

## The Design and Development of an Error Tagging Tool Using the Undergraduate Learner Translator Corpus Error Taxonomy

Noha M. El-Jasser

Received: 05 November 2024

Accepted: 06 December 2024

**Abstract** Learner translator corpus is a promising area in translation studies. However, they require substantial, collaborative, and continuous work to be designed, developed, and exploited. One of the resources in this area is the Undergraduate Learner Translator Corpus (ULTC). It is a parallel, trilingual, bidirectional, and multimodal corpus with more than 55 mln word-tokens. It comprises a main corpus and sub-corpora. To make the corpus more useful, this paper describes the tool designed and developed for error-tagging the ULTC. The tool is web-based, user-friendly, editable, and manageable. A pilot taxonomy has been developed for tagging erroneous as well as positive choices. The designed tagger is currently applied on the main corpus En-ArLTC as a pilot project, to make certain the tool is valid. The data comprises graduation projects, mainly from English into Arabic, produced by undergraduates at Princess Nourah bint Abdulrahman University, SA. This project contributes to filling the gap in representing En-Ar error-tagged translations as there is a lack of available Learner Translator Corpora where Arabic is involved as a pair of languages. The current paper reveals the encouraging attempts towards the application of the error-tagging tool. It yields, however, the necessity of inter-rater reliability. <https://arabicparallelultc.com/>

**Keywords:** learner translator corpora – error-tagging tool – error-annotation – metadata – tagset

---

✉ \*Noha Mohammad El-Jasser  
nmeljaseer@pnu.edu.sa

Department of Translation, College of Languages, Princess Nourah bint Abdulrahman University,  
Riyadh, Saudi Arabia

## 1. Introduction

In the field of translation, the number of available learner translator corpora (LTC) is finite. It becomes even scarce when considering error-tagged corpora. In the case of the Arabic language, there is a lack of representation of the Arabic language on LTCs. The Undergraduate Learner Translator Corpus (ULTC) is the first LTC available with a representative size that deals with the Arabic language (See 2.1.1). This paper aims to describe the error-tagging tool used in the ULTC. According to Granger and Lefer, corpora become more useful when annotated with linguistic information, either automatically (e.g., part of speech tagging) or manually (e.g., semantic features and error tagging) (Granger, S and Lefer, M. -A 2023).

Learner Corpus denotes two types (Štěpánková 2014). First, they refer to texts produced by EFL learners. In this type, it is a resource for analyzing and assessing students' writings in L2. The second type is the LTC, which refers to the translations produced by translation students or trainees where translations are aligned with their corresponding Source Texts (STs). Some LTCs combine both types in the same corpus (PELCRA, CELTraC, and DiHuTra). This paper focuses on translations of students that are aligned with their respective STs. The language pair is English and Arabic, most of which are from L2 >L1. The tagger tool is applied on graduation projects where a student translates a passage of text of around 5,000-word count. The texts' genera are various and the texts' content includes both general and technical writings (Alfuraih and El-Jasser 2024). The corpus lists two versions of translation for a single project: the pre-edited translation and post-edited translation. It traces the phases of the translation process: the first production of texts without the guidance of the supervisor and the final version of translation after discussion with the supervisor. The tagging tool is designed to target both the drafts and the final versions. A taxonomy of erroneous and positive tags has been developed to tag the English–Arabic Learner Translator Corpus (EALTC). Just as identifying errors plays a vital role in improving students' performance, so does identifying positive choices as it highlights students' competence and creativity. The developed tagger tool is customized and incorporated into the corpus website. The corpus data is not used as a means of teaching methods, evaluating students' performances, or assigning grades. In fact, the main purpose of developing an error taxonomy and an error-tagging tool is to inspire and conceptualize didactic and research insights for researchers in the field, instructors, and students.

In this paper, the expressions *error-tagging*, *error annotation*, and *error taxonomy* are used interchangeably to refer to both errors as well as positive tags.

In the literature of Learner Corpora, the term annotation refers to either the linguistic annotations of a corpus or a translation error annotation. Linguistic annotation involves linguistic mark-up of some linguistic features such as part of speech tagging (PoS), lemmatization, tokenization, word tagging, morphosyntactic information, and so on (Mikhailov and Cooper, 2016). Translation error annotation, however, refers to the use of corpus annotation to mark-up translation and linguistic errors. The process of annotation in this paper refers to error annotation of students' translations.

The remainder of the paper is structured as follows: section two explores error-tagged LTCs. The third section introduces the preliminary taxonomy developed for error annotating the EALTC. Section four describes the error tagger tool and section five conceptualizes the potential opportunities drawn from error-tagging translations.

