RiCL

Research in Corpus Linguistics

RiCL 11/2 (2023)

Special Issue

“Corpus-based translation studies (CBTS)"

edited by Sara Laviosa

aelinco
Asociación Española de Lingüística de Corpus

RiCL 11/2 (2023)

Editors

Paula Rodríguez-Puente and Carlos Prado-Alonso

ISSN 2243-4712

https://ricl.aelinco.es/

RiCL Research in Corpus Linguistics

Official journal of aelinco Asociación Española de Lingüística de Corpus

RiCL Research in Corpus Linguistics

11/2 (2023)

RiCL Research in Corpus Linguistics

Introduction: The ascent of corpus-based translation studies

Sara Laviosa
University of Bari Aldo Moro / Italy

Abstract – The pervasiveness of corpus-based research in the broad interdisciplinary field of translation studies is well attested. This editorial briefly reports on some of the most significant academic initiatives undertaken in corpus-based translation studies in recent years. It introduces each of the eight papers selected for this special issue of *Research in Corpus Linguistics* (RiCL). In doing so, the editorial will highlight their distinctive contribution to the interdisciplinarity of translation and interpreting studies.

Keywords – corpus-based translation studies; corpus-based interpreting studies; constrained communication; terminology development; audiovisual translation

Since its first appearance on the scene as a novel approach endorsed and adopted by scholars and scholar-teachers in the empirical study of the product, process and function of translation as well as translator training, corpus-based translation studies —as originally named by Shlesinger (1998)— has developed into a fully-fledged area of scholarly enquiry that engages with multiple disciplines, thus enhancing the interdisciplinarity of translation studies.

The pervasiveness of corpus studies of translation and interpreting in very recent years is amply attested by their substantial contribution to a variety of scholarly initiatives. To name but a few:

1) International symposia and conferences on translation studies, for instance:
   a. *International Symposium on Corpora and Translation Education*, 5–6 June 2021, Hong Kong Baptist University;
   b. 6th edition of the *Using Corpora in Contrastive and Translation Studies Congress* **(**UCCTS), 9–11 September 2021, University of Bologna;
   c. 10th edition of the *European Society for Translation Studies Congress*, 22–24 June 2022, Oslo Metropolitan University and University of Oslo.

2) Special issues in journals dedicated to:
    a. English studies, e.g., *Textus. English Studies in Italy* (Bernardini and Mair 2019);
    b. translation and translanguaging, e.g., *Translation and Translanguaging in Multilingual Contexts* (Dullion 2017; Flores Acuña and Rodríguez Reina 2019);
    c. translation and interpreting, e.g., *MonTI. Monografías de Traducción e Interpretación* (Calzada Pérez and Laviosa 2021);
    d. translation, e.g., *Translation Quarterly* (Laviosa and Liu 2021).

3) Interdisciplinary collected volumes, e.g., *The Routledge Handbook of Translation and Education* (Laviosa and González-Davies 2020), or *The Oxford Handbook of Translation and Social Practices* (Ji and Laviosa 2021), among others.

This special issue testifies to the growing interdisciplinary interest in corpus-based translation studies worldwide. The eight articles included in the issue represent state-of-the-art research that has recently been undertaken by international scholars within the field of corpus-based translation studies and its offshoot, corpus-based interpreting studies. It is worth remembering that the latter was originally advocated and outlined by Shlesinger (1998: 490–491, original emphasis), who, in those early days, set the following goals for the fledgling field of descriptive corpus-based interpreting studies:

> recourse to interpreting as part of corpus-based translation studies may indeed help to focus attention on what sets interlingual mediation apart, *regardless of modality*. By the same token, however, while continuing to explore the common ground, the corpus-based study of interpreting will also help to define what sets it apart. Both aims are very much in keeping with the agenda of its parent discipline, translation studies.

Twenty-five years on, Shlesinger's research agenda is still being followed and expanded, as demonstrated in that three contributions to this special issue deal with simultaneous, consecutive, and dialogue interpreting respectively. The first of these, **Claudio Bendazzoli**'s article, is based on the *Directionality in Simultaneous Interpreting Corpus* (DIRSI; Bendazzoli 2010, 2012), which consists of transcripts and audio-recordings of the source texts (English and Italian) and target texts (English and Italian) collected from three international medical conferences held in Italy. The study investigates the trend of

text compression/expansion for each source speech event and its interpretation. The findings confirm the general trend that interpreted speeches tend to contain a lower number of words than their original speeches, regardless of directionality. However, target texts produced from extremely short source text (under 500 words, typically opening/closing remarks, floor allocation, and announcements) usually contain more words than their corresponding source texts.

Bing Zou and **Binhua Wang**'s study is based on an annotated, self-compiled, and aligned bilingual parallel corpus of *Chinese-English Interpreting for Premier Press Conferences* (CEIPPC). The study investigates non-fluency, namely the different types of pauses including filled/silent pauses, juncture/non-juncture pauses and self-repairs including repetitions, self-corrections, and reformulations. The findings show that most of the interpreters' non-fluencies are significantly related to syntactical structures in the speakers' discourse. Hence, as the authors contend, language-pair specificity should be considered an important variable or parameter for evaluating and assessing interpreters' on-site performance.

