST; Graedler 2013), *Universitat Pompeu Fabra Corpus* (UPF; Espunya 2014), and the *Czech–English Learner Translation Corpus* (CELTraC; Štěpánková 2014). Nevertheless, the *Multiple Italian Student Translation Corpus* (MISTiC) presents post-graduate students' textual productions (Castagnoli 2009).

Corpus metadata "considerably extends the range of research questions that a corpus can readily address" (McEnery and Xiao 2007). Error analysis is the main function of the learner corpora, and from the above-mentioned corpora, STA and MISTiC are the only ones that have not yet been error-tagged. PELCRA is unique by being annotated with a set of positive tags called Positive Feedback that brings attention to the practices of competent learner translators (Uzar and Walinski 2001). Some corpora contain information about individual learners, such as age or native

language (e.g., STA and MeLLANGE), while other corpora provide no information about individual learners. Instead, they focus on the translation task and other contextual information (e.g., UPF). Each corpus appears to have its own purpose and limitations when providing metadata. Standardized, consistent metadata is lacking for translation corpora. New approaches can be incorporated, such as adding the translators' prefaces or reflective essays as meta information that is linked back to the task. Several studies highlight the importance of studying the translators' prefaces as meta-texts that can provide valuable theoretical and practical insights (Jakobsen 2003; Dimitriu 2009; Norberg 2014).

With regards to the translation direction, PELCRA is unique in its focus on the foreign target language, whereas, the target language in the rest of the corpora is native (e.g., STA, RuTLC, KOPTE, and MeLLANGE) or mixed (e.g., ENTRAD). Some corpora contain different source texts aligned with their translations by the same learner (e.g., PELCRA). Other corpora, however, are multiple translation corpora containing several translations of the same source text produced by different learners (e.g., MISTiC, RuTLC, and UTF). In addition, STA, NEST, and MISTiC corpora are not available online. No learner translator corpus is available at this time to provide data about the learners of translation from and into Arabic.

Learner corpora are typically compared to the language of native speakers. Some learner corpora have been subdivided into subcorpora to permit comparisons, or for comparisons to highly proficient users (Izquierdo et al. 2008). MeLLANGE is the only learner translator corpus that contains reference to a professional translation subcorpus. Besides comparisons to professional translators and native writers, comparing a translation and interpretation tasks to foreign language writing and speaking tasks is another possible consideration in this project. Comparisons will enhance the resource representativeness and allow for the replication of the same methodology and hypothesis to different datasets. As Espunya (2014) puts it, research in learner translator corpora is still in "its infancy, judging by the scarcity of publications reporting results or even research programs."

In the past few years, the focus of translation studies has shifted from descriptive translation studies that examine the translation product to empirical translation studies that provide explanatory and predictive models of the translation process (e.g., Jakobsen 2011; Carl and Dragsted 2012; Mesa-Lao 2014; Carl et al. 2015). Tanslog was the first keystroke logging computational tool that elicits objective behavioral data of the translation process (Jakobsen and Schou 1999; Carl 2012). Many studies have used the tool to investigate cognitive, drafting patterns, and other psychological aspects of the translation process (e.g., Hansen 2002; Carl et al. 2012). Serbina et al. (2015) described the features of Tanslog as follows:

> It allows researchers to study intermediate steps of translations by recording all keystrokes and mouse clicks during the process of translation. Based on this behavioral data and the intermediate versions of translations, assumptions with regard to cognitive processing during translation can be made.


**Fig. 1** A Translog linear representation of the translation process that is sentence-aligned with the source and the final product in ULTC. (Color figure online)

> Analysis of translation process data helps explain the properties of translation products, describe potential translation problems and identify translation strategies.

The keystroke logging software Translog will be incorporated in developing the ULTC keystroke logged subcorpus that will align the source and translated texts with the process intermediate tokens. The user's behavioral data will be saved in two files: an XML log file and a Linear View file. Translog uses special characters in a linear representation, as a visualization option provided by the software that allows users to see all subjects' activities in chronological order (Fig. 1). The Pause Plot option represents the keystrokes and pauses to indicate fluent typing and breaks. Statistics can also be used to explore the statistical data to annotate the learner's activity.

The present project aims to fill the gap by creating an available large representative learner translator resource of translations from and into Arabic. It will integrate the previously mentioned approaches of parallel and learner corpora into one resource where translations produced by learners can be complemented with other cross-lingual data. Arabic, English, and French data can be retrieved as parallel or multilingual texts. Comparable data can be extracted from the parallel texts and linked back to original texts and foreign language writing and speaking tasks matching the same sampling method. Interpretation recordings and multimodal videos are transcribed and aligned with their parallel transcripts. The translation process, drafting patterns, and product can also be examined.

The corpus will follow state-of-the-art approaches and best practices in the taxonomies and tagging of translation errors and positive aspects (Uzar and Walinski 2001; MeLLANGE, Castagnoli et al. 2011). It will also develop its own language-specific, corpus-driven taxonomy and tagsets of translation errors and positive aspects. Moreover, consistent comprehensive metadata will be established with an orientation to translation research. The following sections present the first stages of compiling the ULTC, its design, preliminary findings, and future work.

## 4 Compiling the ULTC

The ULTC is a publicly available composite corpus resource that includes different subcorpora reflecting different translation pairs, directions, and modes. It is also comparable to non-translation components. Although the subcorpora might seem to be different in several respects, they are assumed to be shared via a common platform to enhance triangulation, comparability, replication, and representativeness. This project is intended to create a representative resource of learner translators of English/Arabic and French/Arabic. ULTC representativeness and balance are reflected in the tasks covered (e.g., graduation projects, field training tasks, course assignments, exams, and oral transcribed tasks, etc.), proficiency levels (lower intermediate, higher intermediate, advanced) and genre varieties (e.g., technical, legal, scientific, educational, political, social, medical, and journalistic, etc.). Due to the nature of PNU, the population represented in this phase can represent the performance of female learners of Arabic translation. This would provide interesting data to study the features of undergraduate female learners of translation. Nevertheless, adequate and valid generalizations to the whole population of undergraduate learners of Arabic translation cannot be made. To avoid unwanted gender bias, male and female professional translators are included in the ULTC reference corpora. Representative data produced by male learners of Arabic translation in other universities will also be included in the project's future expansion.

The project began in 2014 by representing at its initial stage the graduation projects of students majoring in translation at PNU. The size of the project at its present stage is more than 55 million-word tokens. Data collection, pre-processing, and annotation of data are expected to last until 2025. The beta version of the project is currently available at https://arabicparallelultc.com/. The corpora in this resource fall under three main categories: (1) learner translators' data, (2) learner interpreters' data, and (3) professional translators and native users' reference data. Due to technical and content-related considerations, data from the previously mentioned categories are presented in three main projects: (1) the undergraduate learner translator corpus (ULTC), (2) the undergraduate learner interpreter corpus (ULIC), and (3) the undergraduate learner translator and interpreter reference corpus (ULTIRC). The resource consists of many subcorpora following the design considerations for building learner corpora: (a) learner-related, (b) language-related, and (c) task-related (Tono 2003). Tables 1, 2 and 3 summarize the available corpora in the first phase of the project.