## 2. Error-tagging in learner translator corpus

### 2.1 Related studies

The LTC as a field has emerged from a hybrid of two other fields: learner corpora research and corpus-based translation studies (Granger, S and Lefer, M. -A 2023). The developed LTCs vary in corpus type, size, design, language pairs, collected data, and online availability. There are a number of LTCs such as the *Polish and English Language Corpora for Research and Applications* (PELCRA LTC; Uzar and Walinski, 2001), the *Student Translation Archive* (STA; Bowker and Bennison, 2003), the *Russian Translation Learner Corpus* (RuTLC; Sosnia 2006), the *Multilingual eLearning in LANGuage Engineering Learner Translator Corpus* (MeLLANGE LTC; Kübler, 2008), the *Enseñanza de la Traducción* (ENTRAD; Florén, 2006), the *Multiple Italian Student Translation Corpus* (MISTiC; Castagnoli, 2009), the *Norwegian-English Student Translation Corpus* (NEST; Graedler, 2013), *VARiation in TRAnslation* (VARTRA; Lapshinova-Koltunski 2013), the *Universitat Pompeu Fabra (Barcelona) – Learner Translation Corpus* (LTC-UPF; Espunya, 2014), the *Czech-English Learner Translation Corpus* (CELTraC; Štěpánková, 2014), the *Russian Learner Translator Corpus* (RusLTC; Kutuzov and Kunilovskaya, 2014), *Korpusprojekt zur Translationsevaluation* (KOPTE; Wurm 2016), *The Czech-English Learner Translation Corpus* (CELTraC:English into Czech TLC; Fictumova et al., 2017), the *Italian-Greek learner translator Corpus* (Italogreco; Katerina Florou, 2019). The *Undergraduate Learner Translator Corpus* (ULTC; Alfuraih, 2020), the *Multilingual Student Translation Corpus* (MUST; Granger and Lefer 2020), and *Differences between Human Translations* (DiHuTra; Lapshinova-Koltunski et al., 2022).

Following the scope of this paper, we will consider the error-tagged LTCs and will have an overview of the mechanisms followed for annotating errors.

#### 2.1.1 An overview on ULTC

The ULTC is a massive learner translator corpus with over 55 million word tokens (Alfuraih 2020). It is a parallel, trilingual, bidirectional, and sentence-aligned corpus. It includes a collection of corpora, and it is designed with a main corpus and complementary sub-corpora. It is composed of texts translated by translation department undergraduates, mainly graduation projects. The students are female undergraduates at Princess Nourah bint Abdulrahman University in Saudi Arabia. Arabic is the main language as it is the students' mother tongue. It is the language paired with English or French. Graduation projects are classified into four corpora: the main corpus and 3 sub-corpora. The methodology used in designing the corpora

is unique in that it presents the original text, the pre-edited translation, and the final version of the translation. The corpora are sentence-aligned, i.e., every sentence in the original text is aligned to its equivalent in the target text (the pre-edited translation as well as the post-edited translation). The corpus is not limited to representing the graduation projects only, but also provides the researcher with metadata, such as the translator's preface (It is a preface where students describe their graduation project, the main challenges, and how they overcome those challenges) as well as the student's foreign language acquisition background. The length of a single graduation project ranges between 2,500 and 5,000 words, depending on the graduation project module. The main corpus is called the EALTC. It is a bidirectional, parallel, and sentence-aligned corpus that comprises graduation projects of bachelor students of the English Translation Department. It is the main corpus in the ULTC project as it includes + 25 million word-tokens. Because the module of some of the graduation projects is audio-visual, they were placed in a separate sub-corpus called the Multimodal Learner Translator Corpus. It includes subtitled video clips in addition to their transcriptions of the original extracts, the draft translation, and the final version of the translation. The French–Arabic Learner Translator Corpus is similar to EALTC but the language pair is French–Arabic. The last sub-corpus is the Preface Learner Translator Corpus which includes the preface of the graduation projects, which were written by students about their experience in the journey of translating the graduation project and mentioning the most significant challenges, and ways to overcome them and the skills they learned. Each preface is linked to its graduation project.

Apart from graduation projects, the ULTC project includes sub-corpora as follows. The multi-target Learner Translator Corpus, where displays the original text aligned with its multiple translations. The texts are assignments performed by students in various translation courses and are shorter than graduation projects, as the original text does not exceed 600 words. Another sub-corpus is called the multilingual corpus. It includes an original text in Arabic and its translation into English and French. The ULTC project includes the Multilingual Learner Translator Corpus (MLTC), a corpus of texts written by native speakers of a language. Its purpose is to compare the works written by native speakers of a language with the works written by foreign students. There is also a corpus called the Comparable Learner Translator Corpus, where the researcher can compare the translated student texts available in the main corpus with texts written by native speakers of the language. The last subcorpus is the Undergraduate Learner and Interpreter Reference Corpus, which comprises translations performed by professional translators. Its purpose is to benchmark against learners' performance and evaluation.

As mentioned earlier, the error-tagging system described in this paper is to be applied to the ULTC project. It will target the graduation projects available on the EALTC corpus in the first phase.

### *2.1.2 Error-tagged LTCs*

Early resources of LTCs have emerged with the spread of electronic data at the beginning of 2000s. The main purpose of LTCs was to serve pedagogical purposes by providing elaborate and systematic error analysis of students' translations (PELCRA,

STA, RuTLC, ENTRAD, and MISTiC). In terms of error typologies, almost all LTCs have developed typologies to systematize the process of error annotation. The annotation process was manual and mainly done by teachers to feedback to students on errors. No software was available for error-tagging. In the case of the ENTRAD project, for instance, it was incorporated into translation classes for teaching and evaluation. Though teachers marked errors by using colored codes of the proposed taxonomy, annotations were not machine readable. Teachers had to print the tagged translation texts so that students became aware of their errors. In PELCRA LTC project, though the error typology included a set of positive tags for interesting choices, the typology is from EFL field, not the translation pedagogy field (Espunya 2014). Some LTCs focus on specific features and no error typology was developed (STA, MISTiC, NEST, VARTARA, DiHuTra). The MeLLANGE LTC was the first LTC to introduce a proper typology of 30 tags with a customized version of MMAX2, a manual annotation tool. More LTCs have adopted the method of developing tagging software to facilitate the process of annotating errors. Table 1 showcases that there are 5 LTCs that have developed tagging software for annotating errors (MeLLANGE, RuTLC, CELTraC, ULTC, MUST). In the case of ULTC, an error-tagger tool has been developed. The table also demonstrates that more focus is placed on error annotation, rather than positive annotation. Positive annotation is added as a tag, not as a separate taxonomy. In this section, we will have an overview of those LTCs that have developed software or tool for error annotation:

