The Design of the SauLTC application for the English-Arabic Learner Translation Corpus

Maha Al-Harthi
Princess Nourah University, Riyadh, Saudi Arabia
Amal Alsaif
Al-Imam Mohammad Ibn Saud Islamic University, Riyadh, Saudi Arabia

Abstract

This paper reports on the development of two important tools designed specifically for the project entitled "the Saudi Learner Translation Corpus" (SauLTC): the conversion tool and the SauLTC application. The challenges encountered during the different stages of the project, especially the stage of the corpus alignment, were highlighted. SauLTC is a POS-tagged and error annotated parallel corpus proposed to be 3 million tokens, including translation projects required for graduation at the College of Languages at the Princess Nourah bint Abdulrahman University in Riyadh. It comprises a multi-version corpus featuring linguistic annotation, complemented with an interface for monolingual or bilingual querying of the data. The corpus can be used to identify the students' strategies in translation and analyze their patterns of language use. The paper describes the corpus parameters and compilation process, followed by an explanation of how the textual processing and sentence alignment is being conducted. A detailed description of the SauLTC conversion tool and the application will be provided. Potential uses of the corpus will be suggested for research, training and pedagogical purposes.

1 Introduction

The merge of Learner Corpus Research (LCR) and Corpus-Based Translation Studies (CBTS) was inevitable due to their shared characteristics of interlingual mediation (Granger and Lefer, 2018). Both LCR and CBTS involve assessing the impact of transfer; a kind of transfer from L1 for LCR and from the ST for CBTS (Gilquin et al., 2008). Parallel corpora are also used in a variety of NLP and IE systems, dictionary construction and automatic alignment systems.

In line with these developments, the present paper introduces SauLTC, the first unidirectional, multiversion parallel learner corpus in the Arab world. It consists of final year students' translation projects required for graduation. The structure of the paper is as follows: in the following section, we reviewed the relevant studies in the area of learner translation corpora. Section 3 is devoted to the description of the design and the development of the SauLTC, its structure, participants, corpus compilation and data normalization. The conversion tool developed for the project is described, together with the SauLTC application. Finally, the corpus statistics and general remarks will be provided. The tool and the application could be used for other languages with slight modification of the used POS tagger. Both can be accessed by contacting the authors.

2 Related Work

Corpora that are specifically designed for use in the translation pedagogy have been a significant development. One of the first endeavors is the use of syllabus driven stratified parallel corpora to address specific teaching and learning tasks and train for specialized areas in translation (Tiayon, 2004). Nevertheless, these corpora remained reference corpora that illustrated best case practice. The primary purpose of corpora of learner translations is to provide new possibilities and insights into the translation training process. Before the availability of translation learner corpora, accessing the process of translation and translation training was mainly conducted through think aloud recordings, questionnaires, key-logging and eye-tracking (Göpferich and Jääskeläinen, 2009). While the results of these methods are informative, their collection tends to be limited due to cost and time constraints. Corpora, on the other hand, is relatively easier to compile and access, which encouraged the development of an increasing number of bi- and multilingual parallel learner corpora around the world. To the best of our

knowledge, no English- Arabic leaner translator corpus has been developed so far. However, there are other learner translator language pairs. Historically, we can distinguish two main stages in the development of learner translator corpus research. The first stage comprises early endeavors, which were characterized by being smaller in size and publicly unavailable, while the second stage witnessed more recent projects that are greater in size with their own online interfaces available to the research community at large.

One of the earliest learner translation corpora was compiled in Germany by Robert Spence (1998). Another Learner Translator Corpora are PELCRA (Uzar, 2002) and the student Translation Archive (STA) (Bowker and Bennison, 2003). These early attempts to compile collections of electronically stored learner translations primarily aimed at identifying common problems in learner translations. The Russian Translation Learner Corpus or RuTLC (Sosnina, 2006) is another type that consists of English STs and their translations into Russian as the native language. Like the *PELCRA* corpus, it is also error-tagged, allowing automatic analysis of learner errors. It is used to identify the frequency and distribution of error types in order to detect the most frequent lexical, stylistic and grammatical errors in student translations in order to modify and improve teaching strategies and materials. Finally, Multiple Italian Student Translation Corpus (MISTiC) was developed by Castagnoli (2009) for a corpus-based study on explicitation. Multi-parallel and longitudinal analyses are possible, as there are several translations for each ST and each student contributed more than one translation. Although collecting and analyzing the output of trainee translators can be useful for translation teaching, and research started in the above pioneering projects at the end of the 1990s, these corpora remained exclusive and inaccessible to the wider community of researchers. The corpora also varied in the number of languages they include, the directionality of the translation, and the technologies used.

