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.

Learner translation corpora: Bridging the gap between learner corpus research and corpus-based translation studies

Sylviane Granger and Marie-Aude Lefer
Université catholique de Louvain

1. Introduction

Learner translation corpora (LTC) are corpora made up of translations produced by learners, who can be foreign language learners or translation students, translating into their native language or a foreign language. Although several corpora of this type have been collected in the last twenty years, it must be acknowledged that learner translation corpora remain relatively marginal in the fields of both learner corpus research (LCR) and corpus-based translation studies (CBTS). Apart from the fact that translation exercises are now rarely used in foreign language teaching[1], the main reason for the near-absence of LTC from the LCR scene is that they are not unequivocally recognized as fulfilling the criterion of authenticity a learner corpus is expected to meet. For Sinclair (1996) the default value for corpora is 'authentic': "All the material is gathered from the genuine communications of people going about their normal business" unlike data gathered "in experimental conditions or in artificial conditions of various kinds". To meet this criterion, most learner corpus collections contain data collected as naturally as possible, with as few constraints as possible imposed on the learner or the task (Granger, 2012). As a result, the most popular text types represented are free compositions in the case of writing and interviews in that of speech, both of which allow learners to choose their own wording and leave them a great deal of freedom regarding the ideas they want to express. Corpora that do not meet this criterion are considered as peripheral learner corpora:

> Collections of types of data that have been elicited with procedures exerting more control on the texts produced, such as compositions guided by pictures or student translations, are usually not considered learner corpora. Since the distinction between more or less controlled is, naturally, not clear-cut, such collections might be considered peripheral types of learner corpora" (Nesselhauf, 2004: 128).

For us, it is clear that learner translation corpora are bona fide learner corpora, of a partly different nature from those that are usually collected, but learner corpora nonetheless. From the perspective of LCR, they can admittedly be seen as constrained in the sense that the learner cannot write freely but has to transpose a prior text into another language (cf. Kotze, 2022), but this still leaves a great deal of flexibility regarding the wording used (lexis, grammar, word order, style, etc.). From a CBTS perspective, however, it is inappropriate to characterize translation as controlled and lacking in authenticity. For translation students the task of translating is fully natural and ecologically valid. Another distinctive feature of learner translation corpora is that, besides including translations into a foreign language (L2), they may

[1] However, several scholars have called for translation to be reintroduced in the foreign language classroom (see e.g. Cook, 2010; Koletnik Korošec, 2013; Tsagari & Floros, 2013).

include texts produced by students translating from an L2 into their native language (L1). It is interesting to include these texts because they provide evidence of difficulties encountered by learners, in particular those related to L2-to-L1 transfer.

The objective of this opening article is to provide an overview of learner translation corpus research. By their very nature, learner translation corpora are at the interface between LCR and CBTS. In Section 2 we offer a brief characterization of each field and suggest ways of integrating the two perspectives. Section 3 provides an overview of learner translation corpora and, more particularly, of issues related to corpus design and annotation. Section 4 draws up a catalogue of the main empirical and applied research strands in LTC-based research. The last section gives a brief description of each of the articles included in the special issue.

2. Interfacing learner corpus research and corpus-based translation studies

In this section, we describe the main tenets of LCR and CBTS. We then show how the perspectives they each offer can be integrated in order to advance corpus research on learner productions.

2.1 Learner corpus research

Learner corpus research was initiated in the late 1980s/early 1990s with a view to filling a gap in corpus linguistics, namely the absence of foreign/second language learner varieties from the wide range of language varieties – temporal, geographic, diatypic – already investigated with corpus methods and tools. Learner corpora are not collections of L2 data assembled at random. They are "[s]ystematic computerized collections of texts produced by language learners" (Nesselhauf, 2004: 125). In other words, they are "assembled according to explicit design criteria" (Granger, 2009: 14). Borin and Prütz's (2004: 69) definition makes this very clear: "A learner corpus is a collection of texts – written texts or transcribed spoken language – produced by language learners, and sampled so as to be representative of one or more combinations of situational and learner factors". For learner corpora to be useful for theoretical and applied purposes it is essential to compile them in a very rigorous manner. As pointed out by Gilquin (2015: 16), "[i]n the case of learner corpora, design criteria are especially crucial given the highly heterogeneous nature of interlanguage". As a result, learner corpora, whether spoken or written, are accompanied by a wide set of metadata pertaining to the learner (e.g. age, gender, first language, proficiency level) and the task (e.g. mode, genre, topic, timing). As shown by a comparison of the metadata in the *Cambridge Learner Corpus* (CLC) and the *International Corpus of Learner English* (ICLE) carried out by Barker, Salamoura and Saville (2015: 519), the number of metadata can be quite large (around 20), some of them shared across corpora and others varying in line with the specific aim of the corpus collection.

From its inception, LCR was conceived as having the objective of informing both theory and applications (Granger, 1998: 17):

> By offering more accurate descriptions of learner language than have ever been available before, computer learner corpora will help researchers to get more of the facts right. They will contribute to SLA [Second Language Acquisition] theory by providing answers to some yet unresolved questions such as the exact role of transfer. And in a more practical way, they will help to develop new pedagogical tools and classroom practices which target more accurately the needs of the learner.

Two main characteristics of learner corpora contribute to the achievement of this double objective: their relatively large size and representativeness of L2 learners, and the automated methods of annotation, extraction and analysis afforded by their electronic nature.

The methodology that has been used most extensively to analyse learner corpora is Contrastive Interlanguage Analysis (CIA; Granger, 1996, 2015), which involves two types of comparison: (1) one interlanguage language (IL) variety is compared with one or more corpora of native[2] speaker language (NL); (2) one or more learner varieties are compared with each other. The first type of comparison makes it possible to identify the linguistic features that distinguish learner language from expert language. These features can be errors but also instances of over- or underuse of specific words, phrases or structures. Thanks to the second type of comparison, it is possible to assess the degree of generalizability of distinctive features across learner populations that differ in terms of learner- and/or task-related variables.

The crosslinguistic perspective was prominent from the start in LCR. Many learner corpora are subcategorized on the basis of the learner's native language, the underlying idea being that transfer, i.e. the influence of the learners' L1[3] on their productions in the L2, is likely to result in L1-specific features above and beyond the developmental features shared by several L1 populations. The key role assigned to transfer in LCR was made fully explicit in the Integrated Contrastive Model (ICM) (Granger, 1996; Gilquin, 2000/2001). This model underscores the complementarity of CIA and Contrastive Analysis (CA), which establishes comparisons between different languages. The CA data included in the model are of two types: comparable corpora of original language (OL) in two or more languages and parallel corpora made up of source language (SL) and target language (TL). As represented in Figure 1, the ICM "involves constant to-ing and fro-ing between CA and CIA" (Granger, 1996: 46). CA enables analysts to formulate predictions about interlanguage which can be checked against CIA data. Conversely, CIA makes it possible to identify learners' interlanguage features which can be set against CA data to establish whether they are transfer-related. There have as yet been relatively few studies that implement this model. Those that do usually rely on comparable corpora rather than parallel corpora to carry out the cross-linguistic comparison (e.g. Xiao, 2007). Exceptions include Vanderbauwhede (2012), who uses parallel corpora of Dutch and French for the CA part of her study of demonstratives in L1 and L2 Dutch and French, and Hasselgård and Ebeling (2018), who investigate the translation paradigm of the English noun *people* and its Norwegian equivalents in the bidirectional English-Norwegian Parallel Corpus as a backdrop to investigating the use of these nouns by English learners of Norwegian and Norwegian learners of English.

[2] The second version of CIA (Granger, 2015) makes it clear that the reference corpus need not involve native language but may consist of any expert language variety against which researchers wish to set their IL data.

[3] Although the term *transfer* is generally used to refer to influence from the learner's native language, it can also involve influence from a second or third language.