- A. The MeLLANGE LTC is an aligned, multilingual learner translator corpus of translated texts produced by translation students as well as translation professionals (Kübler 2008). The metadata and texts are stored in the MySQL database. For translation error annotation, an error typology was developed. The hierarchical taxonomy is classified into two main categories: the content-based and the language-based. The two categories are divided into subcategories which are subdivided into error types. Each error type is marked by a code. The codes are used during the process of annotating translation errors. The corpus annotators download the translated texts and corresponding metadata to annotate the translation errors. There are some features available for annotators during the process of annotation. They can add comments or provide appropriate solutions for errors. In case of dealing with an error not classified in the taxonomy, an annotator can suggest an error type. They can mark multiple categories to specific sections of a text. Translation error annotation is executed using a customized version of the manual annotation tool MMAX2 which represents files in XML formatting. Translation error annotation involves 4 levels: the paragraph, the sentence, the content transfer, and language levels. The two former levels are generated automatically but the two latter ones are developed specifically to use the error taxonomy to annotate translation errors across the MeLLANGE corpus.
- B. The Russian Learner Translator corpus (RusLTC) is a large learner translator corpus that is composed of translations produced by Russian under/postgraduates and/or translation trainees from various Russian universities

Table 1. An overview of error-tagging across learner translator corpora

LTC name	Corpus type	Language pairs Directional ity	Linguistic annotation	Error annotation	Positive annotation	Error- tagging tool	Annotated by:	Availability	Public Accessibility
PELCRA LTC (2001)	Multiple / manual alignment	Polish (SL) English L1 > L2	PoS tagging	Manual annotation (feedback)	Positive feedback	No	-	Downloadable	No
(STA) 2003	Parallel	French (SL) Spanish (SL) English L2 > L1	POS, Translation tracking system)	Not reported	-	No	-	Unavailable	-
RuTLC 2006	Electronic/ Alignment is not reported	English (SL) Russian L2 > L1	No	Manual annotation (feedback)	No	No	-	Unavailable	-
ENTRAD 2006	Parallel/ textual alignment	English (SL) Spanish (TL) L2 > L1 L2 > L2 (A few students were French)	No	Manual annotation (feedback)	No	No	-	Online	No
MeLLANG E 2008, 2011	Comporable	SL: German, English, French, Spanish TL: Catalan, German, English, Spanish, French, Italian L2 > L1	PoS tagging, lemmatizatio n, tokenization, self-made mini- context level alignment.	Yes	No	A customized version of manual annotation tool: <i>MMAX2</i>	Annotators	Online	No

LTC name	Corpus type	Language pairs Directionality	Linguistic annotation	Error annotation	Positive annotation	Error- tagging tool	Annotated by:	Availability	Public Accessibility
MISTIC 2009	Multiple	En (SL) French (SL) Italian L2 > L1	PoS tagging	No	-	-	-	Unavailable	-
NEST 2013	Multiple	Norwegian (SL) English L1 > L2	No	No	-	-	-	Unavailable	-
Variation in Translation (VARTRA) 2013	Comparab le/not aligned	German English (SL)  L2 > L1	Yes (PoS, Lemmatized,  Segmented into syntactic chunks and sentences.	No (Out of scope)	-	-	-	Unavailable	-
RuLTC 2014	Multiple	Russian (SL) English(TL ) L2 > 1	Tokenization , POS/self- made	Yes (Hierarchy of 31 types in Content/Langua ge colour-coded Categories)	One tag is labeled as good_job	A customised <i>brat</i> text annotation program	Annotators	Online/ downloadable	Yes
CELTrac Štěpánková 2014,	Parallel	Czech (SL) English L1 > L2	PoS tagging	Yes (Manual error- tagging)	No	<i>Hypal</i> tool	Course teacher+ native proof-reader	Online	No

LTC name	Corpus type	Language pairs Directionality	Linguistic annotation	Error annotation	Positive annotation	Error- tagging tool	Annotated by:	Availability	Public Accessibility
LTC-UPF 2014	Multiple	English (SL) Catalan L2 > L1	Word & PoS tagging, lemma, fine morphologic al features and syntactic functions	Yes (The taxonomy of errors comprises 25 simple categories with no subdivisions)	Markin tool for error annotation/ web- searchable platform IAC	No	Teachers	Online	No
KOPE 2016	Multiple	French (SL) German L2 > L1	tokenization, lemmatization and POS- tagging/ No alignment	Manual annotation	UAM Corpus Tool	Yes (Positive typology)	Researchers annotate Teachers' evaluations	Unavailable	-
(English-to- Czech TLC) 2017	Multiple	Czech (SL) English L1 > L2	PoS tagging, lemmatization No alignment	Yes (Updated error typology of CELTraC's)	Yes (A filter was added to the new typology: <i>positive- feedback</i> )	New version of <i>Hypal</i> tool	Teachers	Online	No
Italian- Greek learner translator Corpus (Italogreco)	Multiple	Greek (SL) Italian (TL) Or English (TL) L1 > L2	No	Manual annotation of grammatical errors	No	No	The researcher	Unavailable	-



LTC name	Corpus type	Language pairs Directional ity	Linguistic annotation	Error annotation	Positive annotation	Error-tagging tool	Annotated by:	Availability	Public Accessibility
MUST 2020	Multiple	Various language (18) L1 > L2 L2 > L1 L2 > L2 (A possible feature)	PoS, Lemmatizati on	Yes	Yes (Plus metatag)	<i>Hypal4MUST</i> platform	Teachers	Online	
ULTC 2020	Composite (Parallel, comparabl e, multiple, reference)	Arabic English French L2 > L1 L1 > L2 (Few)	In progress	Yes	Yes	ULTC web-based tool	Professional annotators	Available	Yes <u>Corpus</u> site
DiHuTra (2022)	Comparab le	English (SL) Croatian, Finnish, Russian Not reported	PoS, Lemmatizati on, Parsed	No	-	-	-	Available	Yes

(Kutuzov and Kunilovskaya 2014). The TTs are aligned with corresponding STs using a *hunalign* library; and then manually edited using the translation memory eXchange *Okapi Olifant*.