The second stage was marked by the availability of online learner translator corpora projects such as the ENTRAD project in Spain, MeLLANGE LTC in the UK, RusLTC in Russia, and CorTrad in Brazil. ENTRAD (see Flore´n Serrano and Lore´s Sanz, 2008) is a text-level aligned corpus that can be queried by the metadata such as the translator's age, gender, and mother tongue. Perhaps the most remarkable achievement in translation learner corpus research was the compilation of the MeLLANGE Learner Translator Corpus (LTC), which was completed in 2007 and provides advanced searchable and user-friendly query interfaces that allow the user to perform an extensive search through rich metadata. The query tool also includes error-tagging system based on prior linguistic annotation.

The one-to-many concordances are used for comparative observations (Castagnoli et al., 2011). The corpus is compiled of originals of four different text types (journalistic, administrative, legal and technical) and their translations produced by learners as well as professional translators (to provide reference translations for comparison with student versions: the trainees). MeLLANGE represents a valuable source of data for universal analyses about translation trainees' performance, due to its availability online and its relatively big size. Like the MeLLANGE LTC, RusLTC (Kunilovskaya and Kutuzov, 2015) is error-tagged and annotated with various metadata about translators and translation situations. It is a multiple learner translation corpus containing English and Russian source texts together with their translations produced by Russian translation trainees. Similarly, CorTrad (Tagnin et al., 2009) is a multiversion English-Brazilian Portuguese corpus that allows a comparison not only between source texts and translations, but also between the different translations of the same source text.

Recent developments in the field of translation learner corpus research include KOPTE (Corpus Project of Translation Evaluation) (Wurm, 2016), UPF LTC (University of Pompeu Fabra learner translator corpus) (Espunya, 2013), and CELTraC (Czech-English Learner Translation Corpus) (Štěpánková, 2014). However, these corpora are not yet made available online.

Sylviane Granger and her team from the Centre for English Corpus Linguistics of Université Catholique de Louvain proposed a new corpus project initiative entitled Multilingual Student

Translation (MUST) in 2016. This corpus can be searchable on a web-based interface, Hypal4MUST, an adapted version of the Hypal tool designed by Obrusnik (2014) for the processing of parallel texts.

The above-mentioned corpora differ mainly in terms of the languages involved, the translation direction, and the techniques/technology employed for corpus creation and querying. Most of these corpora, similar to the present corpus, focus on translations into the students' native language in order to investigate translation related phenomena, whereas the corpora that include student output in the foreign language were considered as a tool for foreign language teaching (Uzar and Waliński, 2001). The principal objective of these projects is to identify common problems and errors in student translations in order to improve teaching materials.

3 SauLTC

SauLTC is a promising project that was initiated to keep up with the latest developments in linguistic research and to make use of the piling archive of the PNU students' translation projects. SauLTC is a multi-version corpus organized in three parallel sub-corpora. The first corpus

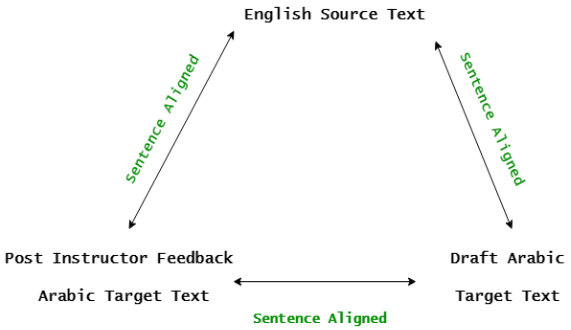

Figure 1: The alignment directions of the SauLTC files

comprises the English source texts. The second and third corpora include two versions of the translations of one source text: First, the draft translation (Version 1 of TT) is the first attempt of the student to translate the source text on her own. Second, the final translation (Version 2 of TT) is the student's same translation after making the necessary changes based on her instructor's feedback.

Each student's contribution includes a learner profile, the source text, the draft translation and the post-instructor feedback final submission. All this information is described in the searchable corpus metadata, with all translations and metadata being anonymized. The corpus is sentence aligned across all three versions (see Fig.1). Thus, one of the functionalities available in our corpus allows end users to examine what a trainee translator produces on her own (draft translation) and the effect of an expert translator's feedback (final translation submission).

