The Multilingual Student Translation corpus: A resource for translation teaching and research

Sylviane Granger & Marie-Aude Lefer

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. https://doi.org/10.1007/s10579-020-09485-6

Abstract

The Multilingual Student Translation (MUST) corpus is a corpus of translations produced by foreign language learners or trainee translators collected collaboratively by a large number of partner teams internationally. The corpus represents a prime example of community sourcing, as the data are collected and shared by the members of the MUST network. Two key characteristics of the corpus are that it involves a large number of language pairs and that each text is accompanied by a rich set of standardized metadata related to the source texts, the translation tasks and the students. The web interface on which the corpus is stored allows the data to be aligned and annotated with a purpose-built translation annotation system. The resulting corpus data lend themselves to a range of applications (translator training, materials design, pedagogical lexicography) and can also be used to advance empirical research in corpus-based translation studies.

Keywords

learner translation corpus, metadata, annotation, standardization, community sourcing, multilingual

1. Introduction

Learner corpus research (LCR) and corpus-based translation studies (CBTS) are two research strands that arose at approximately the same time, in the late 80s/early 90s (Granger 1993, 1994; Baker 1993, 1995). Both fields make use of corpus data to inform theory – second language acquisition theory and translation theory respectively – and to generate more efficient pedagogical applications. LCR relies on learner corpora, i.e. electronic collections of foreign-/second-language learner writing and/or speech, while CBTS relies on comparable and parallel corpora of translated texts. The two focused language varieties – learner language and translated language – have one major characteristic in common: they are "mediated" varieties (Gaspari & Bernardini 2010), i.e. they involve a process of mediation between two languages: the learner's native (L1) and target (L2) language and the translator's source (SL) and target (TL) language. As a result, LCR and CBTS have overlapping research agendas, and several researchers have called for a rapprochement between the two fields. An early study by Granger (1996: 46) states this explicitly: "Far from seeing computerized bilingual corpora as the private ground of translation specialists and typologists and computerized learner corpora as the sole concern of applied linguists, we see the two types of corpora as closely interrelated". Similarly, Johansson (2007: 313) observed that "[n]ew possibilities are afforded by the combined use of learner corpora and multilingual corpora" and Chesterman (2007: 63) suggested adding learner texts to the set of reference texts used in translation studies on the basis that "both translations and learners' texts are produced under particular constraints, and it may be that these constraints have similar effects".

In spite of these calls for greater synergy, the two fields have remained largely separate. Admittedly, learner corpus researchers have sometimes made use of translation corpora to help identify transfer in learners' texts (Cosme 2008, Hasselgård & Johansson 2011), but apart from that researchers in LCR have demonstrated very little awareness of the theoretical tenets and methodological developments of CBTS, and vice versa. A different and arguably more efficient way of interfacing the two fields lies in the compilation of learner translation corpora (LTC), which can be seen as two-in-one resources, as they contain translations produced by foreign language learners or trainee translators. Based on the observation that trainee translators can be seen as a special type of language learner, admittedly with their own specific needs, Bowker & Bennison (2002: 504) consider that an approach to collecting and studying their output similar to that used in LCR "would be highly valuable for both pedagogical and research purposes".

As shown in Table 1, there have already been several initiatives in this direction. The first LTC, the Polish PELCRA corpus and the Canadian Student Translation Archive, emerged in the early 2000s and were followed by several similar initiatives. However, as pointed out by Espunya (2014: 35), "the field is clearly in its infancy, judging by the scarcity of publications reporting results or even research programmes". This may be due to the fact that, apart from the EU-funded MeLLANGE project, the learner corpus collection resulted from local

initiatives maintaining little contact with other teams. In addition, several of these projects were exclusively teaching-oriented.

Corpus	Languages	Publications
Polish and English Language Corpora for Research and Applications (PELCRA)	Polish, English	Uzar & Walinski (2001), Uzar (2002)
Student Translation Archive (STA)	English, French, Spanish	Bowker & Bennison (2002, 2003)
Multilingual eLearning in LANGuage Engineering Learner Translator Corpus (MeLLANGE)	English, Catalan, French, German, Italian & Spanish	Kübler (2008), Castagnoli et al. (2011)
Enseñar a traducir (ENTRAD)	English, Spanish	Florén (2006)
Multiple Italian Student Translation Corpus (MISTiC)	Italian, French, English	Castagnoli (2009)
Norwegian English Student Translation corpus (NEST)	Norwegian, English	Graedler (2013)
Variation in Translation (VARTRA)	German, English	Lapshinova-Koltunski (2013)
Universitat Pompeu Fabra (UPF)	Catalan, English	Espunya (2014)
Czech-English Learner Translation Corpus (CELTraC)	Czech, English	Štěpánková (2014)
Russian Learner Translator Corpus (RusLTC)	Russian, English	Kutuzov & Kunilovskaya (2014)
Korpusprojekt zur Translations Evaluation (KOPTE)	German, French	Wurm (2016)
Undergraduate Learner Translator Corpus (ULTC)	Arabic, English, French	Alfuraih (2019)