A subcorpus of RusLTC is the En-Ru error-tagged subcorpus (Kunilovskaya 2014). It uses the *brat* program for annotation (Stenetorp et al., 2012). To annotate translations, an error typology has been developed by taking into consideration the analysis of students' errors and drawing upon experience in translation quality assessment (TQA). Content errors are classified into three hierarchical taxonomies: semantics, syntax, and pragmatics whereas language errors are classified according to established practices in foreign language education: lexical, morphological, and syntactic errors as well as spelling and punctuation. As the partners of this corpus believe in the importance of understanding the reasons for such errors, they developed two extra tag sets that enable annotators to describe the severity of mistakes (critical, major, and minor) and to allow them to reflect on potential causes of the mistake. The typology is not only confined to errors but also includes tags for creative choices. The tagset adopts colored tags; and the total number of tags is 6471, 236 of which are for positive choices.

- C. The Czech-English Learner Translation Corpus (CELTraC) is a parallel, error-tagged learner translator corpus (Štěpánková 2014). The corpus data are of two sets: written texts and parallel texts. The parallel texts are about translation assignments from Czech into English and are produced by MA students. For annotating errors, the corpus uses *Hybrid Parallel Text Aligner* (Hypal), a corpus annotation tool (Obrusnik 2013). CELTraC aims to test the Hypal tool and investigate challenges that face annotators during the process of tagging errors.

Corpus users submit texts to the Hypal tool which saves data on a database and performs automatic alignments at paragraph and sentence levels. They can edit the alignment manually to correct possible mismatches. For annotating translation errors, the MeLLANGE error-tagging taxonomy was employed. To serve error-tagging purposes, two separate interfaces have been developed: the student and the teacher interfaces. Teachers upload parallel texts and annotate translation errors. They can also view statistical error analysis of the data, a unique feature that enables teachers to figure out common problems and the most challenging areas.

- D. The Multilingual Student Translation Corpus (MUST) is a learner translator corpus composed of translations produced by foreign translation students or trainee translators along with systematic metadata. (Granger, Lefer 2020). As its name suggests, it encompasses STs and their multiple translations produced by students or to-be translators.

The corpus uses *Hypal4MUST* interface for collecting, aligning, and annotating data. Moreover, the MUST corpus developed an error typology of 60 tags to annotate the translated texts on the corpus, called the 'Translation-oriented Translation System'. The corpus contains a feature of tagging positive translation choices. It also contains an optional layer to tag procedures used by students to

solve problems that are not erroneous such as explication, borrowing, and so on.

## *2.2 Limitations and gap in previous research*

This paper addresses the gap in two main features in LTCs. In corpus linguistics, there are two main approaches: the corpus-based and corpus-driven. In the corpus-based research approach, the corpus is a method for testing, approving, or refuting an existing theory or methodology whereas in corpus-driven's the corpus is the source for drawing on novel hypotheses, theories, or analyses (Tognini-Bonelli 2001). Conducting a corpus-based research requires having a source or a resource. Granger and Lefer did a bibliometric survey on corpus-based translation and interpreting studies to draw insights into the current status of corpus-based translation as a field (Granger and Lefer 2022). They examined three trending categories of corpus studies: theory- and methodology-oriented studies, applied studies, and empirical studies. They surveyed '186 corpus studies published in English in twelve top-rated translation and interpreting journals between 2012 and 2019'. Though the study is confined to journals written in English, the study revealed the current reality of this research field. In general, they concluded that corpus-based translation and interpreting studies are still a relatively young research field. Although the scope of the study is much wider than our scope in LTCs, it gives an overview of the status of corpus-based research and that it is under-researched. Considering the aims geared by ULTC, Alfuraih stated that ULTC aims to create a standardized, corpus-driven error taxonomy to support teachers, students, and researchers by providing resources on common translation errors made by undergraduate learners when translating to and from Arabic. Additionally, a core objective of the ULTC is to develop a corpus-based quality assessment framework that evaluates undergraduate translations in terms of competence, creativity, and effective practices. (Alfuraih 2024). Furthermore, by annotating errors in ULTC, research areas typically investigated by researchers interested in LTC are becoming available to researchers interested in translation within the scope of the Arabic language or Saudi undergraduates. Additionally, research will be directed beyond error analysis and quality assessment of translation of students to quantitative representation of errors, based on large data. Tracking students' performance before and after editing will unveil students' progress and will highlight competence aspects. The availability of large, authentic annotated translations will contribute to enhancing students' critical thinking. Hence, tagging errors will inform theory and practice.