3.1 Participants

The creation and compilation of the SauLTC involved three types of participants. The first and main type of participant are the students who are Arabic native language speakers. Their explicit participation consent is documented on the profile forms in addition to other background information they provided. The second type of participants are the instructors who provided feedback on the students' drafts and later assessed the students' final submission. The third type of participant is the alignment verifiers, qualified translators, who were later enlisted to double-check the automatic sentence alignment. Each verifier aligned at least three students' projects which entails the double-checking of nine parallelization.

3.2 Corpus Compilation

Currently, there are 186 student participants, 47 instructor participants, and 17 alignment assistants (see section 3.8). As mentioned above, each student's contribution consists of three Microsoft Word files and a learner profile in a Word template. The leaner profile includes detailed information about English language exposure, together with the student's consent form. The source texts are chapters or booklet extractions from extracted fiction, self-help, biography, history, health, psychology, religion, culture, management or science. The source texts are 6000 words on average. All three documents are collected in one folder under the student's name in addition to the student's profile information. We designed naming convention for this first

version of the corpus as following: each folder is named 'SauLTC_V1_Seq-No4digits_Year_SsmesterNo', for example, SauLTC_V1_0008_2016_S2; the four Word files in the folder have the same naming that end with one of the following depending on its type (_source, _draft, _final _metadata). We also separately collected metadata information of supervisors and alignment verifiers that were recruited to manually double-check the automatic sentence alignment in online forms. This information include the educational background, professional experience, the consent and work commitment.

3.3 Data Normalization

Due to the large number of illustrations, tables, diagrams and figures that the source and final translated texts included and the various ways that students used to deal with, all texts require a prior stage of normalization to minimize the challenges

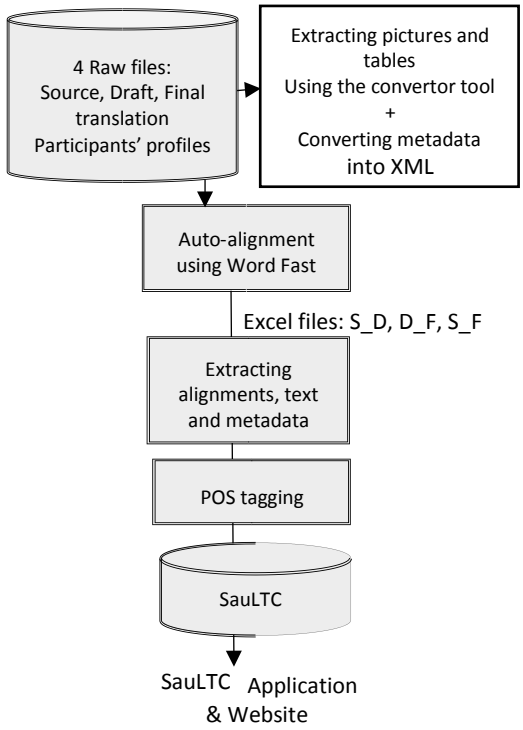

Figure 2: The pipeline process of constructing the SauLTC

that the alignment process may face. Some students excluded these illustrations from their draft target texts and subsequently their final target texts. Others translated and recreated them in the target texts. Since these illustrations and tables as well as the strategies followed by students in dealing with them are an integral part of the translation process, we decided to include them in the searchable database. However, the automatic sentence aligner can only handle running text. tables had to be removed and saved separately in order to be manually aligned.

Due to the large number of student folders, we developed our own tool, SauLTC XML Conversion tool, to extract all these illustrations (See section 3.4). The tool is effective, but it is not able to distinguish automatically in-text illustrations and diagrams from irrelevant paragraph lines, borders and other decorative embellishments. These superfluous additions have to be deleted manually from each student's folder post-extraction. The relevant diagrams and tables are then added to the database automatically to be accessed by the researcher when needed.