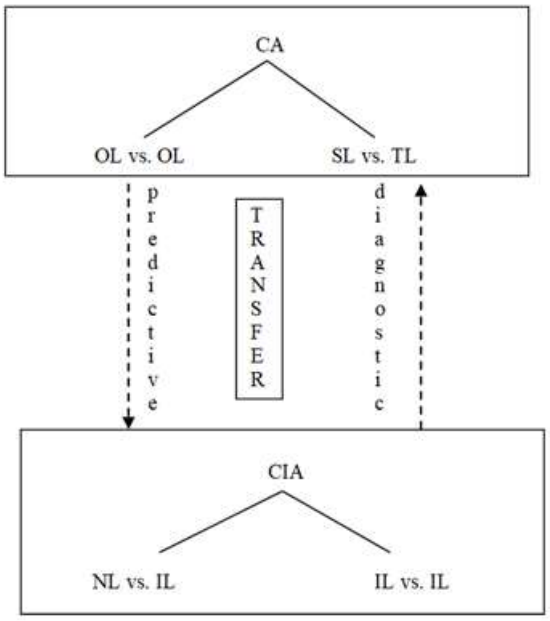

Figure 1: Integrated Contrastive Model (Granger, 1996: 47)

Learner corpus data can be raw, i.e. devoid of any form of annotation, or enriched with information about linguistic aspects of the texts. Although great benefit can be gained from using raw learner corpora, their usefulness is considerably increased when the corpora are linguistically annotated. This can be done automatically using part-of-speech (POS) taggers and parsers. While many learner corpora are available in POS-tagged format, parsing is still quite rare but is clearly gaining ground (see e.g. Schneider and Gilquin, 2016). However, not all types of annotation can be performed automatically. A range of semantic and discourse features, in particular, need to be annotated manually. This is time-consuming but leads to a considerable gain in time in subsequent analysis of the data. One type of annotation that is particularly relevant for LCR is error annotation. Whether to assess the degree of accuracy of interlanguage from a theoretical perspective or to identify errors that need to be remedied in teaching practice, it is useful to annotate errors using a standardized error annotation taxonomy and error annotation tool. Computer-Aided Error Analysis has become very popular in LCR: several annotation systems have been designed, as well as error annotation tools which allow researchers to annotate text files on the basis of their own error taxonomy (Díez-Bedmar, 2021).

One particularly important benefit of LCR is that it has brought to the fore aspects of learner language that had previously been under-researched. While SLA studies have tended to prioritize morphology and syntax, LCR has also devoted much attention to phraseology (including lexico-grammar) and discourse. The prevalence of single-word and multiword lexical units is a characteristic of all corpus studies and is due to both the ease with which words and phrases can be investigated on the basis of electronic data and the profound influence of John Sinclair's phraseological view of language (Herbst, Faulhaber and Uhrig, 2011). The types of phraseological unit that have been investigated the most in LCR are collocations and lexical bundles, which have proved to be extremely problematic for learners (for a survey, see Granger, 2019). Studies of discourse centre on cohesion and, more particularly, logical connectors (e.g. Leedham and Cai, 2013; Van Vuuren and Berns, 2018), which can be extracted automatically from learner corpora, and for which learner corpus data provide the type of continuous discourse necessary for their correct interpretation.

One of the main strengths of learner corpus research is that it helps to quantify learner language. For a long time researchers lacked a quantitative model of learner-specific characteristics and "were left to make do with approximations based on impressions, anecdotes and manual counts of small samples" (Milton and Tsang, 1993: 215-216). The quantifying objective of LCR needs to go hand in hand with a high degree of rigour in analysing the quantitative findings. This entails using statistical tests which over the years have progressively become highly sophisticated (Gries, 2015). Paquot and Plonsky's (2017) survey of LCR publications from 1991 to 2015 shows that there has been substantial progress over time in the statistical treatment of the data but that there is a need for improvement, as many studies still present shortcomings in the use and reporting of statistics.

2.2 Corpus-based translation studies

Corpus-based translation studies emerged at approximately the same time as learner corpus research with a series of papers by Baker (1993, 1995, 1996). In these early publications, which mark the birth of CBTS, Baker promoted the compilation and use of electronic corpora as a basis for investigating translational behaviour and the typical traits of translated texts in a systematic fashion. Until then, translations had been sidelined from mainstream corpus compilation projects because they were seen as instances of 'deviant' language use (mainly because of interference from the source language). Going against this traditional view, Baker (1993) convincingly argued that translations are authentic instances of communication in their recipient culture. She stressed that translated language is a variety in its own right, worthy of rigorous scientific investigation, thereby paving the way for intensive corpus research in translation studies.

From the very start, Baker (1993: 235) placed particular emphasis on the descriptive and theoretical objectives of CBTS:

> Large corpora will provide theorists of translation with a unique opportunity to observe the object of their study and to explore what it is that makes it different from other objects of study, such as language in general or indeed any other kind of cultural interaction. It will also allow us to explore, on a larger scale than was ever possible before, the principles that govern translational behaviour and the constraints under which it operates. Therein lie the two goals of any theoretical enquiry: to define its object of study and to account for it.

Building on pre-corpus translation research from the 1980s, Baker (1993) put forward the central construct of the translation universal. Translation universals were then defined as "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker, 1993: 243). In other words, they are recurrent characteristics of translated language, irrespective of the language pair or register under scrutiny, which are inherent in the translation process rather than the result of source-language influence or crosslinguistic contrasts. They include explicitation, normalization (standardization), simplification and levelling out (convergence). We will return to these features in more detail in Section 4.1. At this stage, however, it is important to point out that the notion of the translation universal has gradually made way for that of the translation feature (or feature of translated language), as corpus work in the last thirty years has clearly demonstrated that the universal nature of these properties does not hold.

The comparative analysis of translated vs non-translated (original) language promoted by Baker's 'translation universals' agenda implied that new corpora needed to be collected, namely corpora of translated texts. One such example is the *Translational English Corpus* (TEC)[4]. TEC is made up of fiction, biography, news and inflight magazines translated into English from a range of European and non-European source languages. It is enriched with metadata about the translators represented in the corpus (e.g. gender, main occupation, language background). In early corpus translation studies such as Laviosa (1998) and Olohan and Baker (2000), data extracted from TEC were typically compared with data drawn from comparable portions of the *British National Corpus*. Importantly, source texts were not included in TEC because Baker insisted that translations be studied in their own right, i.e. without reference to the prior text. This type of approach in CBTS is generally referred to as the monolingual comparable approach. However, it soon became clear that the lack of access to the source texts of the translations jeopardized the interpretation of corpus findings. Without access to the source texts, it is impossible to determine whether a given trend is inherent in the translation process or, rather, triggered by a certain phenomenon in the source text (see e.g. Laviosa, 1998: 9). As a result, the monolingual comparable approach has gradually given ground to more complex corpus designs, which typically include as their central components a parallel corpus, i.e. a corpus that contains translations aligned with their source texts, together with a comparable corpus of original texts in the target language. This type of corpus design, which combines a monolingual comparable component and a multilingual or bilingual parallel component, is graphically represented in Figure 2.

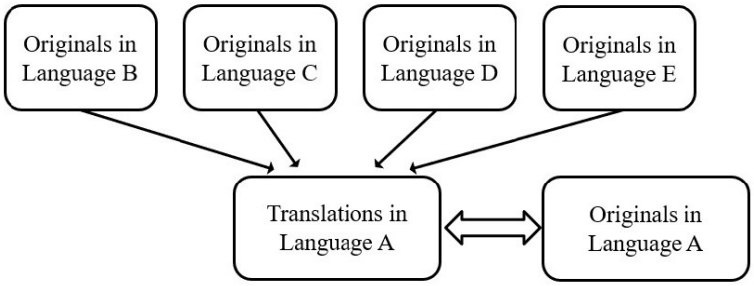

Figure 2: The monolingual-comparable-cum-parallel corpus design (Lefer, 2020: 267)

While Baker's early work laid particular emphasis on the theoretical relevance of corpora for translation studies, many CBTS researchers highlight the double-sided nature of the field. Like LCR, CBTS aims both to inform theory and to improve teaching. As early as 1998 Zanettin demonstrated the relevance of corpora for translator training, and this applied strand has grown increasingly active over the years (see e.g. Zanettin, Bernardini and Stewart, 2003; Beeby, Rodríguez-Inés and Sánchez-Gijón, 2009; Frankenberg-Garcia, 2015).