As for the population represented, tasks in most of the ULTC corpora have been taken from every student in the course who agreed on the use of her data for

**Table 1** ULTC available corpora in the first phase of the project

| The undergraduate learner translator corpus (ULTC) | | | | | | | |
|---|---|---|---|---|---|---|---|
| Corpus | Design and annotation | Current size (approx.) | Task type | Data type | Genre | Language(s) | Proficiency level |
| English–Arabic learner translator corpus (EALTC) | Parallel<br>Extensive searchable metadata<br>Error annotation and positive aspects tagging are in progress | 30 million running words (main corpus) | Graduation projects<br>Translation Technology and CAT projects | Source, draft and final product or source and target | Technical, legal, scientific, educational, political, social, medical, and journalistic | Bidirectional; English and Arabic with Arabic as the main target language | Levels 3, 6, and 8 from the second, third, and fourth years of the Translation Program |
| Multimodal Learner Translator Corpus (MumLTC) | Parallel<br>Extensive searchable metadata<br>Error annotation and positive aspects tagging are in progress | 5 million running words | Graduation projects and Scientific Translation course projects | Source, draft and final product aligned with the subtitled video | Technical, legal, scientific, educational, political, social, medical and journalistic | Bidirectional; English and Arabic with Arabic as the main target language | Levels 6 and 8 from the third and fourth years of the Translation Program |
| French–Arabic learner translator corpus (FALTC) | Parallel<br>Extensive searchable metadata | 3 million running words | Graduation projects | Source, draft and final product | Technical, legal, scientific, educational, political, social, medical and journalistic | Bidirectional; French and Arabic with Arabic as the main target language | Level 8 from the fourth year of the Translation Program |

**Table 1** continued

| The undergraduate learner translator corpus (ULTC) | | | | | | | |
|---|---|---|---|---|---|---|---|
| Corpus | Design and annotation | Current size (approx.) | Task type | Data type | Genre | Language(s) | Proficiency level |
| Multi-target learner translator corpus (MutLTC) | Parallel<br>Extensive searchable metadata<br>Error annotation and positive feedback tagging are in progress | 7 million running words | Specialized Translation, Scientific Translation, and Selected Translation tasks | Source, multiple learners' translations and model translation provided by the instructor for each text | Technical, legal, scientific, educational, political, social, medical, literary, and journalistic | Bidirectional; English and Arabic with Arabic as the main target language | Levels 3, 4, and 6 from the second and third years of the Translation Program |
| Comparable Learner Translator Corpus (ComLTC) | Comparable<br>Extensive searchable metadata | 5 million running words | ULTC and ULIC tasks are linked back to foreign language writing and speaking tasks | Target texts are linked to foreign language writing texts and speaking transcripts | Technical, legal, scientific, educational, political, social, medical, and journalistic | English and Arabic | Levels 8, 6, 4, and 3 from the second, third, and fourth years of the Translation Program |
| Multilingual Learner Translator Corpus (MLTC) | Parallel<br>Extensive searchable metadata | 1 million running words | Field training tasks | Source and target in two languages | Technical, legal, scientific, educational, political, social, medical and journalistic | English, Arabic and French | Level 8 from the fourth year of the Translation Program |

**Table 1** continued

| The undergraduate learner translator corpus (ULTC) | | | | | | | |
|---|---|---|---|---|---|---|---|
| Corpus | Design and annotation | Current size (approx.) | Task type | Data type | Genre | Language(s) | Proficiency level |
| Learner Translator Preface Corpus (LTPC) | Monolingual used as metadata linked back to the translation tasks in EALTC, FALTC and MumLTC | 500,000 running words | Reflective essays | Translator's preface and reflective essays about the translation experience | Self-reflective | Arabic and English | Levels 6 and 8 from the third and fourth years of the Translation Program |

**Table 2** ULIC available corpora in the first phase of the project

| The undergraduate learner interpreter corpus (ULIC) | | | | | | | |
|---|---|---|---|---|---|---|---|
| Corpus | Design and annotation | Current size (approx.) | Task type | Data type | Genre | Language(s) | Proficiency level |
| English–Arabic Learner Interpreter Corpus (EALIC) | Parallel<br>Extensive searchable metadata<br>Error annotation and positive feedback tagging are in progress | 7 million running words | Consecutive, sight and bilateral tasks and exams from Introduction to Interpreting, Consecutive and Sight and Bilateral courses | Source and target transcripts time-aligned with their audio recordings | Technical, legal, scientific, educational, political, social, medical, and journalistic | Bidirectional; English and Arabic | Levels 4, 6, and 7 from the second, third, and fourth years of the Translation Program |
| French–Arabic learner interpreter corpus (FALIC) | Parallel<br>Extensive searchable metadata | 300.000 running words | Consecutive, and sight tasks | Source and target transcripts time-aligned with their audio recordings | Technical, legal, scientific, educational, political, social, medical, and journalistic | Bidirectional; French and Arabic | Levels 6 and 8 from the, third and fourth years of the Translation Program |

**Table 3** ULTIRC available corpora in the first phase of the project

| The undergraduate learner translator and interpreter reference corpus (ULTIRC) | | | | | | |
|---|---|---|---|---|---|---|
| Corpus | Design and annotation | Current size (approx.) | Task type | Data type | Genre | Language(s) |
| English–Arabic reference corpus (EARC) | Parallel<br>Extensive searchable metadata<br>Error annotation and positive feedback tagging are in progress | 5 million running words | Published translated books and articles by professional translators | Source and target | Technical, legal, scientific, educational, political, social, medical and journalistic | Bidirectional; English and Arabic |
| French–Arabic reference corpus (FARC) | Parallel<br>Extensive searchable metadata | 100.000 running words | Published translated books and articles by professional translators | Source and target | Technical, legal, scientific, educational, political, social, medical and journalistic | Bidirectional; French and Arabic |

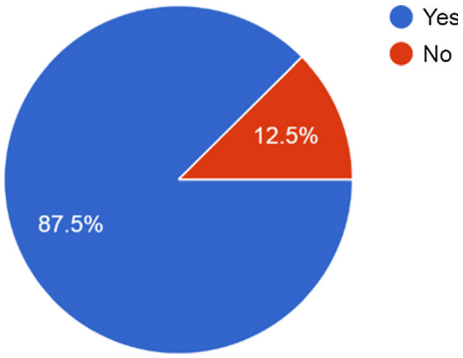


**Fig. 2** Percentage of the total number of students who opted out of the use of their translations for research purposes. (Color figure online)

research purposes. Figure 2 presents the percentage of learners who opted out of the use of their translations for research purposes. In 2014, the corpus researcher addressed the Translation Department at PNU to facilitate the process of data collection. Course instructors and coordinators collaborated in submitting all tasks produced by the students in the following courses: graduation project, introduction to interpretation, sight and bilateral interpretation, consecutive interpretation, specialized translation, scientific translation, selected translation, computer-aided translation technology (CAT), computer-aided language learning (CALL), and Field Training. It was not possible to collect data from all students in some courses as some instructors were not approachable.

### 4.1 ULTC corpora

The ULTC is designed to be a parallel bidirectional corpus. Nevertheless, Arabic is the main target language in the corpus. In addition, it is designed to be comparable in subcorpora that are linked back to ULTC target texts to writing and speaking tasks and to original texts in each language that have the same genre, topic, and communicative function. English–Arabic tasks represent the main corpus; however, French–Arabic and multilingual tasks are presented in different subcorpora. Furthermore, multimodal tasks that contain subtitled videos aligned with their parallel transcripts are in a separate subcorpus. Tasks that have multiple targets from different learners of the same source text are distinguished from those in the main corpus that have only one source and one or two targets from the same learner. The reflective essays and translator's prefaces written by the students were archived in a separate subcorpus that can be retrieved as meta-textual data linked back to the texts that are available in the main corpus. The following is a description of the available corpora in the first phase of the project.

The *English–Arabic Learner Translator Corpus* (EALTC) is the main corpus in the ULTC. It contains graduation projects and computer-aided translation technology (CAT) tasks by students in the English translation program. It presents 60% of the data in the current ULTC. Each graduation project consists of three files: a source text, the pre-edited draft, and the post-edited final product, annotated with the

grade after receiving feedback from the instructor. For multimodal projects, the same files were collected in addition to the subtitled videos. Some projects from different courses and other proficiency levels, such as the level-three Translation Technology course from the second year of a bachelor's program in Translation and the level-six CAT course from the third year of a bachelor's program in Translation, were collected following the same steps. Projects from the last two courses were translated using Trados: a translation memory system downloadable from https://www.sdltrados.com/.