3.4 SauLTC XML Conversion tool

One of the main obstacles of the initial automatic processing of the student folders is the diversity of the translation genres and the formatting. This lack of uniformity led us to develop a converter, SauLTC XML Conversion tool. The tool is able to convert any word text file (English or Arabic) into XML standard format with ability of extracting all figures, tables and formatting shapes separately. Figure 3 shows the main screen of the tool where the user should upload the three Word files: source, draft translation, and final translation with the translation learner metadata. Before the XML conversion, the tool recognizes and removes headers, footers and decorating shapes that student may include in their submission. The tool also clean up the text from any extra spaces. Once the conversion process is

complete, the statistical information of the number of paragraphs, sentences, words, unique tokens, tables and images will be shown for all the three files. These statistics could be used to check the quality of the translation and how the student modifies the final version compared to the draft version.

The tool also offers browsing and editing facilities on the extracted text and save the new editing into the XML format. The user also can browse the extraction of metadata and modify any field before converting it into XML format. Due to the inconsistency on filling the earlier metadata form manually by the student, for example, the student may remove some fields or add unwanted information, which make the automatic extraction difficult and need a kind or normalization to ensure that the selected fields are correctly filled. The resulted XML and JPEJ files will be transferred into the next process of the corpus manipulation as in Figure 2.

3.5 Corpus Parallelization

The SauLTC is a sentence aligned bilingual corpus. For more efficiency, our alignment process runs in two stages: automatic alignment (English-Arabic and Arabic-Arabic pairs) and manual verification, as in Figure 2. The Auto-aligner at WordFast Anywhere was utilized for the automatic parallelization of the source text, draft text, and final submission text. WordFast Anywhere is a free web-based set of translation memory products.

The manual verification of the automated aligned files is all handled by the verifiers who should receive at least three student folders; each has the four Word files (source, draft, final translations and the metadata) and follow the instructions in the SauLTC alignment guidelines. We offer a tutorial video to ensure that each verifier had everything explained in a step-by-step format with online assistance by the corpus team. After they complete their double-checking and report comments for any unusual dealings, they fill in an online short form indicating an approximation of the number of hours it took and the number of mis-alignments they found. The parallelized three excel sheets (source_draft, source_final, draft_final) are uploaded to the SauLTC team. These excel sheets are then converted into XML files and used to create the online parallel searchable database automatically (See section 3.7 for more details about the tool).

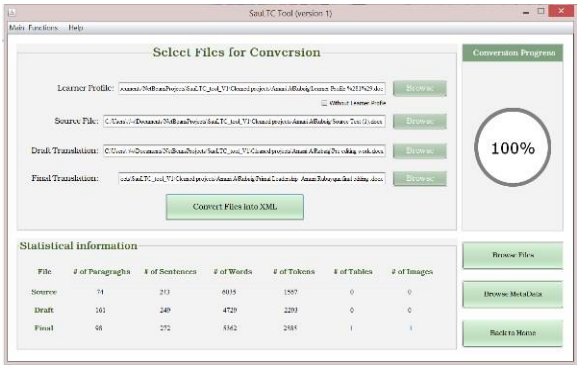

Figure 3: The SauLTC XML Conversion tool main interface

3.6 Part of Speech (PoS) Tagging

To maximize the benefit of using the SauLTC corpus in research, all sentences in the three versions are morphologically tagged using powerful POS automatic taggers for both languages. For the English source texts, the Stanford Automatic Tagger (Toutanova et al., 2003) was used, due to its availability as a powerful open source English tagger. For Arabic, MADAMIRA (Pasha et al., 2014) was used to tag the Arabic texts.

In fact, the two tag sets are not identical which led to a mismatching problem while comparing the word classes in source and target texts in any parallel grammatical investigations. For instance, the user should use the actual POS tags when exploring the corpus in our engine. To overcome this issue, we propose a general tag set to map the two different tag sets: SauLTC General tags. Then, the end user is able to use a specific tag within either the English source files or the Arabic target files using unique tags, while he is able to use the more specific POS tags as well by specifying the language in corresponding files.

The POS tagging is run on sentence level for more accurate tags and this process requires more text processing including tokenization and combining words with corresponding tags. The tags are saved in our database and can be extracted in the XML files using our application and the website.