## **3. ULTC pilot error taxonomy**

Corpus error annotation requires setting out a defined error typology. Using state-of-the-art studies across error-tagged LTCs, ULTC has developed its pilot error taxonomy. The taxonomy is web-based, user-friendly, editable, and expandable. Granger stated that in the literature on LTCs there are four principles for an annotation system: the annotation system has to be manageable, well documented, and makes

the correction task easy and the typology should be hierarchical (Granger 2020). In the case of ULTC, the principles of the annotation system are largely met. The methodological framework focuses on semiotic features of a linguistic sign: syntax, semantics, and pragmatics (Charles Peirce and Charles Morris 1983). This is the typology method used in the RusLTC (Maria Kunilovskaya 2016). Most error-tagged LTCs classify errors into a hierarchical scheme of language errors and content transfer errors (MeLLANGE (2014), CELTraC (2014), MUST (2020). Regardless of language pairs, the scope of translation errors falls under lack of linguistic expression or improper content transfer (Kunilovskaya 2014). Likewise, the scope of ULTC error typology comprises errors in language, language mechanics, and content transfer. The developed tagset is inspired by experience in dealing with students' translations, the Arabic language features, and the common practices across LTCs. The tags are of types: positive and error tags. Figure 1 depicts error tags and positive tags on the administrator website.

The current taxonomy is directed towards written texts, not interpretations. It includes around 100 tags. However, to make the tagset concise and manageable, the tags that are incorporated on the error-tagging tool are 20 for errors and 8 for positive tags. Each error or positive tag is coded and colored. Coding is based on the initial letters of a word. For example, [LA-SY- DERI] stands for derivation. It is classified under the category *Language*, and the subcategory is *Syntax*. The error type is *Derivation*. To avoid having a complex taxonomy, errors that undergo a broad classification are assigned the same color. For example, the yellow color is assigned to the syntactic category, and pink for lexis. Hence, ULTC users can easily figure out an error category once they glance at the color. Then, they can specify which error

#	Name	Code	Color	Mode	Action
1	Addition	[AD]	Pink	Written	[+]
2	Omission	[ER-OM]	Blue	Written	[-]
3	Wrong meaning	[ER-WRMG]	Red	Written	[x]
4	Indecision	[ER-IN]	Blue	Written	[x]
5	Structure	[STR]	Yellow	Written	[x]
6	Derivation	[DERV]	Yellow	Written	[x]
7	Case	[CASE]	Yellow	Written	[x]
8	Action verb	[ACTV]	Yellow	Written	[x]
9	Gender	[GND]	Yellow	Written	[x]
10	Plurality	[PLR]	Yellow	Written	[x]
11	Tense	[TNS]	Yellow	Written	[x]
12	Word Order	[WDO]	Yellow	Written	[x]
13	Passive voice	[PSVOC]	Yellow	Written	[x]
14	Form	[FME]	Yellow	Written	[x]
15	Wrong Word	[WRW]	Pink	Written	[x]
16	Punctuation Marks	[PUNCT]	Orange	Written	[x]
17	Comprehension	[COMF]	Red	Written	[x]
18	Style	[STY]	Cyan	Written	[x]
19	Culture	[C]	Brown	Written	[x]
20	Pragmatics	[PRG]	Grey	Written	[x]
21	Spelling	[SP]	Olive	Written	[x]
22	Morphology	[MR]	Black	Written	[x]
23	Register	[R]	Teal	Written	[x]

#	Name	Code	Color	Mode	Action
1	Bilingual Competence	BC-S	Grey	Written	[+]
2	Bilingual Competence	BC-P	Brown	Written	[+]
3	Bilingual Competence	BC-M	Green	Written	[+]
4	Bilingual Competence	BC-L	Brown	Written	[+]
5	Pragmatic Competence	PC-P	Purple	Written	[+]
6	Pragmatic Competence	PC-S	Green	Written	[+]
7	Profession-related competence	PRC-E	Purple	Written	[+]
8	Profession-related competence	PRC-TT	Red	Written	[+]


Figure 1. Error and positive tags.

type by reading the code. The error types that are selected to be on the taxonomy are common among undergraduates and expected to come across during the process of error-tagging. More focus is placed on language error tags as error annotation in this phase is at the sentence level. Tags are displayed within a sentence after the tagged word.

The aim of the taxonomy developed is to provide an automated source of identifying and describing translation errors committed by translation students or translators-to-be systematically and comprehensively. This will contribute to a breakthrough in research across a number of translation fields: translation pedagogy, translation competence, 'qualitative/quantitative' error analysis, translation criticism, and curriculum design. Researchers, instructors, and students will benefit from viewing the tagged errors displayed on the screen. This will result in identifying errors as patterns and working on developing strategies to avoid such patterns of errors. If well devised, it is anticipated that translation pedagogy will develop more effective curricula and students' levels will be with higher outcomes.

As only trivial attempts are directed towards positive tags, the taxonomy makes this feature available as it will open up novel directions in the fields of translation competence and TQA. It will enable researchers to find a resource of representative data for systematic analysis to draw new findings. By reviewing Table 1, good or creative choices of translation are referred to by adding a tag that denotes a positive choice. In ULTC, however, developing a stand-alone taxonomy of positive choices is one of the ambitious objectives (See 2.2).

#### 4. ULTC error tagging tool

The error-tagging tool is incorporated on the ULTC website. On the administrator website, there is an icon for *tags* where annotators can input, edit, or delete tags by clicking on the *setting* button . A window will pop-up to add the required information. To add a new tag, an annotator has to select the error type from the drop-down list and define whether it is an error or a positive tag. Also, the mode has to be selected whether it is written or oral (Figure 2).

In the boxes that follow, the annotator will type in the category and name of error-type; then, s/he will assign a code for the error type. The error-tagging tool is designed to allow future modifications on tags. It is possible for annotators to add new modes and can modify an existing tag by clicking on the *setting* icon (Figure 3). For an easier visual representation, each category is assigned with a color (Yellow is for syntax and morphology, pink is assigned to lexis, purple is for cohesion, and so on). To maintain short and straightforward codes, language levels are reflected by their assigned color on a given code (Figure 1).