Table 1: Overview of learner translation corpora

These initiatives have played a key role in laying the foundations of the field, but a comparative overview shows that the collected corpora share a number of features that limit their usefulness, including the following: limited set of metadata, small number of language pairs and restriction to direct translations (i.e. into the translator's mother tongue). There was therefore an opening for a new initiative that addressed some of the weaknesses of earlier collections. In March 2016 the Multilingual Student Translation (MUST) project[1] was launched and met with immediate success. The MUST network currently involves 38 partner teams in 19 countries[2]. The main objectives of the project were to collect a large multilingual student translation corpus, to design a rich set of standardised metadata, to create a standardised annotation system for translated language and to make all the data available to the contributing partners via a web-based interface for research and teaching. The interface designed for the project, Hypal4MUST, is an adapted version of Hypal (Obrusník 2014, Fictumova et al. 2017), a user-friendly tool for the collection, alignment and linguistic annotation of student translations, which features both a teacher interface and a student interface.[3]

The remainder of this paper is structured as follows. Section 2 describes the data collection process. Section 3 focuses on the standardised metadata that are collected and stored alongside the data. Section 4 presents the principles underlying the translation annotation system. Section 5 sketches some of the main research and teaching benefits that can be gained from learner translation corpora, and Section 6 concludes with plans for future work.

2. Data collection

MUST data are made up of source texts (the texts that students are asked to translate) and the corresponding target texts (student translations). MUST is a multiple translation corpus, i.e. it contains several translations of each source text, which allows for comparison of translation solutions used by various students rendering the same text. Texts translated by a single student (e.g. dissertation projects) are not included in MUST. Students

[1] https://uclouvain.be/en/research-institutes/ilc/cecl/must.html

[2] For a list of partners, see https://uclouvain.be/en/research-institutes/ilc/cecl/must-partners.html

[3] The project-specific interface, Hypal4MUST, is not available outside the MUST project but the generic Hypal interface can be used by researchers outside MUST (see https://hypal.eu).

contributing translations to the corpus are mostly foreign language learners (e.g. Slovenian-speaking learners of French) and trainee translators (e.g. Italian-speaking students learning Chinese-Italian translation).

MUST source texts are between 150 and 1,000 words long. All source texts are stored in a shared source text database, a collaborative repository of source texts where teachers can submit their own source texts and/or choose existing ones from the shared database. The source text database is a perfect example of *community sourcing*, i.e. the cumulative development of resources resulting from the active participation of virtual community members (Branzov 2016). MUST source texts cover a wide range of general and specialized language genres (see Table 2 for an overview[4]). Reference (expert) translations, whether produced by professional translators or teachers, can also be added to the source text database.

General language	
Journalistic texts	News/reportage, opinion article, interview, review, sport article
Fiction	Novel, short story, poetry, play, comic, graphic novel, children's literature, song
Other	Essay, biography, autobiography, personal e-mail/letter, popularized scientific text, encyclopaedia entry, textbook/handbook, press release, political speech, proceedings of parliamentary debates
Specialized language	
Academic prose (research article, abstract, monograph, research report, call for papers), instructional text (instruction manual, recipe), promotional text (advertisement, tourist guide/brochure, product presentation), consumer information leaflet, patient information leaflet, technical specification, certificate, contract/agreement, pleading, judgment, resolution, rules and regulations, annual report, letter to shareholders (financial statement, corporate social responsibility report), balance sheet, professional/business e-mail/letter, job advertisement	

Table 2: Genres represented in MUST (general and specialized language)

Student translations can be direct (into the student's mother tongue: L2>L1) or inverse (into the student's foreign language: L1>L2). Translations between foreign languages (L2>L2) can also be included in MUST. It is possible to collect several translations per student (e.g. across several semesters or academic years), which allows for longitudinal studies. Translation tasks range from untimed in-class activities (whether graded or not) to timed examinations, with tools and resources (e.g. bilingual dictionaries, online bilingual concordancers, computer-aided translation software) allowed or not.