Bart Defrancq and **Sofie Verliefde**'s analyze the *Interpreter-mediated Police Interviewing with Drafting Corpus* (IMPID; Verliefde and Defrancq 2022), which consists of 12 interpreter-mediated police interviews conducted in Belgium from 2014 to 2019. Their study investigates the frequency of use and the semantic and interactional functions of *dus*, which is the most common Dutch marker of consequence. The findings show that, compared to simultaneous interpreting, dialogue interpreting seems to incentivize interpreters more to add the connective *dus* to their interpretations. Nearly 90 percent of the occurrences have no equivalent in the corresponding source utterances. With regard to functions, turn management (turn taking and turn yielding) prevails (40% of the cases). Rephrasing and filler functions jointly account for a third of the occurrences, and consequential and inferential *dus* amounts to almost a fifth of the examples. The authors discuss a number of cases of untriggered uses, placing them in a wider context of interpreter strategies. They conclude that, while explicitation seems to be at play, the bulk of occurrences fulfils interaction coordination purposes.

Moving on from corpus-based interpreting studies to corpus-based translation studies, **Shuangzi Pang** and **Kefei Wang**'s contribution offers an insightful overview of the evolution of this expanding interdisciplinary area of research over the last two decades, assesses the state of the art, and points to future directions. Interestingly, the

authors identify three major trends that have emerged in the field. Firstly, translations are being viewed as contact varieties that are influenced by a range of constraining factors. Secondly, the field has diversified its research perspectives going beyond linguistics, thus reconciling translation and cultural studies. Finally, there is a growing interest in creating multilingual and diachronic composite corpora and conducting multivariate statistical analyses.

Ho Ling Kwok, **Sara Laviosa** and **Kanglong Liu**'s paper, "Lexical simplification in learner translation: A corpus-based approach," is in line with the emerging trend of viewing translated texts, including trainees' translations, as forms of constrained communication. Their study is based on two comparable corpora: the *International Corpus of English in Hong Kong* (ICE-HK; Nelson 2006) and the *Parallel Learner Translation Corpus* (PLTC)[1] compiled at The Hong Kong Polytechnic University. The aim of the study is to test the lexical simplification hypothesis in translations by students. The findings show that Chinese-to-English translations are not lexically simpler than writing in English as a Second Language (ESL), as indicated by the four parameters of: 1) lexical density (which indicates informativeness), 2) standardized type-token ratio, 3) core vocabulary coverage, and 4) list head coverage (which indicates lexical diversity). Moreover, student translations are found to be lexically denser than ESL writing. The authors discuss the motivations for these results from the perspective of constrained communication, the language background of writers and translators, source language influence, and comparable corpus construction.

Still within the field of descriptive corpus studies of translation, **Yu Zhai** and **Bin Xu**'s contribution is based on a small aligned parallel corpus of six Chinese translations of children's literary books winners of the prestigious *Newbury Medal Award*.[2] The study investigates the phenomenon of *embellishment*, a stylistic feature that characterizes Chinese translations of contemporary English children's literature. Embellishment, which is viewed by the authors as a form of lexical over-explicitation, is found in five out of the six translations included in the corpus. The authors conclude that the phenomenon can be explained in terms of translators' choices and editors' preferences.

Within the field of applied corpus studies of translation, **Koliswa Moropa** and **Bulelwa Nokele**'s provide an overview of the state of the art of specialized multilingual

[1] https://cerg1.ugc.edu.hk.
[2] https://www.britannica.com/art/Newbery-Medal

parallel corpora constructed as resources for terminology development and terminological aids for researchers, as well as prospective and practicing translators working from English into several official indigenous African languages, namely IsiZulu, IsiXhosa, IsiNdebele, SiSwati, Tshivenda, and Xitsonga. Focusing on current research that is being carried out at the University of South Africa (Unisa), the authors demonstrate how, the skillful use of the different functions provided by the software tool *ParaConc* (Barlow 2003) in the *University of South Africa Multilingual Parallel Corpus*[3] (UNISA) —which is constantly being expanded to include more and more subject-specific domains— is a highly valuable resource for enriching the target languages. UNISA is contributing significantly to implementing the national language policy adopted since 1994, which is aimed at developing and intellectualizing the African indigenous languages for teaching, learning, and research.

Finally, within the burgeoning interdisciplinary research area of translation in language learning and teaching, **Ruska Ivanovoska-Naskova**'s examines various pedagogic uses of translation tasks in undergraduate degree programmes in modern languages. Focusing on the use of interlingual subtitling from Italian as language B to Macedonian as language A, the author reports on an observational study carried out by herself as a participant observer in her class, which is intended for Macedonian-speaking students learning Italian at advanced level. Students used a variety of language resources and translation tools including paper and electronic dictionaries, as well as terminological databases. They also created small-size bilingual comparable corpora with texts retrieved from online sources, which they then searched with the software *AntConc* (Anthony 2023). Students also conducted free research on the web. The software *Subtitle Workshop*[4] was used for the subtitling. In line with the data-driven learning approach, the translation was undertaken collaboratively and students were free to start and conduct their research as they thought was best for a given term. Students' feedback at the end of the teaching unit was overall positive. Most of the class shared the view that they would like to have more activities of this kind in their studies.

[3] https://repo.sadilar.org/handle/20.500.12185/489?show=full
[4] http://subworkshop.sourceforge.net

References

Anthony, Laurence. 2023. *AntConc* (Version 4.2.2). Tokyo: Waseda University. https://www.laurenceanthony.net

Barlow, Michael. 2003. *ParaConc: Concordance Software for Multilingual Parallel Corpora*. Houston: Rice University. http://www.mt-archive.info/LREC-2002-Barlow.pdf

Bendazzoli, Claudio. 2010. *Corpora e Interpretazione Simultanea*. Bologna: Asterisco.