Multimodal graduation projects are presented in the *Multimodal Learner Translator Corpus* (MumLTC) as a subcorpus where the source, draft, and final translations can be retrieved while aligned with the subtitled video. Moreover, the *French–Arabic Learner Translator Corpus* (FALTC) is another subcorpus comprised of the graduation projects of students in the French translation program. These three corpora incorporate the same patterns of drafting and revision of the translation tasks as they contain the source, draft, and post-edited final product. It is expected that many research questions related to translation studies can be addressed by studying the drafting patterns, errors, and positive aspects in EALTC MumLTC and FALTC. Moreover, the *Learner Translator Preface Corpus* (LTPC) is a subcorpus that contains introductory reflective essays and translators' prefaces linked back to each task in the three above corpora. In these essays, translation learners reflect on the translation process, theoretical position, critical opinion, and some reference points.

Unlike the translation tasks available in the above-mentioned corpora, the *Multi-Target Learner Translator Corpus* (MutLTC) is a multiple translation corpus that contains bidirectional translation tasks from level-three and level-four selected translation, scientific translation, and specialized translation courses from the second and third years of a bachelor's program in Translation. MutLTC consists of source text(s) aligned with their + 500 multiple targets translated by different learners. A modal translation produced by the instructor is also aligned with each source text.

The English–French–Arabic texts that are translated by trainee translators and aligned at a sentence-level are available in the *Multilingual Learner Translator Corpus* (MLTC). The trainee translators translate the same original Arabic texts into English and French in some tasks. They translate English or French translated texts into their native language Arabic in other tasks. They also translate translated Arabic tasks by their colleagues into their nonnative language. Thus, this corpus provides a unique design of presenting backtranslation of translated texts across the three languages.

Finally, the last available corpus in the ULTC project is the *Comparable Learner Translator Corpus* (ComLTC) where learner translated texts from the ULTC and ULIC corpora are compared to foreign language writing tasks and speaking transcripts produced by the same population and represented in the learner parallel corpora. Texts are linked and matched according to the sampling method (e.g., genre, gender, topic, and communicative function).

### 4.2 ULIC corpora

Interpretation data is available in a subproject where the interpreter's recordings are time-aligned with their parallel transcripts. The Undergraduate Learner Interpreter Corpus (ULIC) is a multiple translation speech corpus where interpretation recordings are aligned with their transcripts. It follows the design of the MutLTC where the source recording(s) are time-aligned with + 500 multiple target recordings and their transcripts. It is divided into two corpora: the *English–Arabic Learner Interpreter Corpus* (EALIC) and the *French–Arabic Learner Interpreter Corpus* (FALIC). Due to its expected size, diversity of interpretation tasks (i.e., consecutive, sight, simultaneous, and bilateral, etc.), and heterogeneity of data (i.e., English/Arabic, French/Arabic), the ULIC is presented as a subproject within the main ULTC project. Thus, another approach in the analysis can be performed by using the interpretation data in this subproject.

### 4.3 ULTIC corpora

As stated in the literature review, a reference corpus in the area of learner corpora can be used as a yardstick against which the learners' performances can be evaluated. The productions of native speakers and highly proficient users serve as benchmarks in learner corpus research. According to Mikhailov and Cooper (2016):

> A research corpus is a collection of data that the researcher wishes to study; a reference corpus is a collection of data compiled for the purpose of comparison… The reference corpus could be any set of data believed to be different from the research data. Also, in order to produce reliable results, it is important for the reference corpus to be much bigger than the research corpus, but sometimes a suitable reference corpus is simply not available (p. 133).

Thus, published translations by professional translators and interpreters are presented in the undergraduate learner translator and interpreter reference corpus (ULTIRC). The *English–Arabic Reference Corpus* (EARC) is the first available corpus in this subproject. It is a parallel corpus where published articles and books are sentence-aligned. The *French–Arabic Reference Corpus* (FARC) is the second available corpus, following the same design for the different language pairs—French–Arabic. The productions of native speakers will be added in the *Comparable Reference Corpus* (CRC). Interpretation data by professional interpreters will be presented in the *English–Arabic Interpretation Reference Corpus* (EAIRC). The *Multilingual Reference Corpus* (MRC) will include sentence-aligned French, Arabic, and English published texts. Because of its expected size and the plan for future expansion, the ULTIRC is presented as a subproject.

### 4.4 Annotation

One of the aims of this project is to standardize the learner translation corpora to enhance their effectiveness, reflectiveness, and resourcefulness. In the literature review, each available learner translator corpus was highlighted as having its own functions and limitations that create gaps when addressing some of the research questions related to the learner, translation task, or instructor. Therefore, the major specifications related to undergraduate learners, instructors, and the translation tasks, as opposed to post-graduate learners or professional translators, must be considered.

All learners who contributed to the project and their instructors were asked to complete a form with information about the course, learner, instructor, translation task, and source and target text such as their native language(s), age, parents' native language, GPA, level, letter grade in the project, number of years they learned English or French, and whether or not they stayed in an English/French-speaking country. Information about the tasks were also to be provided by the learners, such as title of the text, author's name, genre, subgenre, condition, source and target language, year of translation and publication, and reference tools used in the translation process. Such questionnaires are called "learner translator profiles." Written consent was also obtained from the learners to use their translation tasks for research purposes. The identity of the students and instructors was preserved.

"Learner translator profile", "instructor profile", "professional translator profile", and "transcriber profile" are four separate forms developed in this project to collect metadata about each undergraduate learner translator, instructor, professional translator, and transcriber of the interpreting recordings. In the instructor profile, instructors are asked about their major, degree, academic experience, years of experience in teaching translation, years of experience in teaching the current course, and their position. The professional translator profile has been designed for the reference corpus to elicit information about professional translators, source and target texts, and the author, including specialty, gender, nationality, level of education, number of previous translated works, and years of translation, etc. The transcriber profile is designated for the team working on the transcription of the interpretation recordings in the ULIC to obtain information that might affect the transcription quality, such as level of education, specialty, tools used in the transcription process, etc.

The resource annotation has been classified into three layers of metadata about the learner, the task, and the instructor. The corpus has been annotated with the metadata obtained from the learner and instructor profiles, and is searchable via filtering options in the bilingual query interface, based on different external and linguistic factors (Fig. 3). Extra-textual information is included in the form of a header, that is, a separate stand-off XML file making it possible to add several levels of annotation (Fig. 4).