Dozens of parallel corpus collection initiatives have mushroomed in the last thirty years in multiple fields, such as natural language processing, contrastive linguistics and translation studies (see Lefer, 2020 for a detailed survey). The corpus projects initiated by translation scholars mostly cover a single language pair, often in the two translation directions, and a limited set of registers (typically novels, non-fictional prose, news items, administrative or legal documents). Multi-register parallel corpora are quite rare. Exceptions include CroCo for the

[4] https://www.alc.manchester.ac.uk/translation-and-intercultural-studies/research/projects/translational-english-corpus-tec/

German-English pair (Hansen-Schirra, Neumann and Steiner, 2012) and the Dutch Parallel Corpus (DPC) for Dutch-English and Dutch-French (Macken, De Clercq and Paulussen, 2011). Most parallel corpora are sentence-aligned and POS-tagged to allow more complex queries in corpus-linguistic tools. Parsing is also increasingly being used. Generally speaking, few metadata are available. Little is known about the translation conditions (e.g. the tools used), the translators who produced the translations (e.g. language background, translation experience, main occupation) and the translation workflow (e.g. revision). This is in sharp contrast with the rich metadata often included in learner corpora. The situation is improving, however, with the compilation of new-generation parallel corpora, such as the DPC 2.0 (Reynaert, Macken, Tezcan and De Sutter, 2021). Importantly, the vast majority of translation research is based on corpora of professional or expert translations (Lefer, 2020: 260-261). There are comparatively few corpora of learner or novice translations (see Section 3).

In a recent survey of CBTS, Granger and Lefer (2022) found that the linguistic focus of empirical translation studies is mostly on terminology and lexis (including measures of lexical variation and lexical density), grammar (e.g. passives, modals, nominalizations) and discourse (mostly connectors). Translation features still hold centre stage in present-day corpus research, especially explicitation. This notion is used to refer to cases where source-text phenomena are explicitated in translation (e.g. cultural references, logico-semantic links) and instances where translated language encodes grammatical information more explicitly than non-translated language (e.g. optional *that*-complementizer in English).

2.3 Integrating the two perspectives

Although the Integrated Contrastive Model promotes the combined use of learner and parallel corpora, the synergies between LCR and CBTS it has offered so far have been quite limited. The reason for this is that translations, when they are used, are not analysed in order to understand translational behaviour or typical features of translation products, but rather to identify crosslinguistic differences with a view to establishing transfer in learner language. The use of translations in the ICM is thus clearly linked to contrastive linguistics, rather than translation studies. As stated explicitly by Johansson (2007: 33), in contrastive linguistics the features characteristic of translated texts are usually not discussed. The only translation effect that is taken into consideration is source language influence. This perspective has drawn criticism from translation scholars. Olohan (2014: 27-28), for example, criticizes Altenberg's (1998) study of connectors and sentence openings on the grounds that his study "like much of the contrastive linguistic work of this kind, thus fails to recognize that translators' choices may be motivated by something other than language systemic conventions".

There is clearly scope for a closer integration of LCR and CBTS. In this section we discuss two synergetic pathways: one theoretical (the study of mediated discourse and constrained communication), the other methodological (Granger's 2018 Contrastive Translation Analysis). Other aspects which give evidence of the complementarity between the two fields are dealt with in subsequent sections of this article and in the special issue generally.

Translation is often framed as constrained language use, i.e. "language produced in communicative contexts characterised by particularly conspicuous constraints" (Kruger & Van Rooy, 2016: 27). Constrained communication encapsulates both translation and non-native

language production, including learner language. As stated by Chesterman (2017: 63), "both translations and learners' texts are produced under particular constraints, and it may be that these constraints have similar effects" (see also Chesterman, 2004: 10-11). In a seminal paper devoted to constrained communication, Lanstyák and Heltai (2012) examine two constraint dimensions along which instances of constrained language use can be compared: language activation (whether monolingual or bilingual) and text production (whether the text is mediated, i.e. dependent on a prior text, as in editing and translation, or not, as in L2 free writing). The authors draw several parallels between translation features and similar phenomena in bilingual communication (contact effects, simplification, loyalty to norms, explicitation), which they call language contact universals. Kotze (2019, 2022) has elaborated extensively on Lanstyák and Heltai's (2012) proposal by adding three further constraint dimensions: modality and register (spoken, written, multimodal), proficiency (native/highly proficient user vs learner) and task expertise (expert vs non-expert). In this framework, the comparison of constrained varieties "has the potential to illuminate unique sociocognitive aspects of language and text processing in translation, against the background of similarities with other constrained varieties" (Kotze, 2022: 90). The approach has been gaining ground in CBTS recently, with studies devoted to the systematic comparison of professional translation and L2 novice or expert writing (e.g. De Sutter & Lefer, 2020; Ivaska & Bernardini, 2020; Ivaska, Ferraresi & Bernardini, 2022; Neumann, Kerz & Heilmann, forthcoming). Some studies, such as Ferraresi (2019) and Kajzer-Wietrzny (2022), also tackle spoken language (interpreting vs non-native speech). Clearly, such initiatives can help bring CBTS and LCR closer together.

Substantiating the notion of constrained language requires a large and varied corpus base. A key challenge of this type of approach is obtaining comparable corpus data sets for the constrained varieties under investigation (e.g. in terms of topics, registers, task conditions and language proficiency). In other words, "corpus sourcing is a major challenge" (Ivaska et al., 2022: 134). In this context, the methodological framework put forward by Granger (2018), called Contrastive Translation Analysis (CTA), can be seen as an attempt at designing a multi-corpus empirical basis comprising both comparable and parallel multilingual corpora and learner corpora of the languages being compared. In Figure 3, the model is represented with English as the focal language. It includes large corpora of translated language in French (FR), Dutch (DU) and Chinese (CH), which can be used as wholes, i.e. with no reference to the source texts, but can also be broken down into SL-specific subcorpora, thereby allowing for pairwise comparisons between source and target texts. It also contains large comparable corpora of original, i.e. non-translated, texts in the three languages involved in the analysis. The model also includes learner corpora, broken down according to the learners' mother tongue background. The analysis of the two varieties – translated English and learner English – from different contact settings makes it possible to identify their commonalities. In addition, as shown by Granger's (ibid.: 196-199) analysis of markers of contrast in native, learner and translated texts, the model provides a particularly strong basis for establishing the presence of source language effects. More generally, as both learner and translated texts are broken down in terms of L1/SL, they offer a particularly powerful basis for teasing out L1/SL effects from general features of acquisition/translation, a key issue in both LCR and CBTS.

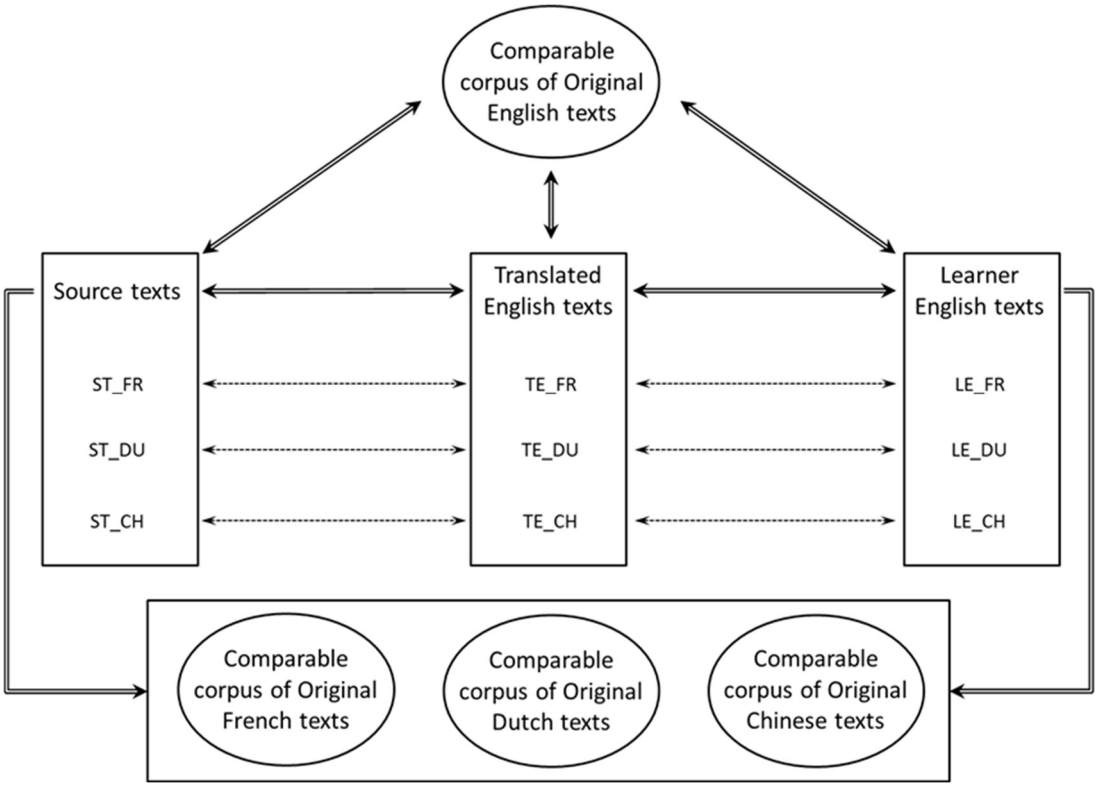

Figure 3: Contrastive Translation Analysis (Granger, 2018: 190)