Bendazzoli, Claudio. 2012. From international conferences to machine-readable corpora and back: An ethnographic approach to simultaneous interpreter-mediated communicative events. In Francesco Straniero Sergio and Caterina Falbo eds. *Breaking Ground in Corpus-based Interpreting Studies*. Frankfurt: Peter Lang, 91–117.

Bernardini, Silvia and Christian Mair eds. 2019. *Investigating Englishes with Corpora: Variation, Contact, Translation.* Special issue of *Textus. English Studies in Italy*. Rome: Carocci editore.

Calzada Pérez, María and Sara Laviosa eds. 2021. *Reflexión crítica en los estudios de traducción basados en corpus / CTS Spring-cleaning: A Critical Reflection*. Special Issue of *Monografías de Traducción e Interpretación* 13. Alacant: Universitàt de Alacant.

Dullion, Valérie ed. 2017. *Between Specialised Texts and Institutional Contexts – Competence and Choice in Legal Translation*. Special Issue of *Translation and Translanguaging in Multilingual Contexts* 3/1. Amsterdam: John Benjamins.

Flores Acuña, Estefanía and Pilar Rodríguez Reina eds. 2019. *Orality, Language and Interpreting Challenges*. Special issue of *Translation and Translanguaging in Multilingual Contexts* 5/3. Amsterdam: John Benjamins.

Ji, Meng and Sara Laviosa eds. 2021. *The Oxford Handbook of Translation and Social Practices*. Oxford: Oxford University Press.

Laviosa, Sara and María González-Davies eds. 2020. *The Routledge Handbook of Translation and Education*. London: Routledge.

Laviosa, Sara and Kanglong Liu eds. 2021. *The Pervasiness of Corpora in Translation Studies*. Special issue of *Translation Quarterly* 101. Hong Kong: Hong Kong Translation Society.

Nelson, Gerald. 2006. *The ICE Hong Kong Corpus: User Manual*. London: University College London.

Shlesinger, Miriam. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta* 43/4: 486–493.

Verliefde, Sofie and Bart Defrancq. 2022. Interpreter-mediated access to the written record in police interviews. *Perspectives* 31: 519–547.

Corresponding author
Sara Laviosa
University of Bari Aldo Moro
Via Garruba 6
70122 Bari
Italy
E-mail: sara.laviosa@uniba.it

RiCL Research in Corpus Linguistics

The impact of directionality and speech event type on target speech compression/expansion in simultaneous interpreting

Claudio Bendazzoli
University of Turin / Italy and University of Las Palmas de Gran Canaria / Spain

Abstract – Simultaneous interpreting is a complex cognitive activity that can be influenced by several factors, including source speech features (e.g., delivery rate), contextual variables, working languages, and directionality (e.g., interpreting from/into one's native or foreign language), among others. Owing to the time constraints inherent in this interpreting mode, simultaneous interpreters must make swift decisions on how to best deliver the original message into the target language. Although explicitation is considered a universal feature of translation and interpreting, it is also true that part of (redundant) information is eventually omitted. In fact, as opposed to translated texts, interpreting corpora show a general trend of interpreted speeches being shorter than source speeches (in terms of number of words). However, a closer look at the *Directionality in Simultaneous Interpreting Corpus* (DIRSI) partially disconfirms such a general trend. The DIRSI corpus consists of three medical conferences mediated by simultaneous interpreters (English/Italian). Each conference is analyzed in terms of speech length to ascertain to what extent directionality and speech event type may have an impact on the interpreters' output. Results show that directionality cannot always be linked to target speech expansion, whereas the type of speech event is likely to play a role. In particular, this applies to the interpretation of source speeches under 500 words, as interpreters adopt optimization strategies to manage politeness, source speech ungrammaticality, and integrate contextual cues.

Keywords –simultaneous interpreting; speech event; directionality; compression; expansion; DIRSI corpus

1. INTRODUCTION

The advent of machine-readable corpora of translated texts has given way to the study of the distinguishing features of "translated literature […] as a system in its own right" (Baker 1993: 238) and, more generally, it opened up the idea of translation universals and norm-oriented features. In this respect, particular attention has been put on the level of explicitness of translated texts (i.e., target texts, henceforth TT) compared to original texts (i.e., source texts, henceforth ST), along with simplification, disambiguation, conventional grammaticality, and repetition avoidance in translation and, to a lesser

DOI 10.32714/ricl.11.02.02

extent, in interpreting (Baker 1993: 243–245). These features are largely born out of the constraints inherent in the translation and interpreting process —for instance, space limitations in subtitling or time in simultaneous interpreting (henceforth SI)— and may also be directly linked to the strategic dimension therein (Riccardi 2005). Eventually, many of those features can be subsumed into text compression or text expansion as linguistic items are omitted or added respectively.

Considering the fundamental differences in the constraints and the strategies that can be found in either translation or interpreting, the patterns unveiled in translated texts may not match with the ones that can be observed in interpreted texts. For instance, while translations would seem to be longer than their ST (Frankenberg-Garcia 2009; Abbasi and Koosha 2016), corpus data show that the opposite is generally the case in SI, with TT being shorter than their original speeches. TT compression was observed in SI of European Parliament debates in the *European Parliament Interpreting Corpus* (EPIC; Russo *et al.* 2012; Russo 2018) and the *EP-Poland Corpus* (Bartłomiejczyk 2022; Bartłomiejczyk *et al.* 2022), where interpreters normally work into their native language and, also, in SI of medical conferences in the *Directionality in Simultaneous Interpreting* corpus (DIRSI; Bendazzoli 2010, 2012), where interpreters work bidirectionally, that is, from their foreign working language into their native language and vice versa (also known as interpreting into B or *retour*). However, this is merely a general trend, which results from cumulative data, but does not single out the various speech events making up the communicative situations they originate from.