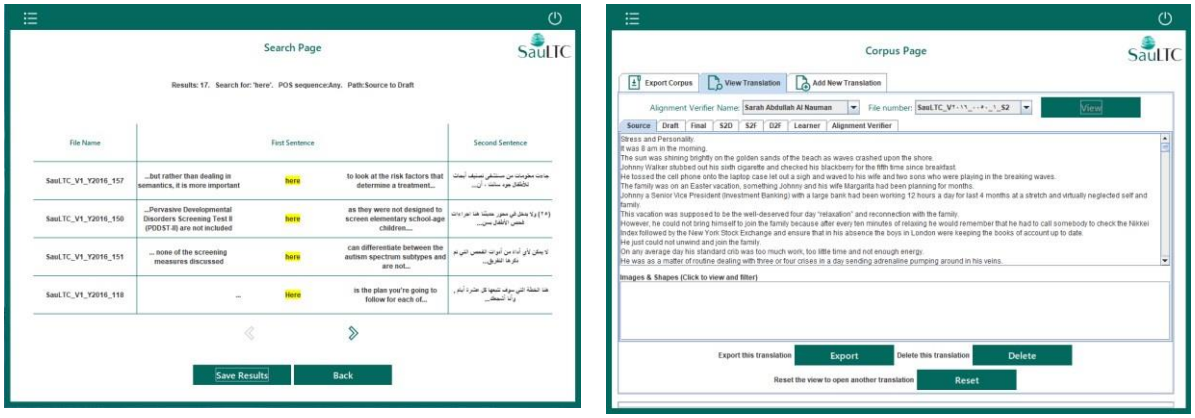

Figure 4: Two samples of the SauLTC desktop application for exploring the corpus and extracting words or phrases.

3.7 SauLTC Application

After completing the parallelization process with the manual verification of each excel alignment file, an automatic extracting desktop Java application has been developed to deal with these files (source_draft, source_final, and draft_final). For each alignment file, the application extracts the text of sentences, tokenizes the words, and extracts POS tags (Arabic and English) along with the general POS tag. We designed a comprehensive database to store all information required and to simplify and fasten the searching and corpus extraction processes. The user may only upload the excel files, the application will continue the remaining process automatically.

In the following section, we list the main features of the SauLTC application:

Importing additional translated files: This feature is to add more translations into the corpus. The application requests all aligned excel files (source-draft, source-final, draft-final) that are produced by WordFast and verified by the verifiers and the folder that has the images and table files. The user assigns a translation learner name and the verifier name from a predefined list. If the names are not listed, the user should add the new names in the editing participants tab.

The text in all files are extracted into the database and segmented further into words and saved with the automatic POS tag using Stanford for English and MADAMIERA for Arabic automatically. Any other figures and illustrations are also saved in the database.

Learner interface: This feature is to add, delete, and edit the learner metadata. Only the user who should be the administrator of the corpus can edit any information of the learner: name, age, demographic information, educational background, translation experience, the use of reference material and so on.

Supervisor interface: This feature is to add, delete, and edit supervisor metadata. The user who should be the administrator of the corpus can edit any information of the supervisor: name, educational background, years of experience in teaching translation, and years of experience in supervising translation projects.

Verifier interface: This feature is to add, delete, and edit alignment verifier metadata. The user who should be the administrator of the corpus can edit any information of the verifier: name, age, translation experience and educational level.

Exporting the corpus: This feature is used to extract the whole corpus or a couple of files that belongs to a specific verifier or learner. The extraction will be in seven XML files: the text tokenized with POS tags and the SauLTC general tags of all the source, draft, and final translation, the metadata of the participants (learner, supervisor, and verifier), source-draft alignment, source-final alignment, draft-final alignment files. All these files share the same naming convention as well as the folder name (i.e. SauLTC_V1_0008_2016_S2).

Exploring the corpus: This is the most powerful feature of our application. The user is able to search according to single or multiple criteria at the same time. First, the user has to select the alignment path (source-draft, draft-final, source-final), and enters a word or a phrase s/he wants to search for. He could specify the POS tags either language specific or the SauLTC tags, in addition to any information from the metadata of the learner, the supervisor and the alignment verifier. Figure 4 presents screenshots of exploring the corpus with all metadata, and how the application process enquires of words or phrases. The user may use regular expressions in the search box along with POS tags. The resulting table could be exported in CVS format for more portability.