Although the studies mentioned in this section give clear evidence of a rapprochement between LCR and CBTS, the links between the two fields remain relatively tenuous. Collecting and analysing learner translation corpora which, by their very nature, integrate the two fields, promises to be an effective way of bridging that gap.

3. Overview of learner translation corpora

Learner translation corpora can be defined as parallel corpora containing authentic translations produced by translation students or foreign language learners in real-life tasks (e.g. in the translation classroom, as opposed to a laboratory setting), aligned with their source texts. LTC are also referred to as 'translation learner corpora' and 'learner translator corpora'. In addition to representing novice, inexperienced translators, LTC differ from the parallel corpora traditionally used in CBTS in that many of them are multiple translation corpora, i.e. they include a number of translations of the same source text, often in large quantities.

The first LTC were collected in the early 2000s, approximately ten years after the emergence of CBTS. Early LTC projects include the *Polish and English Language Corpora for Research and Applications* (PELCRA; Uzar & Walinski, 2001) and the *Student Translation Archive* (STA; Bowker & Bennison, 2003). PELCRA and STA were soon followed by similar initiatives, such as the *Multiple Italian Student Translation Corpus* (MISTiC; Castagnoli, 2009), the multilingual *MeLLANGE Learner Translator Corpus* (Kübler, 2008; Castagnoli, Ciobanu, Kübler, Kunz & Volanschi, 2011), the *Norwegian-English Student Translation* corpus (NEST; Graedler, 2013), the *Universitat Pompeu Fabra (Barcelona) – Learner Translation*

Corpus (LTC-UPF; Espunya, 2014), the *Russian Learner Translator Corpus* (RusLTC; Kutuzov & Kunilovskaya, 2014), the French-German KOPTE (Wurm, 2016), the *Czech-English Learner Translation Corpus* (CELTraC; Fictumova, Obrusnik & Stepankova, 2017) and the *Undergraduate Learner Translator Corpus* (ULTC; Alfuraih, 2020). Interestingly, the majority of LTC projects emanate from the field of translation studies rather than learner corpus research. It should be borne in mind, however, that even in CBTS they can still be considered a niche area. As stated above, present-day corpus translation research mostly relies on corpora of professional or expert translations, to the detriment of other translator profiles.

As is the case with CBTS as a whole, the main objectives of LTC research are both applied and theoretical, as shown by the following quote by Bowker & Bennison (2003) in relation to the STA:

> Student translators can be considered as a highly specialized type of language learner/user. Although their specific needs differ from those of general language learners, a similar approach to collecting and studying the output of student translators would be highly valuable for both **pedagogical and research applications**. With regard to **pedagogy**, a corpus of student translations can provide a means of identifying areas of difficulty that could then be integrated into the curriculum and discussed in class. 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. [our emphasis]

This is echoed by Espunya (2014: 35), who states that LTC have "pedagogical aims, both theoretical, i.e. research into the acquisition of the translating competence and the role of training methodologies, and applied, i.e. developing materials for translator training". Kutuzov & Kunilovskaya (2014) propose a structured research agenda for their RusLTC-based work, also striking a balance between theory (e.g. the issues of variation and choice in translation) and applications (e.g. identification of problem areas).

With the exception of the EU-funded MeLLANGE and the RusLTC (which is collaboratively collected by a consortium of Russian universities), LTC are local projects. As a corollary, they tend to be restricted to a single language pair, often in one direction, and are relatively small in size (they typically contain between 150 and 500 learner translations). They are also very diverse in terms of the range of registers represented (news, fiction, legal texts, administrative texts, etc.) and the type of metadata they include (which range from very basic to highly sophisticated metadata sets). Most projects focus on translations into the students' native language (L2 to L1 translation), with few exceptions.

In terms of corpus annotation, it appears that POS-tagging is not standard practice. What most LCT share, however, is error annotation. In line with the applied objectives of learner translation corpus research, a good number of translation error taxonomies have been specifically designed for LTC. Once again, we see a lot of heterogeneity, with both coarse-grained taxonomies of five error categories and more complex taxonomies containing 50+ categories. Some error taxonomies are well documented, such as the ones developed for the English-Russian and French-German pairs by the RusLTC and KOPTE teams respectively, while others have very limited documentation, which poses serious issues of annotation consistency.

A newcomer to the field is the *Multilingual Student Translation* corpus (MUST; Granger & Lefer, 2020), which is an ongoing international LTC collection initiative that brings together

more than 30 partner teams worldwide. The MUST corpus currently comprises ca 400 source texts (ranging from 150 to 1,000 words in length) and 6,500 student translations produced by ca 2,500 students, with 18 languages represented. In addition to being truly multilingual, MUST is multi-register. It includes numerous text types, both general (news and opinion articles, excerpts from novels, etc.) and specialized (financial reports, tourist guides, instruction manuals, contracts, etc.). The strengths of the MUST corpus include its rich standardized metadata relating to the source texts, translation tasks and learners (40+ metadata rubrics; see Granger & Lefer 2020 for a full overview), and the Translation-oriented Annotation System (TAS) developed collaboratively within the MUST network to support both translator training and research on translation quality across language pairs (Granger & Lefer, 2021). Another recent project is DiHuTra (Lapshinova-Koltunksi, Popović & Koponen, 2022), which contains English news and reviews and their Croatian, Finnish and Russian translations by both professionals and students. Such a corpus design, where the same source texts are translated by expert and novice translators, makes it possible to examine the impact of translation expertise on the linguistic profiles of translational products.

4. Main research strands in LTC-based research

LTC-based research to date has examined a broad range of topics, but two main strands dominate the field. First, there are theory-oriented studies that examine translation features in learner translations (Section 4.1). Alongside this group of studies, there is an applied research strand, mainly focused on computer-aided translation error analysis and translation quality evaluation, with a view to informing translation pedagogy and devising corpus-informed teaching materials and activities (Section 4.2). In recent years some additional trends have started to emerge, greatly contributing to expanding the scope of learner translation corpus research (Section 4.3).

4.1 Translation features

As stated in Section 2.2, translation features, such as explicitation, normalization (standardization), simplification, and levelling-out (convergence), are intensively researched in CBTS. Source-language influence, also referred to as interference or shining-through in CBTS, though not included in Baker's (1993, 1996) initial inventory of translation universals, has also received a lot of attention in the field, often in tandem with normalization (understood as adherence to target-language norms) (cf. Lefer & Vogeleer, 2013). Some researchers resort to the term translationese to refer to these features considered together: "translationese is used as a general non-evaluative term to refer to the quantitative linguistic features of translations that set them apart from non-translations in the same language" (Kunilovskaya, Morgoun & Pariy, 2018: 34). At present the body of corpus work devoted to the topic is largely based on professional translations. Generally speaking, what emerges is that translated texts tend to be more explicit, more standard, simpler (lexically and syntactically) and more homogeneous than their source texts (parallel approach) and/or comparable non-translated texts in the same language (monolingual comparable approach), with variation across language pairs, translation directions, registers, translation modes, etc. The linguistic operationalizations of translation features are quite stable across studies: typical examples include (1) connectors and optional

that-complementizer for explicitation, (2) hapax legomena, *n*-grams and contracted forms for normalization, and (3) lexical density, core vocabulary coverage and average sentence length for simplification. Levelling-out is often investigated through the prism of lexico-syntactic simplification variables.