The present study is based on the DIRSI corpus and aims to ascertain whether the type of source speech events and the directionality of interpreting —that is, whether interpreters work into their native or foreign working language— may still confirm the general tendency to text compression in SI.

The study is organized as follows. Section 2 gives a brief overview of the constraints involved in SI and the observations made with respect to interpreters' strategic behavior, especially with respect to TT compression and expansion. It also examines the range of speech events constituting the conference as a communicative situation. Section 3 provides a description of the DIRSI corpus, which consists of transcribed ST and TT in English and Italian from three international conferences held in Italy. The results are presented in section 4 and discussed in section 5, with a particular focus on the instances of text expansion which were detected in TT originating from very short source speech

events. This result goes against the general tendency of TT compression in interpreting. Finally, Section 6 concludes the study.

2. SIMULTANEOUS INTERPRETING

2.1. Constraints and strategies

SI is a translational activity in which the ST and the TT are produced at the same time. In fact, the interpreter's output is not 100 percent simultaneous, as a minimal unit of meaning from the ST is necessary to start processing the input and get to a meaningful output. Such a time mismatch between ST and TT is known as 'décalage' or 'Ear-Voice-Span' and it can vary depending on a range of factors, such as the individual interpreter's working memory capacity, ST speed, lexical density, delivery (impromptu speech, semi/prepared, read out from a script, etc.), and culture-bound units of meaning (Riccardi 2005). Although this applies to all types of SI —that is, with or without sound-proof booth, and with or without equipment such as headsets and microphone— the data analyzed in the present study refer to SI with a booth and an equipment.

Interpreters' working languages are classified by the *International Association of Conference Interpreters*[1] as language A for one's native language, language B for one's active foreign language (meaning that an interpreter is able to translate both from and into that language) and language C for one's passive foreign language (meaning that an interpreter is able to translate from that language but not into it). Interpreters working at international institutions generally interpret into their native language (with the exception of those language combinations for which there are fewer interpreters available), whereas interpreters working as freelancers —e.g., in Italy's private market— are generally required to cover both directions of a language combination. Depending on the interpreters' directionality —that is, on whether they are translating into their language A or language B— different strategies may be put in place and the scope of language availability may be limited more or less (Gile 2009; Aston 2018; Cresswell 2018).

As is clear from the considerations above, time stands as one of the major constraints in SI. Drawing on a classic definition of interpreting, especially in SI from a booth, "a first and final rendition in another language is produced on the basis of a one-

[1] https://www.aiic.org/

time presentation of an utterance in a source language" (Pöchhacker 2004: 11). Interpreters have virtually no chance to ask for repetitions or clarifications, as they are physically separated from the source speaker and, even if they could do so by making gestures from the booth or voicing a request explicitly, it would nonetheless seriously disrupt the communication flow and make interpreters lose face.

In order to keep up with all the constraints affecting SI, interpreters tend to develop relevant strategies which may be language specific as different language systems pose particular challenges. According to Riccardi (2005), interpreting strategies can be grouped into several categories, namely comprehension, production-oriented, general, and emergency strategies. Among these, compression and expansion are categorized as production-oriented strategies, while omission, paraphrasing and reordering (which may imply text compression or expansion) are listed under the emergency strategy category.

Regardless of the specific nature of each strategy, compression and expansion have been the object of investigation since the early studies in SI research. In the seminal work by Chernov (2004) on text redundancy and anticipation in SI, reference is made to syllabic, lexical, syntactical, semantic, and situational compression. In fact, in some cases, TT compression may be obligatory owing to fundamental differences between two language systems. The same applies to expansion, as some categories may be missing in the target language and more explicit phrasing may be necessary to produce a fully acceptable target output.

Obligatory explicitation results from the differences between two or more language systems whereby it may be necessary to provide more information in the target language than is explicitly available in the source language. A common example of this, when interpreting from a pro-drop language like Italian into a non-pro-drop language such as English, is the use of personal subject pronouns: these can be omitted in Italian but must be mentioned in English. On the other hand, as Frankenberg-Garcia (2009: 49) states, voluntary explicitation:

> can be a result of conscious decision to make the target text easier to understand or even of a subconscious operation inherent to the process of translation.

Gumul (2017) provides a broad overview of explicitation in SI by listing an important number of surface manifestations, such as adding connectives, modifiers, qualifiers, intensifying cohesive ties, inserting hedges, disambiguating lexical metaphors, etc. Her

analysis looks at trainee interpreters' performance and identifies the following factors as having a bearing on explicitation: interpreting strategies (process-oriented and product-oriented), interpreting constraints (time, linearity, unshared knowledge, and memory load), directionality (native vs. *retour*), and idiosyncratic preferences (Gumul 2017: 284).

Various instances of TT compression and expansion have also been observed in previous studies looking at professional interpreters' output, typically using corpus data from European Parliament debates (EP). For example, from a comparable perspective, a higher frequency of the complementizer *that* was observed in English TT with respect to ST delivered in English (Kajzer-Wietrzny 2018). Conversely, from a parallel perspective (Morselli 2018), linking adverbials appeared to be left out more in English TT (from Italian ST), whereas apposition markers were added more in Italian TT (from English ST). Similar results from EP debates concern discourse markers, which were found to be both deleted and added more by interpreters than translators (Defrancq *et al.* 2015).