The Basic SauLTC Statistics	**Total**	
Number of Alignment Verifiers	17	
Number of Learners	209	
Number of Supervisor/Teachers	47	
The SauLTC corpus - Version 1		Avg per translation
Distinct Translation Instances	115	-
Total Files (source, draft, final)	345	
English Sentences in source	36,518	318
Arabic Sentences in draft	32,196	280
Arabic Sentences in final	32,468	282
Total Sentences	101,182	880
Number of English Words	610,370	5,308
Number of Arabic Words	1081,746	9,406
Words in draft	536,177	4,662
Words in final	545,569	4,744
Total Words	1,692,116	14,714
Total Images	1,014	9
The sentence-paralizations		
Source to Draft Translations	30,421	265
Source to Final Translations	30,575	266
Draft to Final Translations	29,628	258

Table 1: The SauLTC Statistics of the first version

3.8 Corpus Statistics

The first version of the SauLTC corpus has 30,421 sentence-parallelization in source_draft, 30,575 sentence-parallelization in source_final, and 29,628 sentence-parallelization in draft_final alignment of only 115 translations in this version. The total number of tokens is 1,692,116, with an average of 14,714 tokens per file, with all corresponding tags: Stanford tags for English words and MADAMIRA tags for Arabic words, in addition to the general tags for both. The total number of translation students in the whole project is 186 who were under the supervision of 47 instructors. The alignment verification is carried out by at least 17 verifiers, participating in the project.

While we have 36,518 English sentences in source files, the aligned sentences are only 30,421 in source_draft alignment, which indicates that there is no one-to-one sentence-parallelization when the students translate the text.

For instance, there are 53 sentences on average in the source file that were merged or deleted when translated into Arabic draft and 52 sentences on average for source_final alignment.

In terms of words, similarly there is around 645 words in English were omitted when translated into Arabic draft, and around omitted 563 words when verified by the supervisors in the final version. In fact, these findings support the claim that Arabic has a richer semantic lexical system than English does, where one Arabic word may be translated into a phrase or multiple words in English to express the same meaning. In addition, the morphological structure in Arabic allows constructing a complete meaningful sentence in one token such as (سنكتبها/we will write it down). There is no significant difference between the number of sentences and words in draft and final versions, both are Arabic. The learner tended to make fewer changes in the final version, following the supervisor's comments; the sentence average in the final translation was decreased by only 22 sentences compared to the draft version. The verifiers provide any significant remarks and comments during the alignment process to assist the researchers to track the changes in the translations.

4 Conclusion

This paper introduces the first version of the SauLTC, together with some of the tools developed specifically for this corpus: SauLTC XML Conversion tool designed to convert word text files into XML standard format, and SauLTC desktop

application which is an automatic extraction tool developed to deal with the alignment excel sheets with automatic POS tagging. SauLTC represents the first learner translator parallel corpus for an English Arabic language pair. It is also one of the first corpora to provide parallelization of pre-edited and post-edited versions of trainee translations. This paper describes the challenges encountered at some of the compilation stages. The most prominent challenge was in the process of text alignment, due to the huge differences in the punctuation mark systems between the English source texts and their Arabic translations in terms of their segmentations, which in turn made automatic alignment imprecise. The practical solution was to hire assistants to manually verify and double check the alignment of sentences between the three documents of the same text. The launching of a website for the corpus and making it available online will be the following stage in the project. This will provide researchers the opportunity of exploring SauLTC with multiple selections of criteria such as extracting specific words or phrases with optional morphological features in translated texts or parallel texts, tracking the errors, investigating strategies followed by translation learners while translating multi-word units and obtaining some statistics of any searchable component. All the features included in the SauLTC application, in addition to some other features will be available in the website.

The corpus was designed to enable researchers to examine the translation process both quantitatively and qualitatively. It is a valuable resource for automatic processing of bilingual text. Translation instructors and translation students and trainees can utilize the corpus for a more data-driven approach to learning and training. Overall, the potential applications of an English-Arabic learner translation corpus are numerous and valuable for research, training purposes of automatic NLP systems such as machine translation, word alignment systems and dictionary construction.

Acknowledgments

This paper is part of a project that was funded by the Deanship of Scientific Research at Princess Nourah bint Abdulrahman University. We are grateful to all the participants who helped throughout the different stages of the project: the students, the instructors and the alignment verifiers. Special thanks goes also to all the research assistants and the team of the project programmers. We are very grateful to the anonymous reviewers for their very insightful comments and suggestions that have greatly improved this paper.

References

Adam Obrusník. 2014. Hypal: A User-Friendly Tool for Automatic Parallel Text Alignment and Error Tagging. Eleventh International Conference Teaching and Language Corpora, Lancaster, pp. 67-69.

Andrea Wurm. 2016. Presentation of the KOPTE Corpus and Research Project. https://www.academia.edu/24012369/Presentation_of_the_KOPTE_Corpus_and_Research_Project.