In recent years, these translation features have also been examined in learner translations. This is a much-needed endeavour as the "typical linguistic features of learner translations as opposed to professional ones are only tentatively described" (Kunilovskaya et al., 2018: 33). One possible approach is to compare learner translations with original texts in the target language. On the basis of English-to-Italian learner translations and comparable texts in original Italian, Castagnoli (2016: 344) sets out to assess whether learner translations comply with typical Italian patterns of interclausal linkage (normalization) or whether they show traces of source-language interference. Her study of connectives reveals that there is strong interference from English in learner translations, with clear signs of normalization as well (compliance with target-language norms). Looking at lexico-syntactic simplification in learner translations into the L1 and the L2 for the English-French pair, Penha-Marion, Gilquin & Lefer (forthcoming) find that learner translations into the L2 display more features of simplification than into the L1. Differences are also observed between students with different degrees of translation experience: the students who are the most inexperienced in translation (in terms of training) produce the simplest outputs.

Other studies examine learner and professional translations contrastively (e.g. Redelinghuys & Kruger, 2015; Kunilovskaya et al., 2018; Lapshinova-Koltunski 2022). Interestingly, there are differences in the assumptions as to expected variation trends across expertise levels. For example, Kunilovskaya et al.'s (2018: 36) expectation is that "there is a gradient in features, which distinguish translations from non-translations and make learner output more pronounced translationese than professional translations". By contrast, Redelinghuys & Kruger (2015) propose alternative starting-point hypotheses whereby some features are expected to be stronger in translations produced by inexperienced translators (simplification through low lexical variation), while others are expected to be more marked in translations produced by experienced translators (explicitation, normalization). Their underlying assumption is that "these features may be the consequence of language processing and/or translation strategies associated with translation expertise" (ibid, 296). Studies so far have uncovered relatively few significant differences between learner and professional translations. This may well be due to the fact that the linguistic operationalizations traditionally used to study translation features do not capture translation expertise and hence that other types of linguistic phenomena need to be examined to better characterize learner translations vis-à-vis professional translations.

4.2 Teaching applications

All LTC projects emphasize the benefits that can be derived from using student translations for translator training.[5] As is the case in LCR, pedagogical applications can be seen to fall into two categories, according to whether they involve immediate or delayed pedagogical use (Granger, 2009: 20-22). In the former case, students' translation data are used in the classroom by the

[5] This section focuses exclusively on translator training, as the use of learner translation data is marginal in the foreign language teaching context.

students who have produced the translations. In the latter, the data are collected cumulatively over time with a view to producing tailored teaching materials and redesigning the translation syllabus. The two functions can be combined: the data can be used by the students who have produced them as well as by students in subsequent years who are following the same curriculum.

The pedagogical benefits are particularly noteworthy if the corpora are annotated for errors, as annotations help teachers "identify the most common difficulties within a given group of learners, thus indicating areas of the learning curriculum where teaching is most needed" (Castagnoli et al., 2011: 239). In most cases, error annotation is integrated into a data management platform which makes it possible to store, manage and query learner translations and accompanying metadata. Fictumova et al. (2017) provide a detailed description of the numerous affordances of such a platform. Teachers draw parallel concordances of specific words and phrases and search for specific error categories. They can also generate error statistics for individual students or student cohorts and, if the data are collected longitudinally, track students' development over a given period of time. Metadata can also be included in the queries and be used by teachers to assess the impact of factors such as task or translation experience on the quality of translations. Error-annotated data are also potentially very useful for students, as they receive structured feedback on their work with well-defined systematic annotations (Granger & Lefer 2020) and can access their own error reports.

The range of classroom activities that can be designed on the basis of LTC data is extremely wide. Kübler (2008) and Kübler, Mestivier and Pecman (2018, 2022) provide examples of activities designed to tackle the main difficulties encountered by students of specialised translation, in particular those related to complex noun phrases, which are extremely frequent in specialised texts and prove to be especially error-prone. Some of the suggested activities require students to consult a corpus of specialized texts in the same domain as their translation task in order to check the acceptability of some of the terms used in the LTC and to identify more appropriate translation solutions. Students can also be presented with concordances of specific error types such as false friends and asked to discuss each error in context and to suggest correct translations. Espunya (2014) describes how she has used error-tagged LTC data to design a whole grammar unit revolving around information packaging mechanisms and argumentative relations.

As shown by Kunilovskaya, Ilyushchenya, Morgoun and Mitkov's (2022) study, a rigorous analysis of learners' errors can set the course for a more empirically motivated educational curriculum. Comparing a set of particularly error-prone SL items extracted from LTC data with the items most focused on in translation textbooks, the authors establish a wide gap between the two sets. For example, the study shows that textbooks tend to focus on grammatical issues while learner difficulties are primarily lexical and often involve multiword units other than the idiomatic/figurative expressions covered in textbooks. Textbooks are also shown to disregard students' difficulties with discourse issues, in particular those related to thematic and information structure.

Although most activities rely on error-annotated data, raw data, i.e. learner translations devoid of any annotations, can also be of great benefit. For example, as suggested by Castagnoli et al. (2011), students can be presented with concordance lines illustrating specific translation problems, and asked to detect the errors and provide alternative solutions. Raw data also allow

for the design of activities that do not involve errors at all. Kübler (2008: 77) describes a 'strategy-oriented approach' intended to trigger "a reflection and a discussion in the classroom about different translation strategies". This approach is reminiscent of that advocated by Seidlhofer (2002) in LCR, itself based on Swain's (1985: 141) reflective approach to students' output. Seidlhofer describes classroom activities that give learners the opportunity to reflect on short texts they have produced and highlights the motivating effect for students of working on their own language productions. One way of transposing this approach to translation is to expose students to multiple learner translations, thereby "triggering reflection on variation and translation acceptability, as students are allowed to analyse pros and cons of different translation solutions at the same time" (Castagnoli et al., 2011: 246). This type of language-awareness activity has the potential to enhance students' assessment and editing skills.

The pedagogical benefits of LTC data extend beyond teaching materials. Reference materials, particularly dictionaries, also stand to gain from insights derived from LTC. In LCR, corpora of learner writing have been used to design usage and error notes which are incorporated into monolingual lexicographical resources, such as the *Macmillan English Dictionary* (Rundell & Granger, 2007) and the *Louvain English for Academic Purposes Dictionary* (Granger & Paquot 2015). Bowker (2003) suggests extending this practice to bilingual dictionaries, using learner translation corpora. Granger and Lefer (2016) provide examples of usage and error notes that can help 'learnerize' bilingual dictionaries, i.e. bring them closer to learners' attested needs.

4.3. Some emerging trends

As shown in Section 4.2, error-based translation quality assessment holds centre stage in applied translation studies based on LTC. However, Kunilovskaya et al. (2018: 35) rightly point out that "the notorious subjectivity in translation assessment can be rooted in the overall textual features of translated texts that are hard to pin down in terms of local translation errors, but which can be described quantitatively". This new line of thinking had led to the development of automated, objective approaches that rely on large sets of linguistic and textual features, rather than error analysis, to assess quality in student translations. The basic tenet of such approaches is that differences between learner translations and comparable original texts in the same language (L1 expert writing) or professional translations are indicative of low translation quality (e.g. De Sutter, Cappelle, De Clercq, Loock & Plevoets,2017; Kunilovskaya & Lapshinova-Koltunski, 2019). Results to date, however, suggest that there is no direct link between linguistic deviations from professional productions (whether free writing or translation) and low quality in learner translations. More research is needed to confirm or disconfirm these initial findings.