Bendazzoli (2019), focusing on the use of the discourse marker *so* by simultaneous interpreters in the DIRSI corpus, revealed that 30 percent of the occurrences were actually generated by the interpreters themselves, sometimes upon evident expansion of the TT with the addition of new information, reiteration of previously given information, and restructuring of the interpreter's output.

It is clear that TT compression and expansion are two sides of the same coin. While redundancy and repetitions in ST can give interpreters the opportunity to take advantage of time-saving strategies as certain items are reduced or omitted, TT expansion or additions can be effective time-gaining strategies whenever interpreters need to receive more units of meaning and figure out how to proceed with their output.

2.2. Conference setting and speech events

In addition to the factors mentioned above, the type of communicative situation where SI is provided also has a strong bearing on the potential constraints and interpreters' strategic behavior, as the rules of procedures applicable to a certain situation may differ considerably from others. For example, speaking time and floor allocation in EP debates may differ considerably from the speech events typically found at scientific or academic conferences (Bendazzoli 2010).

The constituent parts of a conference, considered as a communicative event, were identified in previous studies (Pöchhacker 1994; Riccardi 1995; Russo 1999; Shalom 2002; Ventola 2002), which outlined the structure of a conference into sessions with various functions (e.g., opening session, paper presentation session, poster session, plenary/keynote lecture, panel/roundtable, etc.). Based on these classifications, and thanks to the field observations made during the data collection stage of the DIRSI corpus, it was possible to define the kinds of sections making up a conference, along with its participants' (communicative) roles and main speech events (Bendazzoli 2012). Speech events are particularly relevant to the present study and range from opening remarks to paper presentations, lecture or plenary presentations, floor allocations, procedure, housekeeping announcements, questions, answers, comments, and closing remarks.

It is important to highlight that the presence of simultaneous interpreters requires conference participants to speak with a microphone and one at a time, so as to allow interpreting service users to understand who is actually speaking, and the interpreters to provide their service. When such a procedure is disrupted, interpreters may feel the need to shift to a different speaking person or to verbalize the situation (see Bendazzoli 2023 for an example in the EP).

Conference speech events are also characterized by their total length (in terms of number of words), time duration, and delivery rate. Overall, drawing on the field observation of the interpreter-mediated conferences making up the DIRSI corpus (see specifications in Section 3) and 11 further conferences that were not included in the corpus (yet they are part of the DIRSI multimedia archive), it was possible to determine that few major speech events are embedded in a much larger sequence of shorter speech events. The distribution and relative length/duration of conference speech events are substantially different from those observable in EP debates (Bendazzoli 2012). Based on DIRSI corpus data, the typical ranges of speech event duration (time) and length (number of words) in a conference are as follows: 1) short (up to 15 minutes and less than 1,650 words), 2) medium (between 15 and 30 minutes, between 1,650 and 3,300 words), and 3) long (more than 30 minutes and more than 3,300 words).

3. CORPUS DATA AND METHODOLOGY

The DIRSI corpus includes transcripts and audio recordings from three international medical conferences held in Italy and mediated by professional simultaneous interpreters in English/Italian. Two conferences were about cystic fibrosis and were organized by the *Cystic Fibrosis Foundation*[2] in Verona (CFF4 and CFF5) and one conference was organized by a partnership of associations from different countries in ELSA,[3] a European project, and was about the role of foreign carers in assisting elderly people. The following are the official titles of each conference:

1) *IV Spring Seminar. Recent Advances and Future Developments in Cystic Fibrosis Research: Diabetes, Nutrition, and Internet Communication*, held in Verona on 25 May 2006 (CFF4).

2) *V Spring Seminar. Recent Advances and Future Developments in Cystic Fibrosis Research: What Changes in CF, Pharmacotherapy of the Basic Defect, Advances in CF Lung Transplantation*, held in Verona on 11 May 2007 (CFF5).

3) *Participation and Partnership in Local Policies to Support Non-self-sufficient Elderly People and their Family Members*, held in Cesena on 19 October 2006.

The three conferences were open to both experts and non-experts: physicians and patients in CFF4 and CFF5, and project partners and community members in ELSA.

The corpus consists of four sub-corpora: two sub-corpora with all the original speeches —namely one sub-corpus of Italian ST and one sub-corpus of English ST— plus two sub-corpora with interpreted speeches, namely one sub-corpus of TT into English and one sub-corpus of TT into Italian. In total, five professional interpreters are represented in the corpus (one interpreter, IT-01, worked at two conferences). Four interpreters were native speakers of Italian (IT-01, IT-02, IT-03, IT-04) and also had English as an active working language, while one of them was a native speaker of English (UK-01) and had Italian as an active working language.

The speech events in the corpus belong to the opening, presentation, and closing sessions of the conferences. Debates and Q&A sessions were not considered, as their interactional pattern was considerably different from the other sessions, including cases

[2] https://www.cff.org/
[3] In Italian, the ELSA acronym stands for *Politiche di empowerment delle lavoratrici straniere addette alla cura* ('Policies for the empowerment of foreign carers').

of overlapping speech that could not fit in the design of the corpus. In total, the DIRSI corpus contains 10 hours of ST and 10 hours of TT approximately.