As the methodological framework of the ULTC is to provide pre-edited translations and post-edited translations for a single project (Alfuraih and El-Jasser, 2024), the tool is designed to tag both pre-edited and post-edited texts. On the administrator's interface, texts and tags are displayed side-by-side. Each sentence is aligned with the corresponding draft translation and the final version of the translation. Next to each draft sentence, there are two separate boxes (One for error tags and the other

The screenshot shows the ULTC website interface. On the left is a dark sidebar menu with options like Home, ULTC, Reference Corpus, Users, Advanced Search, and EXTRAS (Tags, Papers, Fags, Contact messages). The main content area is titled 'Home / Error tags / Create a new error type'. It features a 'New error tag' form with the following fields: Type (dropdown menu with 'Error Tagging' selected), Mode (dropdown menu with 'Please select mode'), Category (text input), Name (text input), Code (text input), and Color (color picker with a black swatch). A blue 'Save' button is at the bottom of this form. To the right, there is a 'New mode' form with a 'Mode Name' text input and a blue 'Save' button.

**Figure 2.** Adding a new error tag window

The screenshot displays two forms side-by-side. The left form is titled 'Current error tags' and contains: Type (dropdown menu with 'Error Tagging'), Mode (dropdown menu with 'Written'), Category (text input with 'Transfer'), Name (text input with 'Addition'), Code (text input with '[AD]'), and Color (color picker with a blue swatch). It has blue 'Save' and red 'Delete' buttons. The right form is titled 'Edit mode' and contains: Mode (dropdown menu with 'Written'). It also has blue 'Save' and red 'Delete' buttons.

**Figure 3.** Modification of an existing error tag.

for positive tags) for tagging the proposed translation. The same applies for the final version of the translation (Figure 4). On the other hand, it is possible to tag texts offline. Each project is aligned in an Excel file; and annotators can download the text to be annotated, work on tagging the text, and finally upload it to the ULTC website. However, more tests are needed for activating the downloadable feature. In this phase of the project, error annotation is at the sentence level. It is planned to include multi-level and thematic annotation in the future. It is worth mentioning that all students have already passed the course, and the annotation process is not for assessing or evaluating students. However, the project files are stored based on a systematic criterion that does not reveal the names of students. The files, however, are linked to their respective metadata where adding the student's name is optional.

On the ULTC user website, query search could be retrieved with or without annotation, depending on the user's choice. In case the user is interested in an error-tagged query, there is an option where one can click to query with error-annotation.

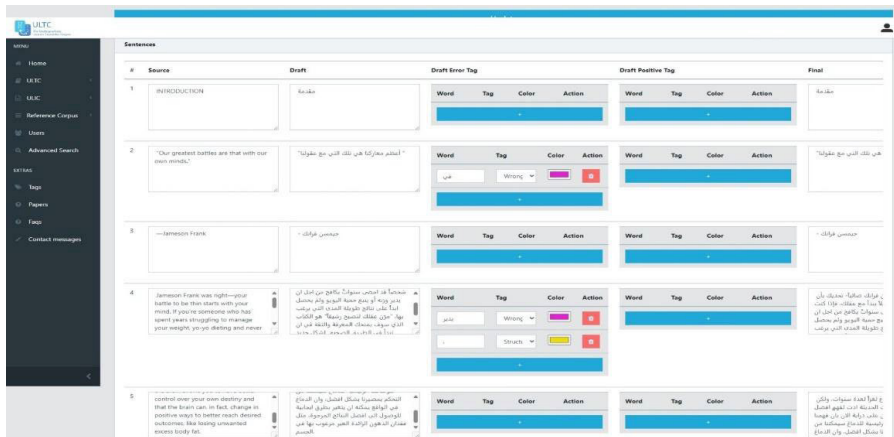


Figure 4. The process of tagging errors and positive choices.

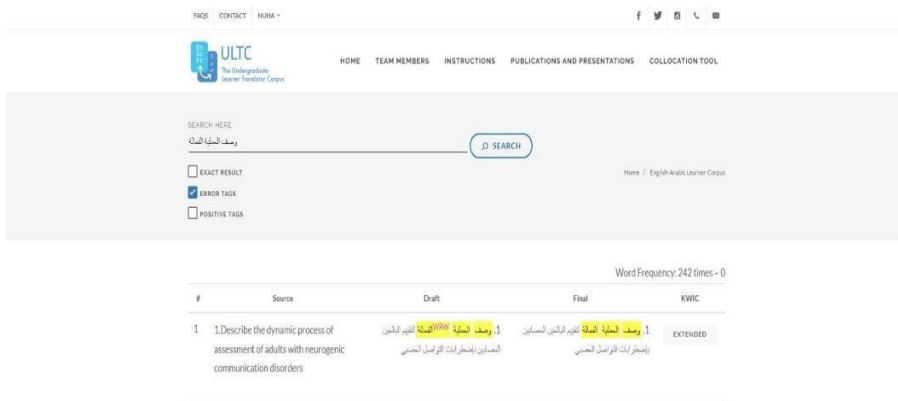


Figure 5. Error-tagged query search.

Error-tagging is tokenized within the text. They are shown next to an erroneous word (Figure 5).

### 5. Opportunities and challenges

This paper attempts at developing a tool to better devise the raw data on ULTC into tagged- ones. It also serves as a preliminary step towards developing an approved and valid taxonomy of errors. The potential advancements are enormous. The availability of the corpus for users: researchers, students, and teachers with no financial subscription requirements makes it available and handy to everyone. The automation of errors will direct research in the field of Translation into more comprehensible, reliable findings.