Technology is also driving new developments in learner translation corpus research. The advent of neural machine translation has marked a major turning point in the translation industry, where language professionals are now frequently asked to post-edit machine translations rather than producing translations from scratch (cf. Ginovart Cid, Colominas & Oliver, 2020). Quite naturally, machine translation post-editing (MTPE) is being increasingly incorporated into translation curricula across the world. This technological turn is driving new initiatives in learner translation corpus research as well, such as the comparison of learner translations with machine translations. Loock (2020), for instance, examines non-canonical word order phenomena such as clefting and subject-verb inversion in English-to-French learner and

machine translations with a view to uncovering translation students' added value over the machine and, more generally, empowering them. Another interesting development is the compilation and error annotation of student post-editing corpora. Focusing on post-editing errors related to complex noun phrases in specialized discourse, Kübler et al. (2022) distinguish between three categories of error: Overconfidence in MT (leaving incorrect MT output unchanged), Underconfidence in MT (modifying correct MT output), and Failure to correct MT (spotting an error in MT and failing to correct it adequately). The most frequent type of error related to noun phrases in their corpus is Failure to correct MT. Along similar lines, Lefer, Piette and Bodart (2022) propose an MTPE annotation system called MTPEAS (Machine Translation Post-Editing Annotation System). MTPEAS consists of seven categories: Value-adding edits, Successful edits, Unnecessary edits, Incomplete edits, Error-introducing edits, Unsuccessful edits, and Missing edits. Combined with TAS (Granger & Lefer, 2021) in order to specify the nature of the erroneous segments still present in the final post-edited texts, the taxonomy makes it possible to systematize annotation of student post-editing corpora.

Alongside these applied initiatives, we also see some new trends emerging in theory-oriented LTC-based research. For example, Wurm (2020) uses KOPTE-derived error annotations to uncover empirical evidence on the development of translation competence. In particular, she is interested in "modelling (…) certain aspects of translation competence for groups of trainees, namely the effects of a stay abroad and media consumption, and some timeline effects in translation competence development, such as (fewer) errors and (more) good solutions" (ibid.: 141). Her findings reveal strong timeline effects, with intensive training leading to fewer errors, more 'good solutions' and/or higher translation speed. By contrast, stays abroad and media consumption do not seem to play a role in translation competence acquisition. Another promising line of enquiry is the study of variation and invariance in translation. Very early on, in the first volume dedicated to CBTS, Malmkjaer (1998: 6) stated that

> Although parallel corpora provide evidence of how languages relate to each other in use, we still only get one individual's introspection on each individual instance contained in the corpus. But (…) translators' opinions may differ on individual instances in individual contexts. We may suspect that where they differ most is where investigation might prove particularly fruitful. As parallel corpora are constructed at the moment, these cases would not come to light.

While it is true that very few professional texts are translated multiple times, most LTC contain several translations of the same source texts and allow systematic research into translation variation and invariance. Relying on the multiple-translation MISTiC corpus, Castagnoli (2020) finds that "full lexical invariance is basically limited to the translation of some concrete nouns, some functional items and numbers, whereas abstract nouns and metaphorical usage trigger more variation" (see also Castagnoli, this volume). Other emerging trends include the comparison of different learner varieties, such as L2 free writing and translation into the L2, with a view to uncovering their commonalities and differences (see Bernardini & Ferraresi, this volume).

5. Overview of the special issue

Although several learner translation corpora have been collected since the early 2000s and several studies have been carried out, to date there has not been any publication entirely devoted to learner translation corpora. The objective of this special issue is to introduce this fast-

developing research field and to illustrate its potential via four empirical studies. The core corpora used in the studies are learner translation corpora, both into the L1 and into the L2, which are complemented with parallel corpora of professional translations or large reference corpora in the target language. Most of the corpora used are enriched with linguistic annotation, such as part-of-speech tagging, parsing and error annotation. The topics investigated range from lexis and phraseology (e.g. evaluative adjectives, terminological collocations, dependency-based bigrams) to syntax and discourse (subject placement).

The first study, by Agnieszka Leńko-Szymańska and Łucja Biel, is situated in the field of legal translation. It focuses on specialised terminology and, more particularly, on verb-object terminological collocations. The main objective is to examine how Polish translation students deal with this type of word combination when translating into their L2. The learner translation corpus is made up of 54 Polish-to-English translations of a 350-word legal text, set against a comparable corpus of translations of the same source text by nine Polish professional translators and two large reference corpora of non-translated English legal texts. A detailed analysis of collocation equivalents used by the translation students and the professional translators brings out a relatively high degree of convergence. The collocational choices prove to be mainly conditioned by properties of the terminological collocations, in particular their degree of congruence. A search in reference corpora of legal texts for collocations used by learners reveals a sizeable proportion of unattested and inadequate combinations. A qualitative analysis of collocation errors shows that language errors are much less frequent than information transfer and naturalness errors, with the former being recurrent and systematic and the latter more idiosyncratic.

The second study, by Gert De Sutter, Marie-Aude Lefer and Bram Vanroy, examines four general cognitive constraints (syntactic priming, cognitive routinisation, markedness of coding and structural integration) and their impact on the linguistic traits of translations produced by learners and experts. To do so, the authors take as a test case subject placement in Dutch (preverbal vs postverbal position), relying on a parallel corpus comprising eight French source texts (news items), each translated by both translation students and experienced professional translators, all working into their native language (L1 Dutch). The four cognitive constraints under scrutiny are operationalized by means of ten variables, such as inter- and intratextual priming, complexity of the verb phrase and subject discourse status. The mixed-effects regression analysis shows that the constraints shape student and professional translators' output similarly: priming and structural integration have the strongest impact on subject placement, while cognitive routinization does not have any significant effect. Despite the many similarities between placement patterns in student and professional translations, striking differences emerge in cases where French source sentences start with an adjunct, which the authors interpret as a difference in automatization when dealing with specific crosslinguistic differences between French and Dutch.

In the third article of the special issue, Sara Castagnoli explores translator choices through the lens of variation and invariance in a multiple learner translation corpus. Specifically, the study is based on an English source text translated into L1 Italian by 35 translation students. The corpus is supplemented with large reference corpora to check the acceptability of the solutions found in student translations. Taking as a starting point specific lexical items in the source text under investigation (multiword units and adjectives), the analysis of parallel corpus data makes it possible to distinguish between sets of source items that trigger variation in student

translations vs those that do not, i.e. items that are translated with the same dominant crosslinguistic equivalent across target texts. In line with previous findings, the study shows that idiomatic multiword units and evaluative adjectives tend to give rise to a wide variety of translation solutions in student productions. The same holds for lexical items that lack a literal equivalent in the target language. By contrast, invariance is mostly observed for unidiomatic multiword units and neutral adjectives.

The last study, by Silvia Bernardini and Adriano Ferraresi, compares two types of constrained communication – translated language and learner language – on the basis of Halverson's (2017) Gravitational Pull Hypothesis. More particularly, it investigates three sources of constrainedness effects: TL salience, SL prominence and connectivity, i.e. the strength of cross-linguistic links. As in Leńko-Szymańska and Biel's study, the focus is on collocations, but in this case the collocations are non-specialised and extracted automatically on the basis of dependency-based syntactic patterns such as adjective-noun or adverb-adjective. The learner and translation corpora consist respectively of 106 essays and 131 translations, all produced by Italian students of English of the same proficiency level and degree of task expertise. However, as the two subcorpora differ in several other respects, the authors have implemented a series of methodological steps to reduce the impact of the comparability issues inherent in naturalistically collected data sets. A quantitative and qualitative analysis of 16 types of dependency collocations shows that translations display higher lexical association scores than essays. Halverson's TL salience is therefore found to be specifically translational. Connectivity and SL prominence, on the other hand, play an equally important role in the two constrained varieties.

Acknowledgements

We would like to thank the two general editors of the *International Journal of Learner Corpus Research* for giving us the opportunity to guest-edit a special issue on learner translation corpora. We also thank the reviewers of the articles included in the issue for their constructive feedback.

References

Alfuraih, R.F. (2020). The undergraduate learner translator corpus: a new resource for translation studies and computational linguistics. *Language Resources & Evaluation*, *54*, 801–830.

Altenberg, B. (1998). Connectors and sentence openings in English and Swedish. In S. Johansson, & S. Oksefjell (Eds.), *Corpora and Cross-Linguistic Research* (pp. 115–143). Amsterdam: Rodopi.

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. 233–50). Amsterdam and Philadelphia: John Benjamins.

Baker, M. (1995). Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. *Target*, *7*(2), 223–243.

Baker, M. (1996). Corpus-based Translation Studies: The Challenges that Lie Ahead. In H. Somers (Ed.), *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager* (pp. 175–86). Amsterdam: John Benjamins.