Each transcript in the corpus comes with a header including metadata about the conference, the participant, and the speech event. The total number of words in each sub-corpus was calculated by extracting the data from the header of each individual transcript. The number of words in the transcripts was obtained with the relevant function in the word processing program *TextPad*.[4] In addition to the automatic extraction of the data under consideration, it was also possible to use the same data included in the Excel document set up to manage the DIRSI multimedia archive. With the use automatic filters, it was possible to query the textual output of individual subjects based on directionality and speech event type. This is presented in graphic form in Section 4.

Using the word count as a unit of measurement for text compression/expansion does not come without problems. As discussed above, there are language-specific features that can determine the use of certain words compulsorily in one language and not in another one. In fact, alternative systems have been proposed —e.g., counting characters, syllables, or morphemes— but they all seem to be affected by similar limitations in determining the explicitness of a TT (Frankenberg-Garcia 2009). Another issue concerns the way words are counted depending on the word processor in use. *TextPad* counts those instances consisting of a word with an apostrophe and the word that follows them as one unit (they were not separated with a space in our transcripts). For this reason, the number of words does not coincide with the number of tokens in the corpus, when transcripts are tokenized. Such critical limitations can nonetheless be counterbalanced thanks to the bidirectional nature of the DIRSI corpus, as its structure allows for both parallel and comparable analyses of Italian and English as source and target languages.

4. RESULTS

The total number of words in DIRSI is 135,835. In more detail, the size of the four sub-corpora is quite balanced: they range from a minimum of 31,500 words (for Italian ST) to a maximum of 37,200 words (for English ST). Both sub-corpora containing English and Italian TT are smaller in size than the sub-corpora containing the respective ST. Table 1 illustrates the size of each sub-corpus: Italian source speeches (ORG-IT), interpretations

[4] https://www.textpad.com/home

from Italian into English (INT-IT-EN), English source speeches (ORG-EN), and interpretations from English into Italian (INT-EN-IT). The second column lists the number of speech events (e.g., opening remarks, paper presentations or lectures, etc.). This is followed by the number of words and the percentage of the corpus covered by each sub-corpus.

Sub-corpus	Number of speech events	Number of words	Percentage of DIRSI
ORG-IT	63	33,412	24.6
INT-IT-EN	63	31,510	23.2
ORG-EN	16	37,249	27.4
INT-EN-IT	16	33,664	24.8
Total	**158**	**135,835**	**100**

Table 1: Total size (number of words) of DIRSI

Looking at the distribution of total words in the four sub-corpora, it is interesting to note that the largest sub-corpus (ORG-EN) only contains a quarter as many texts (speech events) as the other ST sub-corpus (ORG-IT). This apparent disparity can be explained by the fact that the three conferences were held in Italy, they were all organized by Italian subjects, and the role of chair and discussant was always given to Italian speakers.

As discussed in Section 2.2, the course of proceedings at a conference often involves the production of a large number of short speech events (e.g., opening remarks, floor allocation, procedure, and formalities) and a considerably smaller number of longer speech events, such as main lectures. The length of the speaking time is closely related to the number of words produced. Nevertheless, the total speaking time is mostly covered by the few longer speech events. A similar trend is reflected in the length of texts in terms of the number of words produced.

The spoken output in each of the three conferences included in DIRSI is summarized in the set of tables below: Table 2 for Italian ST, Table 3 for English TT, Table 4 for English ST, and Table 5 for Italian TT:

Sub-corpus		Number of speech events	Number of words	Percentage of DIRSI
ORG-IT	CFF4	19	8,707	6.4
	ELSA	18	9,822	7.2
	CFF5	26	14,883	11.0
Sub-total		**63**	**33,412**	**24.6**

Table 2: Size of the ORG-IT sub-corpora in DIRSI (number of words)

Sub-corpus		Number of speech events	Number of words	Percentage of DIRSI
	CFF4	19	9,474	7
INT-IT-EN	ELSA	18	9,228	6.8
	CFF5	26	12,808	9.4
Sub-total		**63**	**31,510**	**23.2**

Table 3: Size of INT-IT-EN sub-corpora in DIRSI (number of words)

Sub-corpus		Number of speech events	Number of words	Percentage of DIRSI
	CFF4	5	15,189	11.2
ORG-EN	ELSA	6	7,836	5.8
	CFF5	5	14,224	10.5
Sub-total		**16**	**37,249**	**27.4**

Table 4: Size of ORG-EN sub-corpora in DIRSI (number of words)

Sub-corpus		Number of speech events	Number of words	Percentage of DIRSI
	CFF4	5	13,500	9.9
INT-EN-IT	ELSA	6	7,628	5.6
	CFF5	5	12,536	9.2
Sub-total		**16**	**33,664**	**24.8**

Table 5: Size of INT-EN-IT sub-corpora in DIRSI (number of words)

Comparing the size of the various sub-corpora of each conference, it is clear that the CFF5 conference makes the largest contribution of Italian ST and English TT, while the amount of words for English ST (and related Italian TT) is similar in CFF4 and CFF5. In contrast, the ELSA conference contributes a good number of words for Italian ST, but has a considerably smaller size for English ST.

In the set of tables below, data on the size of the various sub-corpora are grouped according to the conference to which they refer: Table 6 for CFF4, Table 7 for ELSA, and Table 8 for CFF5.