The data collection and annotation workflow, supported by the Hypal4MUST platform, is as follows:
(1) Data upload (source texts, target texts and corresponding metadata): teachers upload their source texts and the information about translation tasks on the teacher interface. Students can then input their translations and metadata directly on the student interface;
(2) Automatic alignment at paragraph level (and, if need be, manual editing of the automatic alignment);
(3) Data lemmatization and part-of-speech tagging;
(4) Automatic alignment at sentence level (and, if need be, manual editing of the automatic alignment) (see Obrusník 2013 for more details on the alignment algorithm);
(5) Annotation (optional): translations can be annotated on the basis of a standardized taxonomy (see Section 4);
(6) Feedback (optional): annotated translations can then be returned to students.

Students contributing their translations to MUST are asked to fill in and sign an informed consent form at the start of each academic year. The data and metadata included in the MUST corpus are fully anonymized, in accordance with the requirements of the ethical committees of the partner universities. In the annotation interface, only local teachers are allowed access to the student's name, if they wish to know the identity of the student whose translation they are annotating.

3. Standardized set of rich metadata

An important asset of corpora as sources of authentic data for linguistic research is their metadata, which play a key role, among other things, in interpreting corpus findings (see Section 5.2). However, as mentioned in Section 1, the metadata collected together with student translations in learner translation corpora have been fairly

[4] The initial basis for the MUST genre taxonomy is Lee (2001)'s categorization of the genres included in the *British National Corpus*, supplemented with genres identified by the MUST community as being relevant to the project.

limited to date (notable exceptions are NEST and KOPTE). The STA metadata, for example, were limited to course level, subject field (e.g. legal translation), date, translation conditions and students' language background. Likewise, in the EU-funded MeLLANGE project, the metadata were restricted to task type and duration, tools and resources used, translators' language background and university qualifications. This general lack of detailed metadata is quite widespread in the parallel corpora commonly used in contrastive and translation studies, such as the *English-Norwegian Parallel Corpus* (see e.g. Halverson 2017 on this issue). Also, a close examination of the types of metadata collected in the corpora listed in Table 1 reveals that labels and categories vary to a great extent across corpora, thereby making cross-corpora comparisons difficult. For instance, while in MeLLANGE, a distinction is made between three task types (home assignment, in-class and voluntary exercise), RusLTC distinguishes between three different kinds of stress (contest, examination, routine) and two places (home and classroom).

MUST metadata address these two shortcomings: they are rich and standardized (in the sense that all the metadata must be collected in order for the source and target texts to be included in the corpus[5]). In fact, the MUST set of metadata is the result of a collaborative effort: it is largely based on the project partners' input during the kick-off meeting held in 2016. In MUST, we collect three categories of metadata: metadata related to (1) source texts, (2) translation tasks and (3) students (see Tables 3, 4 and 5 for a full overview). Quite logically, both teachers and students contribute metadata: teachers provide the source-text metadata (e.g. intended audience) and part of the translation-task metadata (e.g. whether reference tools are allowed or not, whether the task is timed or marked), while students fill in the other part of the translation-task metadata (e.g. if tools and resources are allowed, students are asked to specify which tools were actually used when translating the ST), as well as translator-related metadata (e.g. language background). As the corpus is to be used in learner corpus research and translation studies alike, attention is paid to both aspects of foreign language learning (e.g. languages of instruction at primary, secondary and tertiary levels, time spent abroad, self-rated proficiency in SL and TL) and translation (e.g. use of a translation brief, use of reference tools such as online bilingual concordancers, translation memories and terminological databases). The metadata rubrics have been described in detail in the MUST documentation distributed to all partners, as some of them require clear definitions. For example, for ST keywords, it is specified that "keywords should be related to the subject of the source text. In other words, they are topic-related. Keywords need to be written in the source language (i.e. the language of the source text)". ST mode is defined as follows: "this refers to the channel of communication used during the source text production. Generally speaking, source texts can be written to be read (e.g. a novel, a newspaper article), written to be spoken (e.g. prepared speeches, dubbing), or correspond to written reproduction of spoken language (e.g. transcribed parliamentary debates)". These rich, standardized metadata make MUST a unique resource for learner translation corpus research (see Section 5.2).