The taxonomy will enable researchers to apply systematic analysis on a wide-scope scale. The taxonomy as well as error-tagging will highlight the strengths and weaknesses areas of students. This will have an impact on error analysis, contrastive analysis, critical thinking, curriculum design, design of exams, developing glossaries, assignments, and in-class activities. The scope of research can be narrowed down to examine specific linguistic or translational features. The impact extends to trendy research practices such as comparison between pre-edited translation and post-edited translation, or comparison between human translation versus machine translation. By viewing and searching authentic texts and translations, students will benefit from being exposed to annotated translations and become more autonomous by developing critical thinking skills and self-directed learning. Instructors can benefit from error-tagged translations by giving authentic examples to students.

Given the fact that the ULTC corpus is 55 mln word tokens, it is a suitable resource for corpus-driven research in which researchers can draw on new methodologies and can make generalizations. Both the taxonomy and the error tagging tool are of paramount importance as they fill the gap in the lack of corpus-driven studies that involve Arabic Language as a language pair. They also fill the gap in corpus-based research to investigate linguistic and translational features, on a limited scope or wider scope. Longitudinal or cross-sectional research could be both implemented using the ULTC as the collected data comprises translations from 2014 to 2018. They will enrich the various fields of study such as translation pedagogy, TQA, translation competence, translation process, translation criticism, training translators, translation evaluation, corpus-based analysis, corpus-driven approaches, machine translation, corpus linguistics, computational linguistics, contrastive (interlanguage) analysis, lexicography, and qualitative and quantitative research. The web-based tagger tool makes the process of error annotation easier for annotators, and does not require annotators to have a computational background. The tagging process becomes also easier and time-saving as the tagset is incorporated into the system where annotators just click on the error type and it will be displayed on the screen with all taxonomy levels. The error-tagging tool is accessible everywhere and manageable. This leads to an opportunity to form a remote team of annotators.

On the other hand, given the fact that the process of error-tagging is manual, this requires plenty of time and effort and might be overwhelming for some annotators. Dealing with such big data requires forming a team to tag the errors across the corpus. Hence, this entails working on a manual to be used by annotators to maintain systematic methodology as done in other projects (The Louvain Error Tagging Manual; Granger et al., 2022). Current challenges have to be addressed such as dealing with duplication of errors and how to tag a single word with more than one error. Defining the demarcations of errors poses another dimension of challenges. The proposed pilot taxonomy has to be tested for its validity and comprehensibility by inter-rater reliability. Elevating the level of error tagging is to be considered to make it available to annotate errors at a textual level, and not be confined to the sentence level. For multimodal corpora, a multi-layered standoff model is to be developed.



## 6. Conclusion

This paper presents the attempts to develop an error taxonomy and error-tagger tool for annotating the error and positive tags in the EALTC. It is the main corpus in the composite corpus ULTC. EALC is a parallel, bidirectional, and sentence-aligned corpus of graduation projects of translations produced by female students at the Translation English Department, Princess Nourah bint Abdulrahman, KSA. Most translations are from English into Arabic. The corpus traces the students' progress by aligning the draft translations and final versions of translations with their respective ST. There are two websites for ULTC. One is for administrators to develop and improve features of the corpus. Another one is for users which is freely accessible. A pilot taxonomy has been developed to classify error and positive tags. It focuses on improper content transfer, language errors, and language mechanics. For tagging translations, each error or positive type is assigned a code and color. The tagger tool is built-in on the ULTC website. This feature makes it easy for annotators to have access everywhere and annotate remotely. The tool is editable and user-friendly. Annotators can view the annotated parts by clicking on an icon *preview* to make sure the tagging process is successful without having to surf the ULTC website, the user interface. Users can access the corpus via the corpus website. They can query either with or without annotation depending on their preference. The current project yields promising advancements in the field. It also requires working on solving problems to challenges that arise to reach core findings. The purpose of error-tagging this corpus is to keep up with advancements in the field and to enrich the Arabic language with a resource for analyzing data, improving translations, enable researchers to conduct research with a variety of comprehensible and reliable approaches that have not been available before.

## References

- Alfuraih, R. F. (2020). The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics. *Language Resources and Evaluation*, 54(3), 801–830.
- Alfuraih, R. F., & El-Jasser, N. M. (2024). Exploitation and evaluation of an Arabic-English Composite Learner Translator Corpus. *International Journal of Arabic-English Studies*, 24, 155–172.
- Alfuraih, R. F. (2024, July). Competence and creativity indicators: a taxonomy of translation positive practices in the undergraduate learner translator corpus. The 11<sup>th</sup> Inter-Varietal Applied Corpus Studies (IVACS) Biennial Conference. Cambridge, UK: University of Cambridge.
- Bowker, L., & Bennison, P. (2003). Student translation archive: design, development and application. In F. Zanettin, S. Bernardini, & D. Stewart (Eds.), *Corpora in Translator Education* (pp. 103–117). London, UK: Routledge.
- Castagnoli, S. (2009). A new approach to the analysis of explicitation in translation: multiple (learner) translation corpora. *International Journal of Translation*, 21(1-2), 89–106.
- Castagnoli, S., Ciobanu, D., Kunz, K., Kübler, N., & Volanschi, A. (2011). Designing a learner translator corpus for training purposes. In Natalie Kubler N. (Ed.), *Corpora, language, teaching, and resources: From theory to practice*, (Vol. 12, pp. 221–248). Lausanne, Switzerland: Peter Lang.
- Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources & Evaluation*, 48, 33–43.
- Fictumova, J., Obrusnik, A., & Stepankova, K. (2017). Teaching specialized translation. Error-tagged translation learner corpora. *Sendebär*, 28, 209–241.