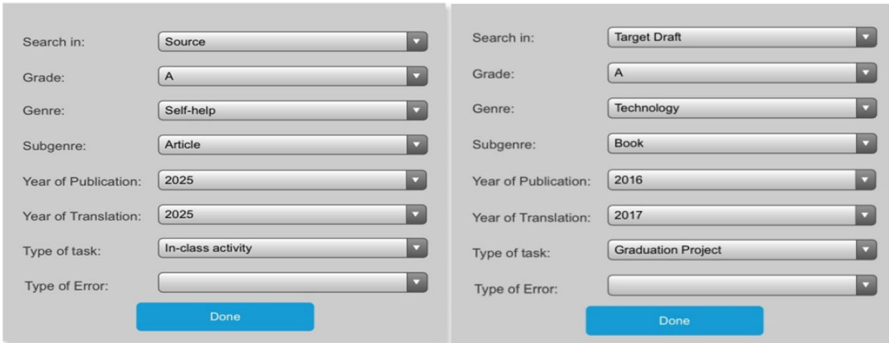


**Fig. 3** ULTC filtering options. (Color figure online)

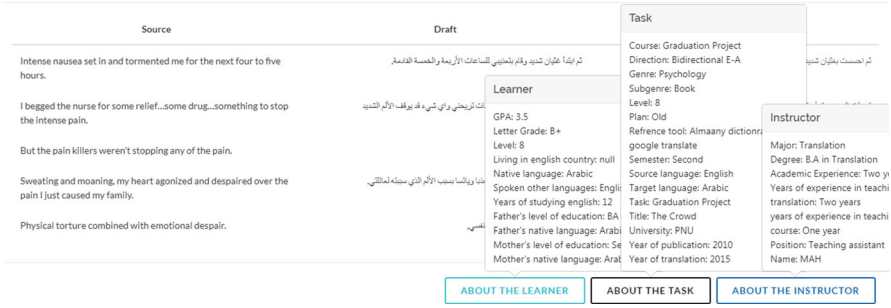


**Fig. 4** ULTC annotation layers. (Color figure online)

## 4.5 Pre-processing and alignment

The current phase involves the pre-processing of data where it needs to be cleaned from any irrelevant data, segmented, annotated, and encoded. The preprocessing database is composed of about 8700 files, containing at least 55 million running words. The EALTC, FALT, and MumLTC corpora provide two translations of each source text by the same learner. The source text, draft, and final translations in these corpora are aligned at the sentence-level (Fig. 5). The MumLTC data is comprised of transcripts of subtitled videos that are time-aligned with the data (Fig. 6).

Furthermore, no alignments are involved with ComLTC and LTPC data. They are merely coded, annotated, archived, and linked back to their matches. ComLTC data can be retrieved by the ULTC query tool, but the LTPC will appear as metadata to tasks in the EALTC, FALT, and MumLTC corpora. The alignment of MutLTC, which contains + 500 multiple translations of the same source text by different learners, has been done at the sentence-level in one file for each source text. On the other hand, MLLTC tasks are aligned and merged to be exploited in parallel as they contain two translations for the same source by different learners in different languages.

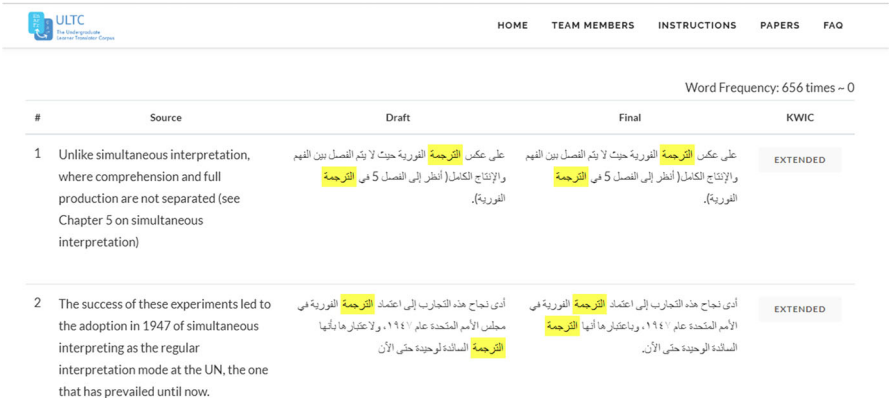


**Fig. 5** ULTC alignment. (Color figure online)

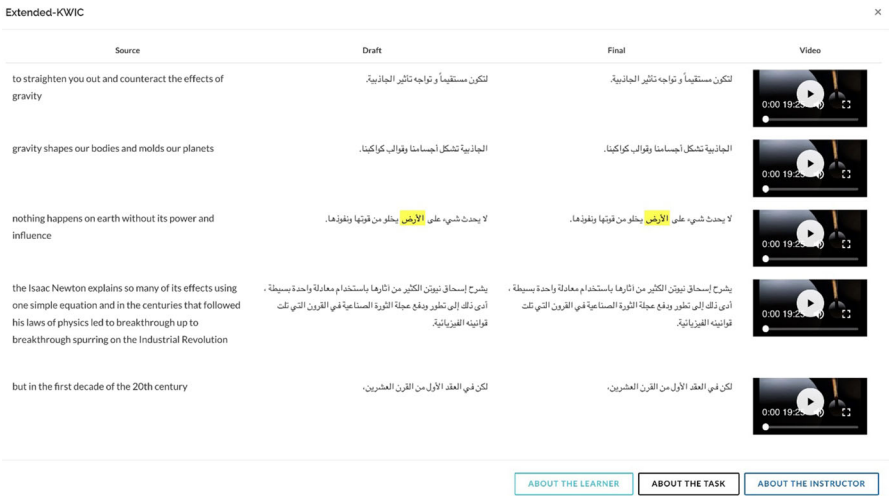


**Fig. 6** ULTC extended key word in context (KWIC) from MumLTC. (Color figure online)

At the present stage, the corpus source texts are aligned with their translations using +Align and Trados alignment tools. Nevertheless, these tools, like most of the available alignment tools, do not support multiple versions of the same text. Therefore, bitexts are merged into multitexts and errors in alignment are corrected manually. They are then converted into XML documents. ULTC data can be retrieved at the sentence-level while extended contexts (Fig. 6) can be obtained to examine the wider parallel or multilingual contexts. Users can download the aligned extended contexts as plain text files. Nevertheless, the whole parallel text cannot be retrieved or downloaded due to source text copyrights.

ULIC transcriptions and recordings are time-aligned and linked back in a systematic way. Following the same approach as used in the MutLTC, ULIC


**Fig. 7** ULIC time alignment of original speaker's and interpreters' transcripts with their recordings in the extended KWIC. (Color figure online)

multiple translations of the same source recording interpreted by different learners are aligned and merged into one document. A user can easily search the ULIC and listen to a portion of a recording that matches the query results (Fig. 7).

## 4.6 Search Interface

The available corpora can be exploited qualitatively and quantitatively by using a user-friendly interface developed for the ULTC project (Fig. 8) and easy-to-use search tools. Due to the variation in corpora design, each corpus has its own web-based query interface and search tools. In addition, the advanced search tool permits users to search all corpora simultaneously. Once the project is complete, the search interface source code can be released under a proprietary license. ULTC parallel concordancers allow users to search in the larger context(s) rapidly and reliably.