Language	
Sampling	Whole text, sample If sample: title of the text from which the sample is taken, sample position (beginning, middle, end, mixed, unknown)
Title	
Author type	Sole, multiple, corporate, unknown
Author's/authors'/organization's name	
Author's/authors' native speaker status	Only for sole and multiple authors
Number of words	
Language type	General, specialized
Super-genre	See Table 2
Genre	
Sub-genre	
Topic keywords	
Publication status	Published, unpublished If published: publication date, reference
Format	Print document/hard copy, audiovisual, electronic document/digital copy
Mode[6]	Written to be read, written to be spoken, written

[5] For this reason, student translations collected outside the MUST project cannot be included in MUST.

[6] Mode, target audience and communicative purpose metadata are based on the categories used for the *Dutch Parallel Corpus* (Macken et al., 2011).

	reproduction of spoken language
Target audience	Internal audience, broad external audience, specialized external audience, other
Communicative purpose	Inform, instruct, persuade, artistic function, mixed communicative purpose, other
ST translation status	Original text, translated text, unknown
Reference translation	Yes, no If reference translation: translator status (professional translator, member of teaching/research staff, mixed), translation directionality (L2>L1, L1>L2, L2>L2, unknown), target language of the reference translation, publication status (published or unpublished at the time of data collection)

Table 3: MUST source-text metadata

Task name	
Task description	
Task deadline	
Access code for the student interface	
Student status	Trainee translator, foreign language learner, other
Level	Undergraduate, graduate
Target language	
Type of task	In-class activity, examination, home assignment, voluntary exercise, mixed
Marking	Marked, unmarked
Duration	Timed, untimed, duration in minutes
Tools and resources	Allowed, not allowed If allowed: monolingual dictionaries, bilingual dictionaries, glossaries and terminological databases, translation forums, bilingual concordancers, SL and/or TL general or specialized corpora, CAT tools with TDB, CAT tools with TM, CAT tools with both TDB and TM, machine translation, internet/web (several tools can be selected)
Feedback and revision	Yes, no If feedback received before submitting the translation: individual teacher's feedback, in-class discussion/collective teacher's feedback, peer-to-peer revision, mixed
Translation brief	Yes, no If yes: requester (commercial company, media, research and education, public enterprise, public service, other), target audience, communicative purpose, format, layout and style guide, use of glossaries and/or translation memories

Table 4: MUST translation-task metadata

Gender	Male, female, non-binary
Age (years)	Possible to select: "I do not wish to share this information"
Mother tongue(s)	Up to two mother tongues can be specified
Foreign language(s)	Up to three foreign languages can be specified
Main language(s) of instruction	Primary school, secondary school, university
Current study background	Translation, interpreting, modern languages/philology, (applied) linguistics, intercultural communication/mediation, science, law, engineering, other
Prior study background	Translation, interpreting, modern languages/philology, (applied) linguistics,

	intercultural communication/mediation, science, law, engineering, other, no prior study/secondary school only
Self-rated proficiency in SL & TL	Intermediate, advanced, native If intermediate or advanced: years studying the SL/TL, stays in SL-/TL-speaking countries and duration of those stays (in months)
Translation experience	Number of semesters spent studying translation, translation internship(s), experience with Computer-Aided Translation tools
SL & TL status	Whether native or foreign language

Table 5: MUST student metadata

As can be seen from Tables 4 and 5, the task- and student-related metadata do not include detailed information about level, whether course level (e.g. basic, intermediate or advanced course) or L2 proficiency level (based, for example, on the Common European Framework of Reference for Languages). The reason for this is that these constructs are operationalized in many different ways in the academic systems and degree curricula of the 19 countries currently involved in MUST. Including precise level information would have been misleading. The project partners therefore decided to do without it. However, it is important to stress that two metadata fields are directly related to level: (1) whether students are enrolled in an undergraduate or graduate programme and (2) students' self-rated proficiency in SL and TL (intermediate or advanced). Level information can also be inferred indirectly through translation experience (number of semesters spent studying translation).

4. The Translation Annotation System

One of the tasks that translation teachers are called on to carry out on a regular basis is correction of students' translations. This activity is highly time-consuming and provides very little return on investment as the teachers' work is scattered and essentially lost. Huge benefits could be gained if both the translations and the corrections were stored in a database, progressively generating a large searchable corpus which could be used to draw up error profiles for particular learners or learner groups and, on that basis, to design tailor-made exercises. Wible et al. (2001) have integrated this type of error tracking system into a web-based English as a Foreign Language writing environment and demonstrated its great usefulness for both teachers and students.