- 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), Paris, France.
- Florou, K. (2019). Learner Translator Corpus: Italogreco or Another Way to Confirm Teachers' Intuitions. *Journal of Education and Learning*, 8(5), 75–80.
- Granger, S., & Lefer, M. A. (2020). The multilingual student translation corpus: a resource for translation teaching and research. *Language Resources and Evaluation*, 54(4), 1183–1199.
- Granger, S., Swallow, H., & Thewissen, J. (2022). The Louvain error tagging manual. Version 2.0. Center for English Corpus Linguistics, Louvain-la-Neuve, Belgium.
- Granger, S. & M. A. Lefer (2022). Corpus-based translation and interpreting studies: a forward- looking review. In S. Granger & M.-A. Lefer (Eds), *Extending the Scope of Corpus-based Translation Studies*. Bloomsbury Advances in Translation series (pp. 13–41). London, UK: Bloomsbury.
- Granger, S. & Lefer, M. A. (2023). Learner translation corpora: bridging the gap between learner corpus research and corpus-based translation studies. *International Journal of Learner Corpus Research*, 9(1), 1–28.
- Hovy, E., & Lavid, J. (2010). Towards a 'science' of corpus annotation: a new methodological challenge for corpus linguistics. *International Journal of Translation*, 22(1), 13–36.
- Javora, Š. (2015). *Defining an error typology: the Case of CELTraC* (Doctoral dissertation, Bachelor's thesis, Masaryk University, Brno, Czech Republic).
- Kübler, N. (2008, May). A comparable learner translator corpus: creation and use. In *Proceedings of the Comparable Corpora Workshop of the LREC Conference* (pp. 73-78), Marrakech, Morocco.
- Kunilovskaya, M., & Ilyushchenya, T. (2014). Russian Learner Translator Corpus in translator training. In *4th Conference Using Corpora in Contrastive and Translation Studies* (pp. 32–33), Lancaster, UK.
- Kunilovskaya, M., Kovyazina, M., & Ilyushchenya, T. (2014). Error-tagging in Russian Learner Translator Corpus and its classroom applications. *didTRAD, Barcelona, Spain (July 8, 2014). Extended paper on Academia. edu > Maria Kunilovskaya.*
- Kunilovskaya, M. (2016). *Description of RusLTC translation error typology*. [https://rusltc.org/static/references/description\\_RusLTC-error-typology2016.pdf](https://rusltc.org/static/references/description_RusLTC-error-typology2016.pdf)
- Kunilovskaya, M., & Morgoun, N. (2016, June). Available corpora and error-annotated student translations in translator education. In *Proceedings of the 6th Conference. The Future of Education* (pp. 121–5), Florence, Italy.
- Kutuzov, A., & Kunilovskaya, M. (2014). Russian learner translator corpus: design, research potential and applications. In *Text, Speech and Dialogue: 17th International Conference, TSD 2014, Brno, Czech Republic, September 8-12, 2014. Proceedings 17* (pp. 315–323). Springer International Publishing, 2014.
- Lapshinova-Koltunski, E. (2013, August). VARTRA: a comparable corpus for analysis of translation variation. In *Proceedings of the sixth workshop on building and using comparable corpora* (pp. 77–86). Sofia, Bulgaria. Association for Computational Linguistics.
- Mikhailov, M., & Cooper, R. (2016). *Corpus linguistics for translation and contrastive studies: a guide for research*. Routledge. Corpus Linguistics Guides. London, UK: Routledge.
- Mikhailov, M. (2024). Corpora, translation studies, and contrastive. In Li, D., & Corbett, J (Eds.), *The Routledge Handbook of Corpus Translation Studies* (p. 121). Milton, UK: Taylor & Francis.
- Obrusník, A. (2013). A hybrid approach to parallel text alignment “ *Dipl. práce., Masarykova univerzita. Dostupné z http://is.muni.cz/th/356468/ff\_b.*
- Rochberg-Halton, E., & McMurtrey, K. (1983). The foundations of modern semiotic: Charles Peirce and Charles Morris. *The American Journal of Semiotics*, 2(1-2), 129–156.
- Russian Learner Translator Corpus (2013, March 14).” *Learner Translator Corpora Related Research*”. [https://rus-ltc.org/static/download/LTC\\_RelatedResearch\\_March2014.pdf](https://rus-ltc.org/static/download/LTC_RelatedResearch_March2014.pdf).
- Sosnina, E. P. (2006). Development and application of Russian translation learner corpus. St. Petersburg, Russia: Papers from the Corpus Linguistics Conference.
- Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., & Tsujii, J. I. (2012, April). BRAT: a web-based tool for NLP-assisted text annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 102–107). Avignon, France. Association for Computational Linguistics.
- Štěpánková, K. (2014). Learner Translation Corpus: CELTraC (Czech-English Learner Translation Corpus). *Dipl. práce, Masarykova univerzita. Dostupné z http://is.muni.cz/th/400362/ff\_b.*
- Tognini-Bonelli, E. (2001). *Corpus linguistics at work*. Amsterdam, The Netherlands: John Benjamins Publishing.

- Uzar, R., & Walin'ski, J. (2001). Analysing the fluency of translators. *International Journal of Corpus Linguistics*, 6, 155–166.
- Wurm,A.(2016).Presentation of the KOPTÉ Corpus and Research Project. [https://www.academia.edu/24012369/Presentation\\_of\\_the\\_KOPTÉ\\_Corpus\\_and\\_Research\\_Project](https://www.academia.edu/24012369/Presentation_of_the_KOPTÉ_Corpus_and_Research_Project).