Sub-corpus	Number of speech events	Number of words	Percentage of DIRSI
ORG-IT	19	8,707	6.4
INT-IT-EN	19	9,474	7.0
ORG-EN	5	15,189	11.2
INT-EN-IT	5	13,500	9.9
Total	**48**	**46,870**	**34.5**

Table 6: Size of CFF4 sub-corpora (number of words)

Sub-corpus	Number of speech events	Number of words	Percentage of DIRSI
ORG-IT	18	9,822	7.2
INT-IT-EN	18	9,228	6.8
ORG-EN	6	7,836	5.8
INT-EN-IT	6	7,628	5.6
Total	**48**	**34.514**	**25.4**

Table 7: Size of ELSA sub-corpora (number of words)

Sub-corpus	Number of speech events	Number of words	Percentage of DIRSI
ORG-IT	26	14,883	11
INT-IT-EN	26	12,808	9.4
ORG-EN	5	14,224	10.5
INT-EN-IT	5	12,536	9.2
Total	**62**	**54,451**	**40.1**

Table 8: Size of CFF5 sub-corpora (number of words)

In this additional comparative analysis of the data on the size of the various sub-corpora of DIRSI, it is interesting to note that there is always a lower number of words in the TT than in the related ST, thus confirming the general tendency discussed above, with the exception of the Italian ST and the English TT in the CFF4 conference. Between these two sub-corpora (CFF4_ORG-IT and CFF4_INT-IT-EN) there is an increase of more than 700 words (roughly +9%) in the English TT. This increase may not be coincidental considering that, in CFF4, one of the two interpreters is a native English speaker (UK-01), so his working directionality in the sub-corpus in question is from language B to language A (English). This finding was checked against the word distribution between the two interpreters, and the higher output of interpreter UK-01 is confirmed when compared to the Italian native interpreter (IT-01) in English TT. Figure 1 shows the number of words in each ST in Italian (grey column) and their TT in English (black column) broken down by interpreter: IT-01 is the interpreter with English as language B and UK-01 is the interpreter with English as language A.

A look at Figure 1 shows that, although the total output of both interpreters does not deviate much from the number of words in the ST, UK-01 has a larger output in all cases regardless of the length of the ST (with only two exceptions out of a total of 14 speech events).[5] The same trend is not attested in the opposite directionality with TT in Italian, for which both interpreters typically produce shorter output than their respective

[5] The difference between the total words in the Italian ST and the English TT by interpreter is -61 words for IT-01 and +875 words for UK-01.

ST. This is in line with the general trend of TT compression attested in all the other sub-corpora and the studies mentioned earlier about SI of EP debates.

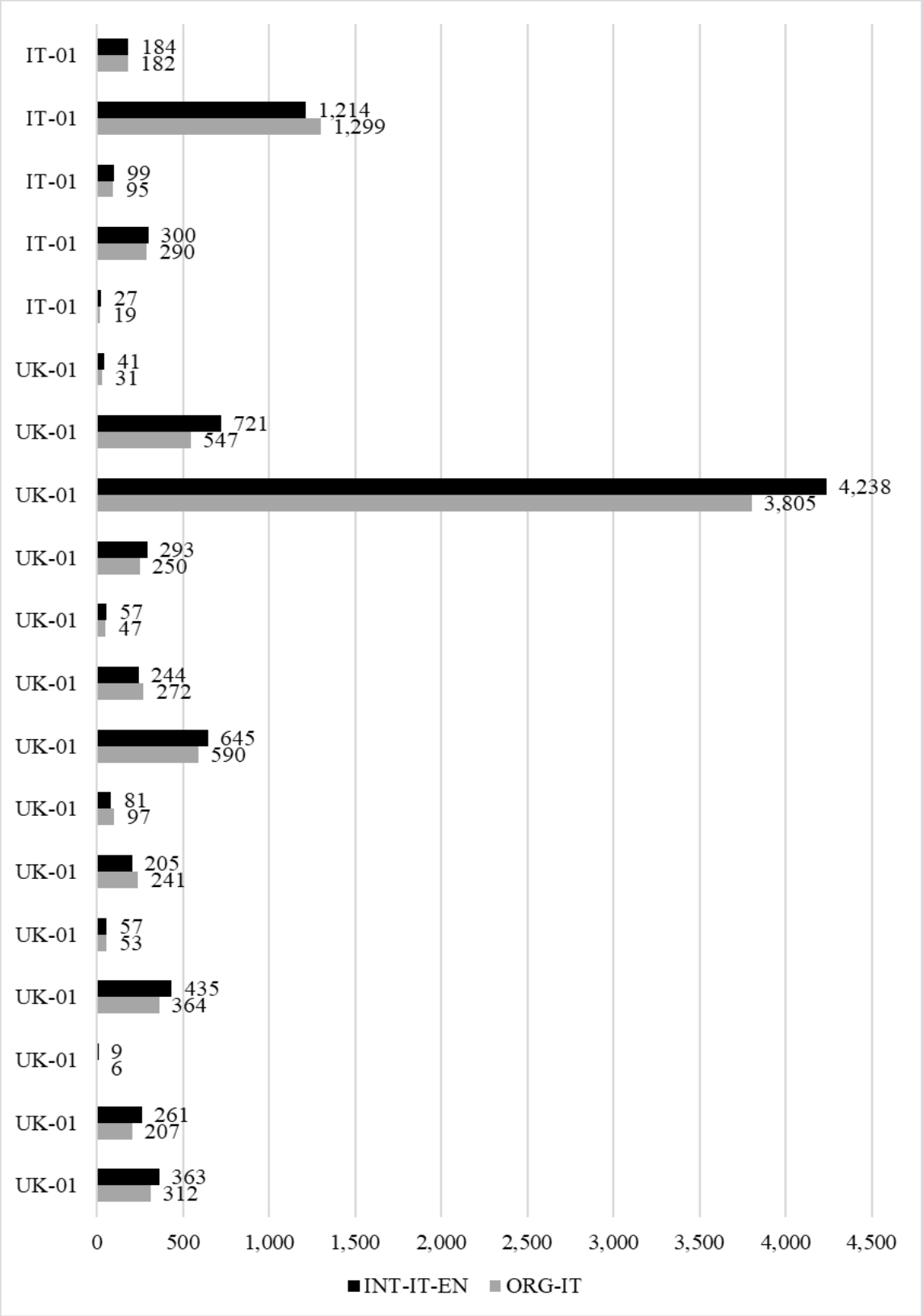


Figure 1: Comparison of the number of words in Italian STs and English TTs in CFF4 by interpreter