With a view to integrating this functionality into Hypal4MUST we needed to develop an error typology specifically designed for translation annotation. To this end we drew inspiration from two main sources. First, we examined the error typologies used in previous learner translation corpus projects. Most of these typologies subdivide errors into two main categories. On the one hand, content transfer errors, i.e. errors that are linked to the translation process, such as omission or distortion errors. On the other, language errors, i.e. errors that are independent of the translation process, such as agreement or tense errors (Kübler 2008, Espunya 2014). Apart from this major divide, the systems differ widely in terms of terminology, overall architecture and degree of granularity. Within a collaborative project like MUST, it was deemed essential to design a single typology to be used by the entire MUST community so as to ensure comparability across tasks and language pairs. A general discussion with the MUST partners at the start of the project led to the elaboration of a preliminary typology that includes 60 tags. Rather than refer to the system as an error annotation system, it was decided to adopt a more neutral/positive term, viz. Translation-oriented Annotation System (TAS). There were two main reasons for this. First, the system includes a 'Plus' metatag to mark positive features, i.e. particularly good translation choices. Second, TAS contains an optional layer of annotation that does not refer to translation errors at all, but aims to identify the translation procedures used by students to solve translation problems (e.g. explicitation, normalization, borrowing). The decision to add this extra layer is in line with the two-pronged objective of the MUST project: far from being restricted to teaching applications, it follows in the footsteps of some previous LTC projects such as the Russian Learner Translator Corpus project (Kutuzov & Kunilovskaya 2014) in having a broad research agenda. As demonstrated by Castagnoli's (2009: 5) study of explicitation, for example, a learner translation corpus is an ideal resource to study "how features which have traditionally been studied in professional translations relate to learner translation".

The second source inspiration for TAS comes from the field of learner corpus research, within which a large number of detailed annotation systems have been developed for the analysis of learners' free writing (for surveys, see Díaz-Negrillo & Fernández-Domínguez 2006 and Lüdeling & Hirschmann 2015). These typologies

are highly relevant for the 'language' section of TAS and are likely to play a key role in the annotation of inverse translations. The learner corpus literature (Dagneaux et al. 1998, Granger 2003) also lists a number of principles that annotation systems would be wise to adopt. The systems should be manageable, i.e. not involve too many tags. The categories should be well defined and there should be no overlap between them. The typology should be hierarchical so as to allow teachers and researchers to annotate texts at different levels of granularity (e.g. grammar, grammar verb, grammar verb tense). This principle is particularly crucial in a highly multilingual project such as MUST, as some languages are likely to require more subcategories than others. German, for example, has a much richer system of inflectional morphology than English. Annotation systems should also be well documented, with ample illustrations and guidelines to be followed, especially for errors that allow of different interpretations. Most importantly, as wisely observed by Espunya (2014: 37), an annotation system "should ease the correction task rather than complicate it". To achieve this objective, the first version of TAS is being tested by the MUST community, and an ecologically valid version will be developed.

5. Research and teaching benefits

As pointed out by Bowker & Bennison (2003) in one of the earliest LTC projects, learner translation corpora have far-reaching potential for both translation pedagogy and research.

5.1. Teaching

The benefits for translator training are obvious. Annotated translations can be used by teachers to highlight attested areas of difficulty and design remedial exercises. They can also be used in data-driven learning activities where students are expected to play a more active role: “Students can be presented with sentences containing a specific error-type and be asked to understand why there is an error, and what could be considered a correct translation. The reference translation can then be used to compare their suggestions with a professional translation” (Kübler 2008: 76). Several versions of the same translation segment can serve as a prompt for lively discussions on the factors that underlie particularly good translations. LTC data can also inform materials design. For example, Espunya (2014) has used her learner translation corpus to inform a grammar unit revolving around information packaging mechanisms and argumentative relations.

The database format of the student data stored in the Hypal4MUST interface is conducive to a wide range of pedagogical applications that are difficult, if not impossible, to implement outside this type of environment. For example, it is possible to identify the difficulties that are shared by a large number of students as against those that are more individual. As students have their own unique identifiers, it is also possible to chart their development across time and identify areas of persistent difficulty.