Barker, F., Salamoura, A., & Saville, N. (2015). Learner corpora and language testing. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 511–533). Cambridge: Cambridge University Press.

Beeby, A., Rodríguez-Inés, P. & Sánchez-Gijón, P. (Eds.). (2009). *Corpus Use and Translating: Corpus use for learning to translate and learning corpus use to translate*. Amsterdam: John Benjamins.

Borin, L., & Prütz, K. (2004). New wine in old skins? A corpus investigation of L1 syntactic transfer in learner language. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and Language Learners* (pp. 67–87). Amsterdam: Benjamins.

Bowker, L. (2012). Meeting the needs of translators in the age of e-lexicography: Exploring the possibilities. In S. Granger, & M. Paquot (Eds.), *Electronic Lexicography* (pp. 379–397). Oxford: Oxford University Press.

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 and New York: Routledge.

Castagnoli, S. (2020). Translation choices compared: Investigating variation in a learner translation corpus. In S. Granger & M.-A. Lefer (Eds.), *Translating and Comparing Languages: Corpus-based Insights. Corpora and Language in Use Proceedings 6* (pp. 25–44). Louvain-la-Neuve: Presses universitaires de Louvain.

Castagnoli, S. (2016). Investigating trainee translators' contrastive pragmalinguistic competence: a corpus-based analysis of interclausal linkage in learner translations. *The Interpreter and Translator Trainer*, *10*(3), 343–363.

Castagnoli, S., Ciobanu, D., Kübler, N., Kunz, K., & 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.

Chesterman, A. (2004). Hypotheses about translation universals. In G. Hansen, K. Malmkjaer, & D. Gile (Eds.), *Claims, Changes and Challenges in Translation Studies* (pp. 1–14). Amsterdam: John Benjamins.

Cook, G. (2010). *Translation in Language Teaching: An Argument for Reassessment*. Oxford: Oxford University Press.

De Sutter, G., Cappelle, B., De Clercq, O., Loock, R., & Plevoets, K. (2017). Towards a corpus-based, statistical approach to translation quality: Measuring and visualizing linguistic deviance

in student translation. *Linguistica Antverpiensia, New Series: Themes in Translation Studies*, *16*, 25–39.

De Sutter, G., & Lefer, M.-A. (2020). On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*, *28*(1), 1–23.

Díez-Bedmar, M. B. (2021). Error analysis. In N. Tracy-Ventura, & M. Paquot (Eds.), *The Routledge Handbook of Second Language Acquisition and Corpora* (pp. 90–104). New York & London: Routledge.

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

Ferraresi, A. (2019). Collocations in contact: Exploring constrained varieties of English through corpora. *Textus: English Studies in Italy*, *1/2019*, 203–222.

Fictumova, J., Obrusnik, A., & Stepankova, K. (2017). Teaching specialized translation error-tagged translation learner corpora. *Sendebar*, *28*, 209–241.

Frankenberg-Garcia, A. (2015). Training translators to use corpora hands-on: challenges and reactions by a group of thirteen students at a UK university. *Corpora*, *10*(3), 351–380.

Gilquin, G. (2000/2001). The Integrated Contrastive Model: Spicing up your data. *Languages in Contrast*, *3*(1), 95–123.

Gilquin, G. (2015). From design to collection of learner corpora. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 9–34). Cambridge: Cambridge University Press.

Ginovart Cid, C., Colominas, C., & Oliver, A. (2020). Language industry views on the profile of the post-editor. *Translation Spaces*, *9*(2), 283–313.

Graedler, A.-L. (2013). NEST—A corpus in the brooding box. *Studies in Variation, Contacts and Change in English*, *13*.

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: Lund University Press.

Granger, S. (1998). The computerized learner corpus: a versatile new source of data for SLA research. In S. Granger (Ed.,) *Learner English on Computer* (pp. 3–18). London & New York: Addison Wesley Longman.

Granger, S. (2009). The contribution of learner corpora to second language acquisition and foreign language teaching: A critical evaluation. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 13–32). Amsterdam and Philadelphia: John Benjamins.

Granger, S. (2012). How to use foreign and second language learner corpora. In A. Mackey & S. Gass (Eds.), *Research Methods in Second Language Acquisition: A Practical Guide* (pp. 7-29). Malden: Blackwell.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24.

Granger, S. (2018). Tracking the third code: A cross-linguistic corpus-driven approach to metadiscursive markers. In A. Cermakova, & M. Mahlberg (Eds.), *The Corpus Linguistics Discourse* (pp. 185–204). Amsterdam: John Benjamins.

Granger, S. (2019). Formulaic sequences in learner corpora: Collocations and lexical bundles. In A. Siyanova-Chanturia, & A. Pellicer-Sanchez (Eds.), *Understanding Formulaic Language: A Second Language Acquisition Perspective* (pp. 228–247). London: Routledge.

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.

Granger, S., & Lefer, M.-A. (2020). The Multilingual Student Translation corpus: a resource for translation teaching and research. *Language Resources and Evaluation*, *54*, 1183–1199.

Granger, S., & Lefer, M.-A. (2021). *Translation-oriented Annotation System manual (Version 2.0). CECL Papers 3*. Louvain-la-Neuve: Centre for English Corpus Linguistics/Université catholique de Louvain.

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* (pp. 13–41). London: Bloomsbury.

Granger, S., & Paquot, M. (2015). Electronic lexicography goes local: Design and structures of a needs-driven online academic writing aid. *Lexicographica - International Annual for Lexicography* 31(1), 118–141.

Gries, S. Th. (2015). Statistics for learner corpus research. In S. Granger, G. Gilquin, & F. Meunier (Eds.), *The Cambridge Handbook of Learner Corpus Research* (pp. 159–181). Cambridge: Cambridge University Press.

Halverson, S. L. (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). Berlin: De Gruyter.

Hasselgård, H., & Ebeling, S.O. (2018). At the interface between Contrastive Analysis and Learner Corpus Research: A parallel contrastive approach. *Nordic Journal of English Studies*, *17*(2), 182–214.

Hansen-Schirra, S., Neumann, S., & Steiner, E. (2012). *Cross-linguistic Corpora for the Study of Translations. Insights from the Language Pair English-German*. Berlin: De Gruyter.

Herbst, T., Faulhaber, S., & Uhrig, P. (Eds.). (2011). *The Phraseological View of Language. A Tribute to John Sinclair.* Berlin & Boston: Walter de Gruyter.

Ivaska, I., & Bernardini, S. (2020). Constrained language use in Finnish: A corpus-driven approach. *Nordic Journal of Linguistics*, *43*(1), 33–57

Ivaska, I., Ferraresi, A., & Bernardini, S. (2022). Syntactic properties of constrained English: A corpus-driven approach. In S. Granger, & M.-A. Lefer (Eds.), *Extending the Scope of Corpus-Based Translation Studies* (pp. 133–157). London: Bloomsbury.

Jantunen, J. H. (2004). Untypical Patterns in Translations. Issues on Corpus Methodology and Synonymity. In A. Mauranen, & P. Kujamäki (Eds.), *Translation Universals - Do they Exist?* (pp. 101–126), Amsterdam: John Benjamins.

Kajzer-Wietrzny, M. (2022). An intermodal approach to cohesion in constrained and unconstrained language. *Target*, *34*(1), 130–162.

Koletnik Korošec, M. (2013). Translation in Foreign Language Teaching. In N. K. Pokorn, & K. Koskinen (Eds.), *New Horizons in Translation Research and Education 1*. Publications of the University of Eastern Finland Reports and Studies in Education, Humanities and Theology, 61–74.

Kotze, H. (2019). Converging what and how to find out why: An outlook on empirical translation studies. In L. Vandevoorde, J. Daems, & B. Defrancq (Eds.), *New Empirical Perspectives on Translation and Interpreting* (pp. 333–370). Abingdon: Routledge.

Kotze, H. (2022). Translation as constrained communication: Principles, concepts and methods. In S. Granger, & M.-A. Lefer (Eds.), *Extending the Scope of Corpus-Based Translation Studies* (pp. 67–97). London: Bloomsbury.

Kruger, H., & Van Rooy, B. (2016). Constrained language: A multidimensional analysis of translated English and non-native indigenised varieties of English. *English World-Wide*, *37*(1), 26–57.