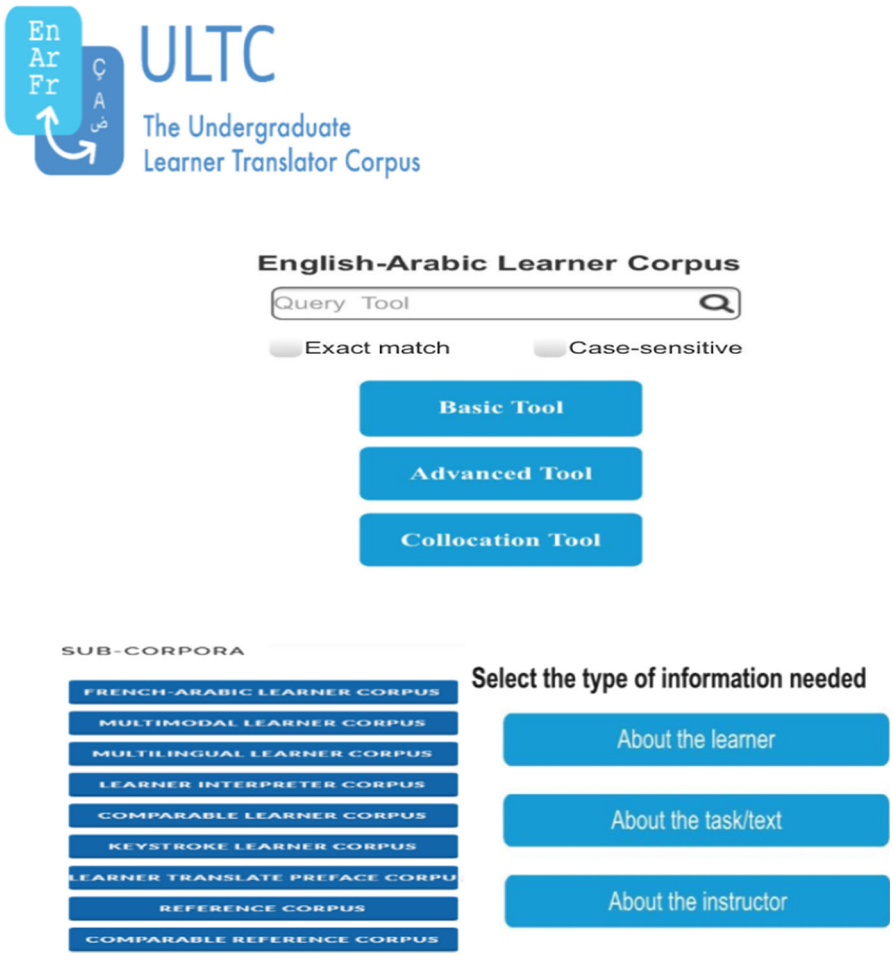


**Fig. 8** A screenshot of the ULTC interface (beta version). (Color figure online)

### 4.7 Word frequency

Word frequency is the first tool available in the ULTC. It specifies rapidly how many times the query occurs in the ULTC in general and in each subcorpus. One of the main functions of the ULTC is to provide frequency tools that enable calculating the token-type ratio of lexical diversity and density in a learner's translated texts. This can be obtained by having the number of types divided by the number of token(s). "Case-sensitive" and "exact match" options for Arabic are available in the ULTC to search for lemmatized tokens, or unlemmatized type lists to be generated by the morphological analysis and disambiguation of Arabic tool MADAMIRA 2.1 (Pasha et al. 2014).

### 4.8 Bilingual and multilingual concordancing tools

A concordancer is 'a tool that operates on a corpus by retrieving all the occurrences of a particular search pattern in its immediate contexts and displaying these in an easy-to-read format such as KWIC display' (Bowker and Pearson 2002). The ULTC allows for a basic KWIC search where the concordances appear as aligned sentences. It also provides an option of exploring the wider context of a given search by retrieving its extended KWIC. According to the different designs of the ULTC corpora, each corpus has its own concordancing tool that generates specified KWIC in different displays. A user can search for a single word, a phrase, or a string of words that will be retrieved in ULTC monolingual (e.g., LTPC), bilingual (e.g., EALTC, FALTC, or MumLTC), or multilingual (e.g., MLLTC) parallel concordancers. On the other hand, MutLTC and ULIC query results are displayed in a different style, where different concordances are retrieved and the + 500 targets are listed under their source text/transcript. Like the rest of the ULTC corpora, a user can select the extended KWIC to examine the wider context (Fig. 9).

### 4.9 Bilingual and multilingual collocation tool

Collocations are 'groups of words which frequently appear in the same context' (Oakes 1998). Collocates are types of formulaic sequences that can be bigrams or trigrams. "Bigrams are defined as sequences of two successive words; trigrams are strings of three successive words in a text" (Schmidt and Wörner 2012). Due to the diversity of the ULTC corpora, a multilingual collocation tool has been developed to allow for searches of bigrams and trigrams. A user can search for learners' translations of collocates within a predetermined span and part of speech to evaluate their strength and equivalence. It can extract all words that co-occur within a specified word span (Fig. 10). The tool can search through all corpora and sort out the results in the parallel contexts with their frequency data. It allows for bigram and trigram collocates that make it possible for n-gram's in the learner data to be

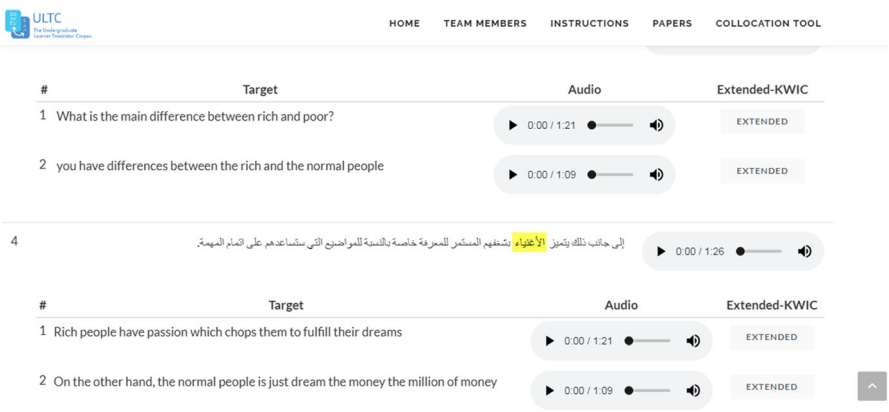


**Fig. 9** ULIC concordancing tool. (Color figure online)

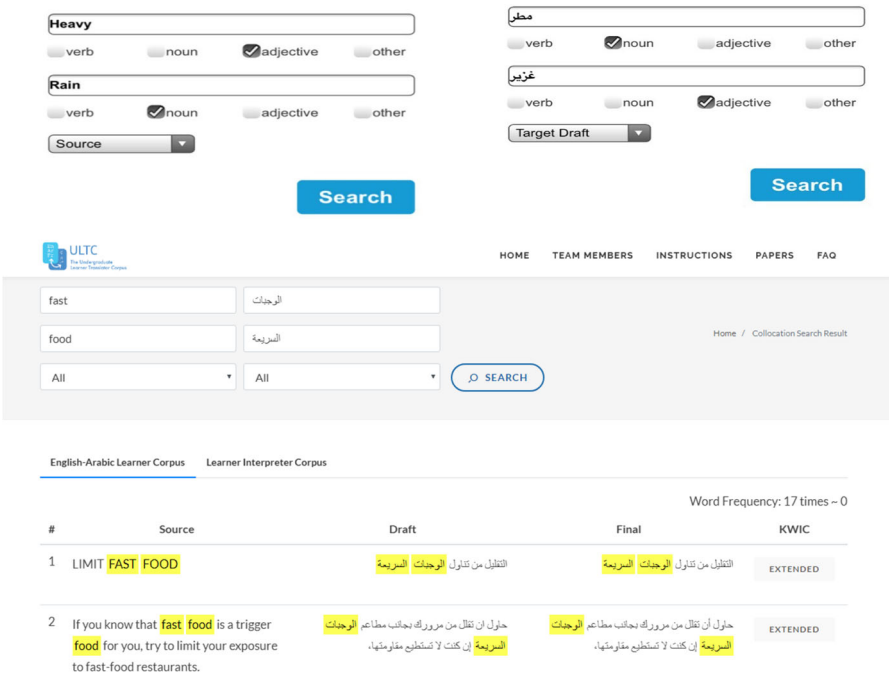


**Fig. 10** ULTC collocation tool. (Color figure online)

compared to their translated equivalents. Once the resource is part of speech tagged, a user can to choose the query part of speech. This tool is assumed to be useful for comparing between native speakers', and professional and learner translators' use of collocation.

## 5 Preliminary findings

A few preliminary findings are presented and discussed in this section as the corpus is currently being launched. Investigating structural differences between source and target segments is a major area of interest within the field of contrastive analysis and translation studies (e.g., Al-Jarf 2007; Al-Momani 2010; Abu Shquier and Abu Shqeer 2012). This section investigates the extent to which the English foreign language word order subject–verb–object (SVO) can influence the production of the learner's native language—Arabic. Verb–subject–object (VSO) is the basic word order in Standard Arabic (SA).