Another potential benefit pertains to the efficiency of teacher-student interactions. Thanks to teachers' use of TAS, students receive structured feedback on their work, with well-defined systematic annotations made available on the student interface. These are arguably more efficient and user-friendly than less structured feedback. Preliminary feedback from students seems generally positive, but more extended experimentation is necessary to evaluate students' reactions to the system.

Another area where learner translation data have a great deal to offer is bilingual lexicography. Here too inspiration can be drawn from the field of learner corpus research, where from quite early on learner corpus data have been used to design usage and error notes and integrate them into monolingual learners' dictionaries (MLD) (Maingay & Rundell 1987, Gillard & Gadsby 1998). The objective of these notes is to draw advanced learners’ attention to important aspects of word/phrase usage that cause them difficulty. Learner translation data could be used to design similar notes and incorporate them into bilingual dictionaries, especially learners' bilingual dictionaries (Granger & Lefer 2016). Unlike the generic notes included in MLDs which target all learners, whatever their mother tongue background, the notes would target the wide range of difficulties that learners experience when they want to translate a particular word or phrase from one language to another, including semantic difficulties (complete or partial false friends, overlapping polysemy) and stylistic mismatches (formal vs informal register).

5.2. Research

Far fewer concrete results have been achieved in the area of translation research than in that of teaching. The main focus of LTC has clearly been on pedagogical applications. Yet learner translation corpora have great potential for translation research. As highlighted early on by Bowker & Bennison (2003), “in terms of research, scholars such as Baker (1995) and Laviosa (1998) have already demonstrated that corpora can be useful for

studying the nature of professionally translated text; we believe that there is also much to be learned about translation process and product by investigating the nature of text translated by students". In relation to the RusLTC, Kutuzov & Kunilovskaya (2014: 315-316) list several avenues of research, including the study of variation and choice in translation (as LTC contain different translations of the same source text), analysis of the typical features of trainee translators' interlanguage, and exploration of the interdependence between translation features and various metadata, such as source text genre.

The MUST corpus opens up a wealth of new ways to study learner translated language. One of the greatest benefits it can offer research lies in its rich, standardized metadata (see Section 3). It has been repeatedly pointed out that detailed metadata in parallel corpora are vital for translation research (cf. Halverson 2017, Lefer forthcoming), as they allow various cognitive, technological and sociocultural aspects of translation to be taken into consideration, and ultimately lead to a better understanding of what translation is. Thanks to the rich MUST metadata, researchers can adopt robust multifactorial research designs to study translational behaviour. Relying on MUST metadata, it is possible, for example, to assess the exact influence of translators' language background, use of reference tools or revision on the translation product. In addition, if used in combination with other corpora, the corpus allows for empirical investigations of student translation as compared with professional translation and/or learner/non-native writing, which can provide invaluable insights into interlingual mediation and constrained communication characterized by bilingual language activation (Lanstyák & Heltai 2012, Kruger 2018). As for the standardized annotation system TAS, it ensures the comparability of annotations across sub-corpora and languages. It is possible, for example, to focus on one specific TAS category (e.g. distortion errors or collocation errors) across sub-corpora or languages, thereby disentangling genre- or language-pair-specific difficulties from more universal areas of difficulty.

6. Conclusion

The enthusiasm that the MUST project has been met with shows that the time was ripe for the compilation of a large, multilingual student translation corpus. One of the reasons why the project has sparked so much interest is that it makes it possible to exploit student translations and their annotations, which are an integral part of many researchers' teaching activities, whether in translation or modern language departments. The standardized metadata and annotation systems allow all these data to be put to use in both research (whether theoretical or applied) and translation and foreign language teaching.

MUST is a truly collaborative project based on the principles of community sourcing (cf. Branzov 2016): MUST partners share data, take part in joint decision-making processes and reflect together on specific aspects of the project (e.g. TAS). The community has taken an active role from the very start of the project, with many fundamental decisions, for instance as regards corpus make-up and metadata, being taken at the kick-off meeting. The project, which brings together foreign language learning and translation scholars, testifies to the value of interdisciplinarity and heralds promising synergies between learner corpus research and translation studies.