Kübler, N. (2008). A comparable Learner Translator Corpus: Creation and use. In *Proceedings of the Comparable Corpora Workshop of the LREC Conference* (pp. 73–78), Marrakech, 28-30 May 2008. http://www.lrec-conf.org/proceedings/lrec2008/workshops/W12_Proceedings.pdf

Kübler, N., Mestivier-Volanschi, A., & Pecman, M. (2018). Teaching specialised translation through corpus linguistics: quality assessment and methodology evaluation by experimental approach. *Meta*, *63*(3), 806–824.

Kübler, N., Mestivier, A., & Pecman, M. (2022). Using comparable corpora for translating and post-editing complex noun phrases in specialized texts: Insights from English-to-French specialized translation. In S. Granger, & M.-A. Lefer (Eds.), *Extending the Scope of Corpus-based Translation Studies* (pp. 237–266). London: Bloomsbury.

Kunilovskaya, M., Ilyushchenya, T., Morgoun, N., & Mitkov, R. (2022). Source language difficulties in learner translation: Evidence from an error-annotated corpus. *Target*. https://doi.org/10.1075/target.20189.kun

Kunilovskaya, M., & Lapshinova-Koltunski, E. (2019). Translationese features as indicators of quality in English-Russian human translation. In I. Temnikova, C. Orasan, G. Corpas Pastor, & R. Mitkov (Eds.), *Proceedings of the 2nd Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT 2019)*, 47–56.

Kunilovskaya, M., & Morgoun, N. (2016). Available corpora and error-annotated student translations in translator education. In *Proceedings of the 6th Conference The Future of Education* (pp. 121–125). Padova: Libreria Universitaria.

Kunilovskaya, M., Morgoun, N., & Pariy, A. (2018). Learner vs. professional translations into Russian: Lexical profiles. *Translation and Interpreting*, *10*(1), 33–52.

Kutuzov, A., & Kunilovskaya, M. (2014). Russian learner translator corpus: design, research potential and applications. In P. Sojka, A. Horak, I. Kopecek, & K. Palak (Eds.), *Text, Speech and Dialogue. Lecture Notes in Computer Science* (pp. 315–323). Berlin: Springer.

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

Lapshinova-Koltunski, E. (2022). Detecting normalisation and shining-through in novice and professional translations. In S. Granger, & M.-A. Lefer (Eds.), *Extending the Scope of Corpus-based Translation Studies* (pp. 182–206). London: Bloomsbury.

Lapshinova-Koltunski, E., Popović, M., & Koponen, M. (2022). DiHuTra: a Parallel Corpus to Analyse Differences between Human Translations. In *Proceedings of the 23rd Annual Conference of the European Association for Machine Translation* (pp. 335–336). European Association for Machine Translation.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, *43*(4), 557–570.

Leedham, M., & Cai, G. (2013). *Besides…on the other hand*: Using a corpus approach to explore the influence of teaching materials on Chinese students' use of linking adverbials. *Journal of Second Language Writing*, *22*, 374–389.

Lefer, M.-A. (2020). Parallel corpora. In M. Paquot & S. Th. Gries (Eds.), *A Practical Handbook of Corpus Linguistics* (pp. 257–282). Cham: Springer.

Lefer, M.-A., Piette, J., & Bodart, R. (2022). *Machine Translation Post-Editing Annotation System (MTPEAS) manual. Version 1.0*. Louvain-la-Neuve: OER UCLouvain. http://hdl.handle.net/20.500.12279/829

Lefer M.-A., & Vogeleer, S. (Eds.). (2013). *Interference and normalisation in genre-controlled multilingual corpora*. *Belgian Journal of Linguistics*, *27*.

Loock, R. (2020). It's non-canonical word order that you should use! A corpus approach to avoiding standardized word order in translated French. In S. Granger, & M.-A. Lefer (Eds.), *Translating and Comparing Languages: Corpus-based Insights. Corpora and Language in Use Proceedings 6* (pp. 69–85). Louvain-la-Neuve: Presses universitaires de Louvain.

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

Milton, J., & Tsang, E.S.C. (1993). A corpus-based study of logical connectors in EFL students' writing: directions for future research. In R. Pemberton, & E.S.C. Tsang (Eds.), *Studies in Lexis* (pp. 215–246). Hong Kong: The Hong Kong Institute of Science and Technology.

Nesselhauf, N. (2004). Learner corpora and their potential in language teaching. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 125–152). Amsterdam: John Benjamins.

Neumann, S., Kerz, E., & Heilmann, A. (forthcoming). Comparing contact effects in translation and second language learning. In H. Kotze, & B. Van Rooy (Eds.), *Constraints on Language Variation and Change in Complex Multilingual Contact Settings*. Amsterdam: John Benjamins.

Olohan, M. (2004). *Introducing Corpora in Translation Studies*. London & New York: Routledge.

Olohan, M., & Baker, M. (2000). Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, *1*(2), 141–158.

Paquot, M., & Plonsky, L. (2017). Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research*, *3*(1), 61–94.

Penha-Marion, L. A. de S., Gilquin, G., & Lefer, M.-A. (forthcoming). The effect of directionality on lexico-syntactic simplification in French><English student translation. In H. Kotze, & B. Van Rooy (Eds.), *Constraints on Language Variation and Change in Complex Multilingual Contact Settings*. Amsterdam: John Benjamins.

Redelinghuys, K., & Kruger, H. (2015). Using the features of translated language to investigate translation expertise: A corpus-based study. *International Journal of Corpus Linguistics*, *20*(3), 293–325.

Reynaert, R., Macken, L., Tezcan, A., & De Sutter, G. (2021). Building a new-generation corpus for empirical translation studies: the Dutch Parallel Corpus 2.0. In V. Wang, L. Lim, & D. Li (Eds.), *New perspectives on corpus translation studies* (pp. 75–100). Singapore: Springer.

Rundell, M., & Granger, S. (2007). From corpora to confidence. *English Teaching Professional*, *50*, 15–18.

Schneider, G., & Gilquin, G. (2016). Detecting innovations in a parsed corpus of learner English. *International Journal of Learner Corpus Research*, *2*(2), 177–204.

Seidlhofer, B. (2002). Pedagogy and local learner corpora. In S. Granger, J. Hung, & S. Petch-Tyson (Eds.), *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* (pp. 213–234). Amsterdam and Philadelphia: John Benjamins.

Sinclair, J. (1996). Preliminary Recommendations on Corpus Typology. Technical report. EAGLES (Expert Advisory Group on Language Engineering Standards). Available at www.ilc.cnr.it/EAGLES96/corpustyp/corpustyp.html

Swain, M. (1995). Three functions of output in second language learning. In G. Cook, & B. Seidlhofer (Eds.), *Principle and Practice in Applied Linguistics* (pp. 125–144). Oxford: Oxford University Press.

Tsagari, D., & Floros, G. (Eds.). (2013). *Translation in Language Teaching and Assessment.* Newcastle upon Tyne: Cambridge Scholars Publishing.

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

Vanderbauwhede, G. (2012). The Integrated Contrastive Model evaluated: The French and Dutch demonstrative determiner in L1 and L2. *International Journal of Applied Linguistics*, *22*(3), 392–413.

Van Vuuren, S., & Berns, J. (2018). Same difference? L1 influence in the use of initial adverbials in English novice writing. *IRAL*, *56*(4), 427–461.

Wurm, A. (2020). Translation quality in an error-annotated translation learner corpus. In S. Granger, & M.-A. Lefer (Eds.), *Translating and Comparing Languages: Corpus-based Insights*. *Corpora and Language in Use Proceedings 6* (pp. 141–162). Louvain-la-Neuve: Presses universitaires de Louvain.

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

Xiao, R. (2007). What can SLA learn from contrastive corpus linguistics? The case of passive constructions in Chinese learner English. *Indonesian Journal of English Language Teaching*, *3*(1), 1–19.

Zanettin, F. (1998). Bilingual comparable corpora and the training of translators. In S. Laviosa (Ed.), *Meta*, *43*(4). Special issue: *The Corpus-based Approach: A New Paradigm in Translation Studies*, 616–630.

Zanettin, F., Bernardini, S., & Stewart, D. (Eds.). (2003). *Corpora in Translator Education*. London: Routledge.