Anna Espunya. 2014. The UPF learner translation corpus as a resource for translator training. Language Resources and Evaluation 48(1): 33-43.

Arfath Pasha, Mohamed Al-Badrashiny, Mona Diab, Ahmed El Kholy, Ramy Eskander, Nizar Habash, Manoj Pooleery, Owen Rambow, and Ryan M. Roth. 2014. MADAMIRA: A Fast, Comprehensive Tool for Morphological Analysis and Disambiguation of Arabic. Proceedings of the 9th Language Resources and Evaluation Conference (LREC'14): 1094-1101

Charles Tiayon. 2004. "Corpora in translation teaching and learning". Language Matters, 35 (1): 119-132.

Ekaterina Sosnina. 2006. Development and application of Russian translation learner corpus. In *Proceedings of corpus linguistics—2006, St. Petersburg, Russia:* 365–373.

Gaëtanelle Gilquin, Szilvia Papp, and María B. Díez-Bedmar (eds.). 2008. Linking up Contrastive and Learner Corpus Research. Amsterdam/Atlanta: Rodopi.

Kelly Washbourne. 2015. Learning to Fail: Unsuccessful Translations as Pedagogical Resource. Current Trends in Translation Teaching and Learning E, 2: 285–320.

Kirsten Malmkjær. 2017. The Routledge Handbook of Translation Studies and Linguistics. Abingdon & New York: Routledge.

Kristina Toutanova, Dan Klein, Christopher D. Manning, and Yoram Singer. 2003. Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency

Network. In Proceedings of HLT-NAACL, 252-259.

Kristýna Štěpánková. 2014. Learner Translation Corpus: CELTraC (Czech-English Learner Translation Corpus). Bachelor's Diploma Thesis.

Lynne Bowker and Peter Bennison. 2003. Student Translation Archive: Design, development and application. In F. Zanettin, S. Bernardini & D. Stewart (eds.) Corpora in Translator Education. London & New York: Routledge, 103-117.

Malcolm Williams. 1989. The Assessment of Professional Translation Quality: Creating Credibility out of Chaos. In TTR (Traduction, terminologie, Rédaction) 2 (2): 13-33.

María C. F. Serrano. 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).

Maria Kunilovskaya and Andrey Kutuzov. 2015. A quantitative study of translational Russian (based on a translational learner corpus). In Corpus Linguistics 2015. Proceedings of the 7th International Conference: 33-40.

Natalie Kübler. 2008. A comparable Learner Translator Corpus: Creation and use. LREC 2008 Workshop on Comparable Corpora: 73-78.

Rafal Uzar and Jacek T. Waliński. 2001. Analysing the fluency of translators. International journal of corpus linguistics, 6: 155-166.

Rafal Uzar. 2002. A corpus methodology for analysing translation. In Tagnin, S.E.O. (Ed.), Cadernos de Tradução: Corpora e Tradução, 1(9): 235–263.

Robert Spence. 1998. "A corpus of student L1-L2 translation". In Granger S. and Hung J. (eds.). Proceedings of the First International Symposium on Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching, 110-112.

Sara Castagnoli, Dragos Ciobanu, Kerstin Kunz, Natalie Kübler, and Alexandra Volanschi. 2011. "Designing a learner translator corpus for training purposes." Corpora, Language, Teaching, and Resources: From Theory to Practice 12: 221-248.

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

Stella E. O. Tagnin. 2014. The CoMET Project: Corpora for Teaching and Translation. In T. Sardinha & T. Ferreira (Eds.), Working with Portuguese Corpora, 201-214. Bloomsbury.

Stella E. O. Tagnin, Elisa Duarte Teixeira, Diana Santos. 2009. CorTrad: a multiversion translation corpus for the Portuguese-English pair., Teixeira, E., and Santos, D. (2009). CorTrad: a multiversion translation corpus for the Portuguese-English pair., Teixeira, E., and Santos, D. (2009). CorTrad: a multiversion translation corpus for the Portuguese-English pair. 28th. conference on Lexis and Grammar, Bergen.

Susanne Göpferich and Riitta Jääskeläinen. 2009. Process research into the development of translation competence: Where are we, where do we need to go? Across Languages and Cultures, 10 (2): 169-191.

Sylviane Granger and Marie-Aude Lefer. 2018. MUST: A collaborative corpus collection initiative for translation teaching and research. CECL Papers: 72-73.