Acknowledgements

We would like to thank the MUST local coordinators - Silvia Bernardini, Łucja Biel, Mario Cal Varela, Cem Can, Sara Castagnoli, Madalina Chitez, Elisa Corino, Julie Deconinck, Gert De Sutter, Margherita Dore, Gaetano Falco, Jonė Grigaliūnienė, Sandra Louise Halverson, Ruska Ivanovska-Naskova, Marlen Izquierdo, Xu Jiajin, Gurgen Karapetyan, Natalie Kübler, Efi Lamprou, Magnus Levin, Adriana Mezeg, Christine Michaux, Marina Morbiducci, Adriane Orenha Ottaiano, Adriana Orlandi, Heloísa Orsi Koch Delgado, Jun Pan, Anastasia Parianou, Gill Philip, Éric Poirier, Juan Pedro Rica Peromingo, Carola Strobl, Jenny Ström Herold, Olympia Tsaknaki, Jurgita Vaičenonienė, Susana Valdez, Heidi Verplaetse, Andrea Wurm - for contributing their translation data to the MUST project as well as for their helpful and enthusiastic support.

We would also like to thank the two anonymous reviewers for their helpful suggestions and comments.

References

Alfuraih, R.F. (2019). The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics. *Language Resources & Evaluation*. https://doi.org/10.1007/s10579-019-09472-6

Baker, M. (1993). Corpus linguistics and Translation Studies. Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology. In Honour of John Sinclair* (pp. 233–250). Amsterdam: John Benjamins.

Baker, M. (1995). Corpora in Translation Studies: An overview and some suggestions for future research. *Target,* 7(2), 223–243.
Bowker, L., & Bennison, P. (2002). Translation Tracking System: A tool for managing translation archives. *Proceedings of the Third International Conference on Language Resources and Evaluation* (pp. 503–507). Las Palmas, Canary Islands, 29-31 May 2002.
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 & New York: Routledge.
Branzov, T. (2016). Community-sourcing in virtual societies. *Serdica Journal of Computing*, 10(3-4), 263–284.
Castagnoli, S. (2009). *Regularities and variations in learner translations: A corpus-based study of conjunctive explicitation*. Unpublished PhD Thesis. Pisa University.
Castagnoli, S., Ciobanu, D., Kunz, K., Kübler, N., & Volanschi, A. (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.
Chesterman, A. (2007). Similarity analysis and the translation profile. *Belgian Journal of Linguistics*, 21, 53–66.
Cosme, C. (2008). Participle clauses in learner English: the role of transfer. In G. Gilquin, S. Papp, & M. B. Díez-Bedmar (Eds), *Linking up contrastive and learner corpus research* (pp. 177–198). Amsterdam & New York: Rodopi.
Dagneaux, E., Denness, S., & Granger, S. (1998). Computer-aided Error Analysis. *System: An International Journal of Educational Technology and Applied Linguistics*, 26(2), 163–174.
Díaz-Negrillo, A., & Fernández-Domínguez, J. (2006). Error tagging systems for learner corpora. *Revista Española de Lingüística Aplicada*, 19, 83–102.
Espunya, A. (2014). The UPF learner translation corpus as a resource for translator training. *Language Resources and Evaluation*, 48, 33–43.
Fictumova, J., Obrusník, A., & Stepankova, K. (2017). Teaching specialized translation error-tagged translation learner corpora. *Sendebar*, 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).
Gaspari, F., & Bernardini, S. (2010). Comparing non-native and translated language: Monolingual comparable corpora with a twist. In R. Xiao (Ed.), *Using Corpora in Contrastive and Translation Studies* (pp. 215–234). Newcastle: Cambridge Scholars Publishing.
Gillard, P., & Gadsby, A. (1998). Using a learners' corpus in compiling ELT dictionaries. In S. Granger (Ed.), *Learner English on Computer* (pp. 159–171). London & New York: Addison Wesley Longman.
Graedler, A.-L. (2013). NEST - a corpus in the brooding box. *Studies in Variation, Contacts and Change in English*, 13. http://www.helsinki.fi/varieng/series/volumes/13/graedler/
Granger, S. (1993). The International Corpus of Learner English. In J. Aarts, P. de Haan, & N. Oostdijk (Eds.), *English Language Corpora: Design, Analysis and Exploitation* (pp. 57–69). Amsterdam & Atlanta: Rodopi.
Granger, S. (1994). The learner corpus: a revolution in applied linguistics. *English Today* 10(3), 25–33.
Granger, S. (1996). From CA to CIA and back: an integrated contrastive approach to computerized bilingual and learner corpora. In K. Aijmer, B. Altenberg, & M. Johansson (Eds.), *Languages in Contrast. Text-based cross-linguistic studies* (pp. 37–51). Lund Studies in English 88. Lund: Lund University Press.
Granger, S. (2003). Error-tagged learner corpora and CALL: a promising synergy. *CALICO*, 20(3), 465–480.
Granger, S., & Lefer, M.-A. (2016). From general to learners' bilingual dictionaries: Towards a more effective fulfilment of advanced learners' phraseological needs. *International Journal of Lexicography*, 29(3), 279–295.
Halverson, S. (2017). Gravitational pull in translation. Testing a revised model. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds), *Empirical Translation Studies: New methodological and theoretical traditions* (pp. 9–45). Trends in Linguistics. Studies and Monographs. De Gruyter Mouton.
Hasselgård, H., & Johansson, S. (2011). Learner corpora and contrastive interlanguage analysis. In F. Meunier, S. De Cock, G. Gilquin, & M. Paquot (Eds), *A Taste for Corpora. In honour of Sylviane Granger* (pp. 33–62). Amsterdam: John Benjamins.
Johansson, S. (2007). *Seeing through Multilingual Corpora. On the use of corpora in contrastive studies.* Amsterdam & Philadelphia: John Benjamins.
Kruger, H. (2018). Expanding the third code: Corpus-based studies of constrained communication and language mediation. In S. Granger, M.-A. Lefer, & L. Penha-Marion (Eds.), *Book of Abstracts. Using Corpora in Contrastive and Translation Studies Conference (5th edition)*. CECL Papers 1 (pp. 9–12). Louvain-la-Neuve: Centre for English Corpus Linguistics/Université catholique de Louvain.
Kübler, N. (2008). A comparable Learner Translator Corpus: Creation and use. *LREC 2008 Workshop on Comparable Corpora*, 73–78.