> For traditional Arab grammarians, VSO is the normal syntactic word order. According to generative grammar, VSO is the basic word order and SVO is derived through subject movement. The VSO order is unmarked for focus, emphasis, and information distribution (Abu Shquier and Abu Shqeer 2012).

### 5.1 Methodology

A sample of English verbal source sentences from MutLTC and EALIC has been analyzed as these corpora have the same current size and present multiple targets by different students of the same source segments. Five verbal source sentences from MutLTC, a multiple translation subcorpus that contains written translations, have been randomly chosen from five different texts from different genres (political, medical, and social). The same number of verbal sentences has been selected from the EALIC; a speech corpus of interpreting transcripts (Tables 4, 5). For each source sentence, 80 random target segments from the bilingual concordancers were

**Table 4** English verbal source sentences from MutLTC

| MutLTC source verbal sentences | Percentage of SVO target structure (%) | Percentage of VSO target structure (%) |
|---|---|---|
| 1. Saudi foreign minister Adel Al-Jubeir discussed Syria's future with Russian foreign minister Sergei Lavrov in Moscow on Wednesday | 21.5 | 78.5 |
| 2. The Human Resources Department helps support the continued development of the company through providing it with distinctive employees | 31.42 | 68.58 |
| 3. The red blood cells donate the oxygen to the cells and pick up the carbon dioxide produced by the cells. | 14.29 | 85.71 |
| 4. Education frequently takes place under the guidance of others | 50 | 50 |
| 5. It ensures that sensitive information is only disclosed to authorized parties (confidentiality) | 77.15 | 22.85 |

**Table 5** English verbal source sentences from EALIC

| EALIC source verbal sentences | Percentage of SVO target structure (%) | Percentage of VSO target structure (%) |
|---|---|---|
| 1. Parkinson's involves the malfunction and death of vital nerve cells in the brain, called neurons | 80 | 20 |
| 2. The well-known US coffee shop Starbucks just opened its largest store in the land of tea: China! | 97 | 3 |
| 3. Stress affects everyone | 90 | 10 |
| 4. The Obama administration put all this data in one place as a way of providing answers to questions about higher education | 95 | 5 |
| 5. Our bones do most of their growing when we are young so it's especially important for children to get plenty of calcium so that we grow up with strong bones | 96.25 | 3.75 |

**Table 6** EARC source verbs

| EARC source verbs | Percentage of SVO target structure (%) | Percentage of VSO target structure (%) |
|---|---|---|
| 1. Discussed | 8 | 92 |
| 2. Helps | 11.5 | 88.5 |
| 3. Donate | 56 | 44 |
| 4. Takes place | 62 | 38 |
| 5. Ensures | 24 | 76 |

analyzed manually and the number of the times the learners used the deviant SVO sentence structure was compared to the number of times they made the necessary shift to the VSO sentence structure.

To enhance the comparison, the verbs from the MutLTC chosen sentences, "*discussed*, *helps*, *donate*, *takes place,* and *ensures,*" have been compared to 80 random available examples of them from the same genres in the English–Arabic reference corpus, which contains English/Arabic published written translations by professional translators (Table 6).

## 5.2 Results

The deviant SVO structure occurred in the results of many target segments. This supports the findings of previous research that explored the negative transfer from English as a foreign language into Arabic as a native language (e.g., Al-Jarf 2007; Al-Momani 2010). An average of 38.87% of the examined MutLTC translations misplaced the verb after the subject. Nevertheless, the deviation from the unmarked VSO structure increased dramatically in the investigated EALIC speech transcripts as they included an average of 91.65% deviant segments, which is higher than the MutLTC targets. Figure 11 compares the averages of using the two structures in both corpora.

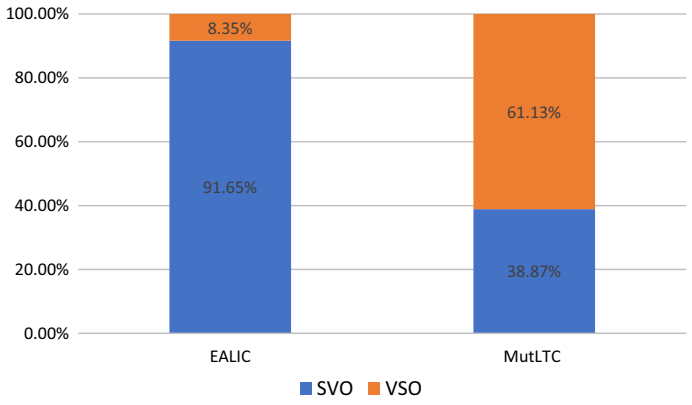


**Fig. 11** Occurrence of the deviant SVO in EALIC and MutLTC. (Color figure online)

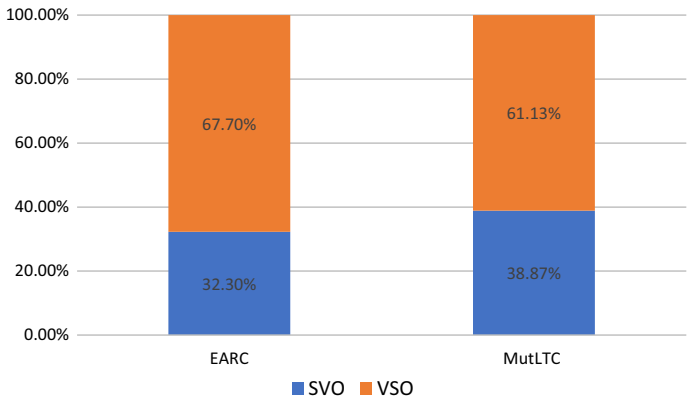


**Fig. 12** Occurrence of the deviant SVO in EARC and MutLTC. (Color figure online)

On the other hand, the deviant SVO structure occurs in 32.30% of the investigated examples from the published translations in the EARC, compared to 38.87% from the MutLTC. Both corpora contain English/Arabic written translations from the same genres. The results (Fig. 12) indicate that the use of the unmarked VSO structure in the EARC is higher than that of the MutLTC. Nevertheless, the VSO structure was often found in the EARC and the MutLTC.

### 5.3 Discussion

As a possible explanation for the deviation from the unmarked structure in both learner corpora, the strategy of literal translation by the learners as they translated the sentence directly word by word following the source structure without considering its overall meaning may have been misused. The phenomenon of diglossia may also explain the highest ratio of the deviant structure in the EALIC transcripts. Thus, the use of two distinct varieties of the same language by the same speaker may have been according to the social context (Ferguson 1959). In dialectal spoken Arabic, SVO is the basic word order. Arabic dialects are "grammatically and lexically less complex, have an exclusively oral form and are hardly ever written" (Horn 2015). The gap between Standard Arabic and the colloquial varieties is assumed to cause a challenge for students of translation when they translate a Standard Arabic task, which leads to lower learning achievements because of the diglossic switching as their unmarked spoken variety is always dialectical.

Furthermore, some insights can be obtained by analyzing the results from comparing the MutLTC and EARC examples. Although the results show that EARC published translations do not contain a high percentage of deviation from the unmarked VSO structure, they still reflect the existence of the deviant structure in 32.30% of the examined data. This suggests that the use of deviant structure could be because of the process of nominalization or grammatical metaphor, which is a common characteristic of written discourse; "that is, where actions and events are presented as nouns rather than as verbs" (Paltridge 2012). Moreover, published

translations may contain examples of interference as do learners' translations. Therefore, translation error driven from the reference corpora will be taxonomized and tagged. Nevertheless, a detailed analysis of the structural differences is beyond the scope of this manuscript and will be left for future research.