Kutuzov, A., & Kunilovskaya, M. (2014). Russian Learner Translator Corpus: Design, Research Potential and Applications. In P. Sojka, A. Horák, I. Kopeček, & K. Palak (Eds), *Text, Speech and Dialogue* (pp. 315–323). Lecture Notes in Computer Science. Springer.

Lanstyák, I., & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures*, 13(1), 99–121.

Lapshinova-Koltunski, E. (2013). VARTRA: A comparable corpus for analysis of translation variation. *Proceedings of the 6th Workshop on Building and Using Comparable Corpora* (pp. 77–86). Sofia, Bulgaria, 8 August 2013.

Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, 5(3), 37–72.

Lefer, M.-A. (forthcoming). Parallel corpora. In M. Paquot, & S. Th. Gries (Eds), *Practical Handbook of Corpus Linguistics*. Springer.

Lüdeling, A., & Hirschmann, H. (2015). Error annotation systems. In S. Granger, G. Gilquin, & F. Meunier (Eds), *The Cambridge Handbook of Learner Corpus Research* (pp. 135–157). Cambridge: Cambridge University Press.

Macken, L., De Clercq, O. & Paulussen, H. (2011). Dutch Parallel Corpus: A Balanced Copyright-cleared Parallel Corpus. *Meta*, 56(2), 374–390.

Maingay, S., & Rundell, M. (1987). Anticipating learners' errors – implications for dictionary writers. In A.P. Cowie (Ed.), *The Dictionary and the Language Learner* (pp. 128–135). Tübingen: Niemeyer.

Obrusník, A. (2014). Hypal: A User-Friendly Tool for Automatic Parallel Text Alignment and Error Tagging. *Eleventh International Conference Teaching and Language Corpora* (pp. 67–69), Lancaster, 20-23 July 2014.

Obrusník, A. (2013). *A hybrid approach to parallel text alignment*. Bachelor thesis. Masaryk University.

Štěpánková, K. (2014). Learner Translation Corpus: CELTraC (Czech-English Learner Translation Corpus). Bachelor's Diploma Thesis. Masaryk University.

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

Uzar, R., & Waliński, J. (2001). Analysing the fluency of translators. *International Journal of Corpus Linguistics*, 6, 155–166.

Wible, D., Kuo, C.-H., Chien, F.-Y., Liu, A., & Tsao, N.-L. (2001). A Web-based EFL writing environment: Integrating information for learners, teachers, and researchers. *Computers & Education*, 37, 297–315.

Wurm, A. (2016). Presentation of the KOPTE Corpus and Research Project. https://www.academia.edu/24012369/Presentation_of_the_KOPTE_Corpus_and_Research_Project