## 6 ULTC timeline and future work

The present project is expected to last until 2025. Future data collection will be extended to cover the other gender in addition to other universities and countries as ULTC corpora are constantly expanding. Because manual revisions of alignment and annotations are time-consuming, an automatic tool will be developed to support Arabic language translation resources. Consequently, 2014–2025 has been chosen as the time-frame for the ULTC (Fig. 13).

The focus of the next stage will be on POS tagging of Arabic data using MADAMIRA 2.1 (Pasha et al. 2014). In addition, different corpus-driven error taxonomies will be created and tagged, as error typologies for translation from and into Arabic are not available for pointing out problematic areas in the translations of learner translators, learner interpreters, and professional translators from and into Arabic. The existing framework will be considered for developing ULTC positive and error tagsets (MeLLANGE, Castagnoli et al. 2011). Inspired by the PELCRA LTC competence model (Uzar and Walinski 2001), positive practices will be taxonomized and tagged to highlight the competence and creativity indicators in the performances of undergraduate learners of translation from and into Arabic. PNU translation instructors and experts in the field will be invited to collaborate on the best error and competence taxonomies for written and oral translations from and into Arabic. The corpus will be fully tagged and the phase of error tagging will last until the end of 2025.

In the near future, a keystroke logging learner translator subcorpus that provides documentation of learner translators' behavioral data and drafting and revision patterns in the translation process will be launched, incorporating the most recent version of Translog II (Carl 2012). Moreover, the learners' translated tasks will be compared to non-translation tasks produced by learners of translation in the learner comparable corpus. Another important project will be to build the reference multilingual corpus to represent the performance of professional translators from

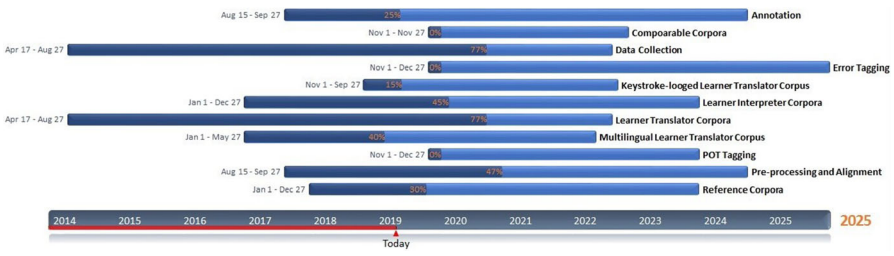


**Fig. 13** ULTC timeline. (Color figure online)

and into Arabic. The last stage will be to compile the comparable reference corpus that will present original and translated texts by native writers and professional translators.

## 7 Conclusion

This paper introduces the ULTC as an ongoing composite corpus resource with three main projects that include many subcomponents reflecting different translation pairs (Arabic, English, French), directions, and modes. It includes learner translator corpora, interpreting corpora, reference corpora, and comparable non-translation components. The corpus is intended to be a representative resource with many applications such as translation pedagogy, research on the translation process, machine translation, lexicography, and contrastive and comparative linguistics. Most of the available corpora are parallel; that is, sentence-aligned source texts with their translations. Furthermore, multimodal and interpreting corpora exist, with transcripts aligned with video takes containing subtitles or audio files. The ULTC focuses on providing extensive metadata, both on internal information (text type and field of specialization) and external factors in acquiring translation-related skills (foreign-language skills, course-level, grade, instructor, etc.). This paper reports on the development of ULTC analysis tools and discusses some technical solutions to the complexities of retrieving results from different available subcomponents, such as making multiple versions ready for bilingual concordance searches. The ULTC aims to create a large-scale error-annotated corpus of English-to-Arabic and French-to-Arabic translations.

Overall, all texts collected so far are from female students as PNU is a public women's university located in Riyadh, Saudi Arabia. Many corpus-driven/based studies can be explored using ULTC data such as the difficulties of Arabs dealing with another language and culture. This resource is expected to inform textbooks on what should be presented initially to undergraduate translation learners at early stages to be proficient and skillful. It is also expected to assist teachers on what should be tested and considered in translation and language testing. Researchers from different backgrounds are invited to use this resource to develop different potential computational applications and tools, as in automated scoring, native language identification, and neural, statistical or rule-based machine translation systems. Dictionary and lexicon builders are also welcome. The project may seem ambitious, but it is an attempt to fill a perceived gap in Arabic translation resources. It is hoped that it will provide interesting data for researchers in translation from and into Arabic.

**Acknowledgements** The author would like to thank the anonymous reviewers for the detailed and constructive review that helped to clarify many points and improve the structure of the manuscript. The author is greatly indebted to PNU instructors, course coordinators, and learners for their contributions.

## References


Abu Shquier, M. M., & Abu Shqeer, O. (2012). Words ordering and corresponding verb-subject agreements in English–Arabic machine translation, An enhancement approach. *The International Arab Journal of Information Technology (IAJIT), 2,* 49–60.

Afli, H., Lohar, P., & Way, A. (2017). MultiNews: A web collection of an aligned multimodal and multilingual corpus. In *Proceedings of the first workshop on curation and applications of parallel and comparable corpora.* Taipei, Taiwan.

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

Al-Jarf, R. (2007). SVO word order errors in English–Arabic translation. *Translators' Journal, 52,* 299–308.

Al-Momani, I. (2010). Does the VP node exist in Modern Standard Arabic? *Journal of Language and Literature*, ISSN: 2078-0303, May 2010.

Alotaibi, H. M. (2017). Arabic–English parallel corpus: A new resource for translation training and language teaching. *Arab World English Journal, 8*(3), 319.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 17–45). Amsterdam: John Benjamins.

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

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

Bowker, L., & Peter, B. (2003). Student translation archive and student translation tracking system: Design, development and application. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in translator education* (pp. 103–119). Manchester: St Jerome Publishing.

Carl, M. (2012). Translog-II: A program for recording user activity data for empirical reading and writing research. In *Proceedings of the eighth international conference on language resources and evaluation, European Language Resources Association (ELRA)*, Istanbul, Turkey. http://www.lrec-conf.org/proceedings/lrec2012/summaries/614.html.

Carl, M., Bangalore, S., & Schaeffer, M. (2015). *New directions in empirical translation process research: Exploring the CRITT TPR-DB*. Cham: Springer. **(New Frontiers in Translation Studies)**.

Carl, M., & Dragsted, B. (2012). Inside the monitor model: Process of default and challenged translation production. *Translation: Corpora, Computation, Cognition, 2*(1), 127–145. **(Special issue on the crossroads between contrastive linguistics, translation studies and machine translation)**.

Carl, M., Dragsted, B., Elming, J., Hardt, D., & Jakobsen, A. L. (2012). The process of post-editing: A pilot study. In B. Sharp, M. Zock, M. Carl, A. L. Jakobsen (eds.), *Proceedings of the 8th natural language processing and cognitive science workshop* (Copenhagen studies in language series, Vol. 41, pp. 131–142).

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

Castagnoli, S., Ciobanu, D., Kunz, K., Volanschi, A., & Kübler, N. (2011). Designing a learner translator corpus for training purposes. In N. Kübler (Ed.), *Corpora, language, teaching, and resources: From theory to practice* (pp. 221–248). Bern: Peter Lang.

Cettolo, M. (2016). An Arabic–Hebrew parallel corpus of TED talks. In *Proceedings of the AMTA 2016 workshop on Semitic machine translation (SeMaT).* Austin, US-TX.

Dimitriu, R. (2009). Translators' prefaces as documentary sources for translation studies, Perspectives. *Studies in Translatology, 17*(3), 193–206.

Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources & Evaluation, 48,* 33.

Ferguson, C. A. (1959). Diglossia. *Word, 15,* 325–340.

Florén, C. (2006). ENTRAD, an English Spanish parallel corpus created for the teaching of translation. Paper presented at *the 7th teaching and language corpora conference (TALC 2006).*

Fung, P., & Cheung, P. (2004). Mining very-non-parallel corpora: Parallel sentence and lexicon extraction via bootstrapping and em. In *Proceedings of EMNLP*, vol. 2004.

Graedler, A. L. (2013). Nest – a corpus in the brooding box. In M. Huber & J. Mukherjee (Eds.), *Corpus linguistics and variation in English: Focus on non-native Englishs*. Studies in Variation, Contacts and Change in English, University of Giessen.

Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung, S. Petch-Tyson (eds.), *Computer learner corpora, second language acquisition and foreign language teaching.* Amsterdam & Philadelphia: Benjamins.

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.

Hansen, G. (Ed.). (2002). *Empirical translation studies: Process and product* (Copenhagen studies in language, vol. 27). Denmark: Samfundslittera-tur.

Hewavitharana, S., Vogel, S. (2011). Extracting parallel phrases from comparable data. In *Proceedings of the 4th workshop on building and using comparable corpora: Comparable corpora and the web* (pp. 61–68). Association for Computational Linguistics.

Horn, C. (2015). Diglossia in the Arab world. *Open Journal of Modern Linguistics, 5,* 100–104.

Hu, K., & Tao, Q. (2013). The Chinese–English conference interpreting corpus: Uses and limitations. *Meta, 58*(3), 626–642. https://doi.org/10.7202/1025055ar.

Izquierdo, M., Hofland, K., & Reigem, Ø. (2008). The ACTRES parallel corpus: An English–Spanish translation corpus. *Corpora, 3*(1), 31–41.

Izwaini, S. (2003). Building specialised corpora for translation studies. In *Workshop on multilingual corpora: Linguistic requirements and technical perspectives, corpus linguistics.* (pp. 17–25). , Lancaster University, UK. http://www.coli.uni-sb.de/muco03/izwaini.pdf.

Jakobsen, A. (2003). Effects of think aloud on translation speed, revision and segmentation. In F. Alves (Ed.), *Triangulating translation: Perspectives in process oriented research* (pp. 69–95). Amsterdam: Benjamins.

Jakobsen, A. L. (2011). Tracking translators' keystrokes and eye movements with Translog. In C. Alvstad, A. Hild, & E. Tiselius (Eds.), *Methods and strategies of process research integrative approaches in translation studies* (pp. 37–55). Amsterdam: John Benjamins Publishing.

Jakobsen, A. L., & Schou, L. (1999). Logging target text production with Translog. *Copenhagen Studies in Language* (Vol. 24, pp. 9–20). Copenhagen: Samfundslitteratur.

Kumar, G., Cao, Y., Cotterell, R., Callison-Burch, C., Povey, D., & Khudanpur, S. (2014). *Translations of the CALLHOME Egyptian Arabic corpus for conversational speech translation.* IWSLT.

Kutuzov, A., & Kunilovskaya, M. (2014). Russian learner translator corpus. In P. Sojka, A. Horak, I. Kopecek, & K. Pala (Eds.), *Text, speech and dialogue (Lecture Notes in Computer Science)* (Vol. 8655, pp. 315–323). Berlin: Springer.

Li, X., et al. (2013). *GALE Arabic-English parallel aligned treebank – broadcast news*. Part 1 LDC2013T14. Web Download. Philadelphia: Linguistic Data Consortium.

McEnery, A. M., & Xiao, R. Z. (2007). Parallel and comparable corpora: What are they up to? In G. Anderman, & M. Rogers (Eds.), *Incorporating corpora: Translation and the linguist.* Retrieved from http://eprints.lancs.ac.uk/59/.

Mesa-Lao, B. (2014). Gaze behavior on source texts: An exploratory study comparing translation and post-editing. In S. O'Brien, L. W. Balling, M. Carl, M. Simard, & L. Specia (Eds.), *Post-editing of machine translation* (pp. 219–245). Newcastle Upon Tyne: Cambridge Scholar Publishing.

Mikhailov, M., Cooper, R. (2016). *Corpus linguistics for translation and contrastive studies: A guide for research.* Routledge. Corpus Linguistics Guides. London & New York: Routledge.

Norberg, U. (2014). Fostering self-reflection in translation students. *Translation & Interpreting Studies, 9*(1), 150–164.

Oakes, M. (1998). *Statistics for corpus linguistics*. Edinburgh: Edinburgh University Press.

Paltridge, B. (2012). *Discourse analysis* (2nd ed.). London: Bloomsbury.

Pasha, A., Al-Badrashiny, M., Diab, M., El Kholy, A., Eskander, R., Habash, N., Pooleery, M., Rambow, O., & Roth, R. M. (2014). *MADAMIRA: A fast, comprehensive tool for morphological analysis and disambiguation of Arabic*. Language Resources and Evaluation Conference (LREC 2014).

Rafalovitch, A., & Dale, R. (2009). United Nations General Assembly resolutions: A six-language parallel corpus. In *Proceedings of the MT summit XII*. (pp. 292–299, Ottawa, Canada).

Russo, M., Bendazzoli, C., Sandrelli, A., & Spinolo, N. (2012). The European parliament interpreting corpus (EPIC): Implementation and developments. In S. F. Straniero & C. Falbo (Eds.), *Breaking ground in corpus-based interpreting studies* (pp. 53–90). Frankfurt am Main: Peter Lang.

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

Schmidt, T., & Wörner, K. (Eds.). (2012). *Multilingual corpora and multilingual corpus analysis* (p. 407). Amsterdam/Philadelphia: John Benjamins.

Serbina, T., et al. (2015). Development of a keystroke logged translation corpus. In C. Fantinuoli & F. Zanettin (Eds.), *New directions in corpus-based translation studies* (pp. 11–34). Berlin: Language Science Press.

Shlesinger, M. (2008). Towards a definition of interpretese: An intermodal, corpus-based study. In G. Hansen, A. Chesterman, & H. Gerzynisch-Arbogast (Eds.), *Efforts and models in interpreting and translation research* (pp. 237–253). Amsterdam/Philadelphia: John Benjamins.

Smith, J. R., Quirk, C., & Toutanova, K. (2010). Extracting parallel sentences from comparable corpora using document level alignment. In *Human language technologies: The 2010 annual conference of the North American chapter of the Association for Computational Linguistics*, pp. 403–411. Association for Computational Linguistics.

Sosnina, E. P. (2006). *Development and application of Russian translation learner corpus*. St. Petersburg: Papers from the Corpus Linguistics Conference.

Stefanescu, D., Ion, R., & Hunsicker, S. (2012). Hybrid parallel sentence mining from comparable corpora. In *Proceedings of the 16th conference of the European Association for Machine Translation* (pp. 137–144).

Štěpánková, K. (2014). *Learner translation corpus: CELTraC* (Bachelor's thesis).

Temnikova, I., Abdelali, A., Hedaya, S., Vogel, S., & Al Daher, A. (2017). *Interpreting strategies annotation in the WAW corpus. RANLP*, p. 36.

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.

Tono, Y. (2003). Learner corpora: Design, development and application. In *Proceedings of the corpus linguistics 2003 conference* (pp. 800–809). Lancaster, UK, 28–31 March 2003.

Uzar, R., & Walinski, J. (2001). Analyzing the fluency of translators. *International Journal of Corpus Linguistics, 155*(166), 12.

Wurm, A. (2013). Eigennamen und Realia in einem Korpus studentischer Übersetzungen (KOPTE); in: *transkom, 6*(2); 381–419. http://trans-kom.eu.

Xiao, R., & McEnery, T. (2002). A two-level approach to situation aspect. Paper presented at the *5th chronos colloquium on tense, aspect and modality*, Groningen, Netherlands.

Zaidan, O. F., & Callison-Burch, C. (2014). Arabic dialect identification. *Computational Linguistics, 40*(1), 171–202.