Given the variety of speech events making up each conference, the next step is to provide a detailed study of the trend of text compression/expansion for each source speech event and its interpretation within all the DIRSI sub-corpora, so as to check whether the general trend remains constant in all cases. In order to facilitate this kind of analysis and to effectively manage the amount of data at hand, the data from DIRSI-ORG-EN and DIRSI-INT-EN-IT (i.e., from the ST in English with the related TT in Italian), present in all the three conferences, are grouped into a single graph. By contrast, the ST in Italian and the related TT in English are presented separately for each conference and according to two total levels of output. A first group includes texts with less than 500 words, while a second group includes texts containing more than 500 words.

As Figure 2 clearly shows, contrary to the trend identified from a global observation of the data, in texts of shorter length (up to about 2,000 words) the number of words is slightly higher in TT than in ST. In contrast, in texts classified as medium (1,650–3,300 words) and long (> 3,300 words) in length, the general trend in which TT have fewer words than the corresponding ST is confirmed, as shown in Table 9.

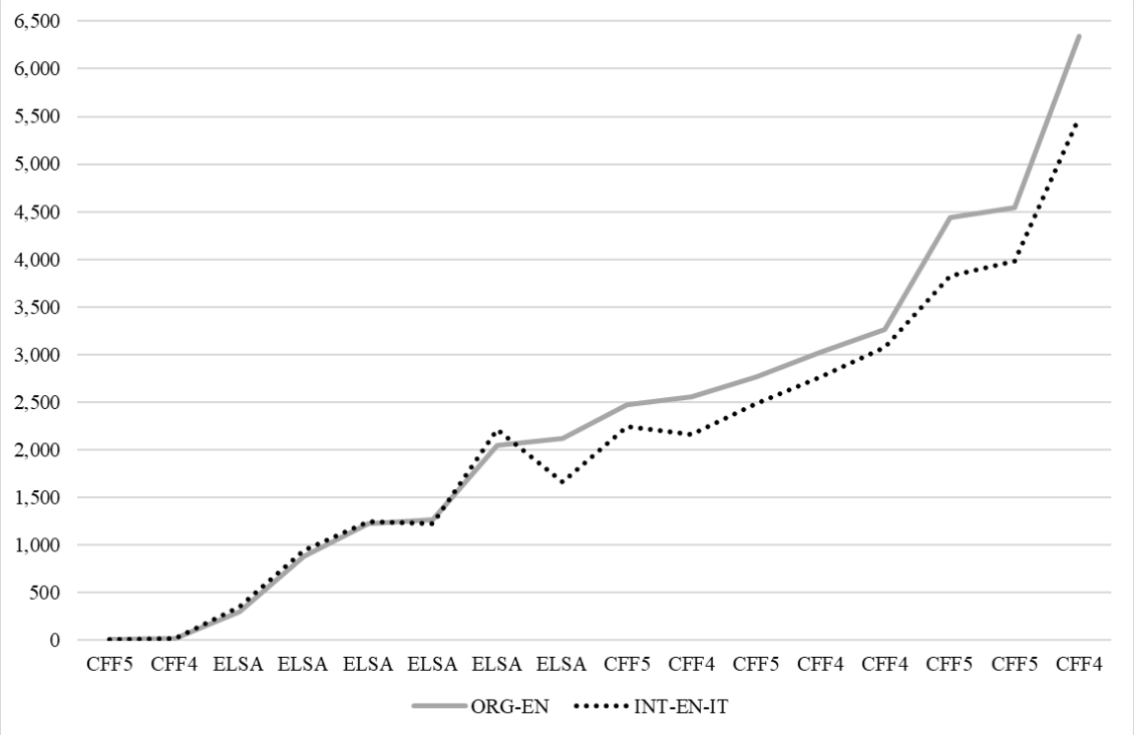


Figure 2: Number of words in English ST and Italian TT in DIRSI

Sub-corpus	Number of words in ORG-EN	Number of words in INT-EN-IT
CFF5	6	4
CFF4	18	14
ELSA	297	347
ELSA	879	938
ELSA	1,228	1,243
ELSA	1,269	1,225
ELSA	2,045	2,216
ELSA	2,118	1,659
CFF5	2,472	2,241
CFF4	2,550	2,160
CFF5	2,763	2,478
CFF4	3,019	2,763
CFF4	3,264	3,077
CFF5	4,439	3,827
CFF5	4,544	3,986
CFF4	6,338	5,486

Table 9: Number of words in English ST and Italian TT in DIRSI

As explained earlier, the analysis of the Italian ST and their corresponding TT in English was carried out with two subsets of data, that is, separating all the source speech events with less than 500 words from those with a higher number of words in each conference. Figures 3 and 4 display the trend for each group respectively in the CFF4 conference. The vertical axis lists the number of words while the horizontal axis lists the speech events from the shortest to the longest in their category.

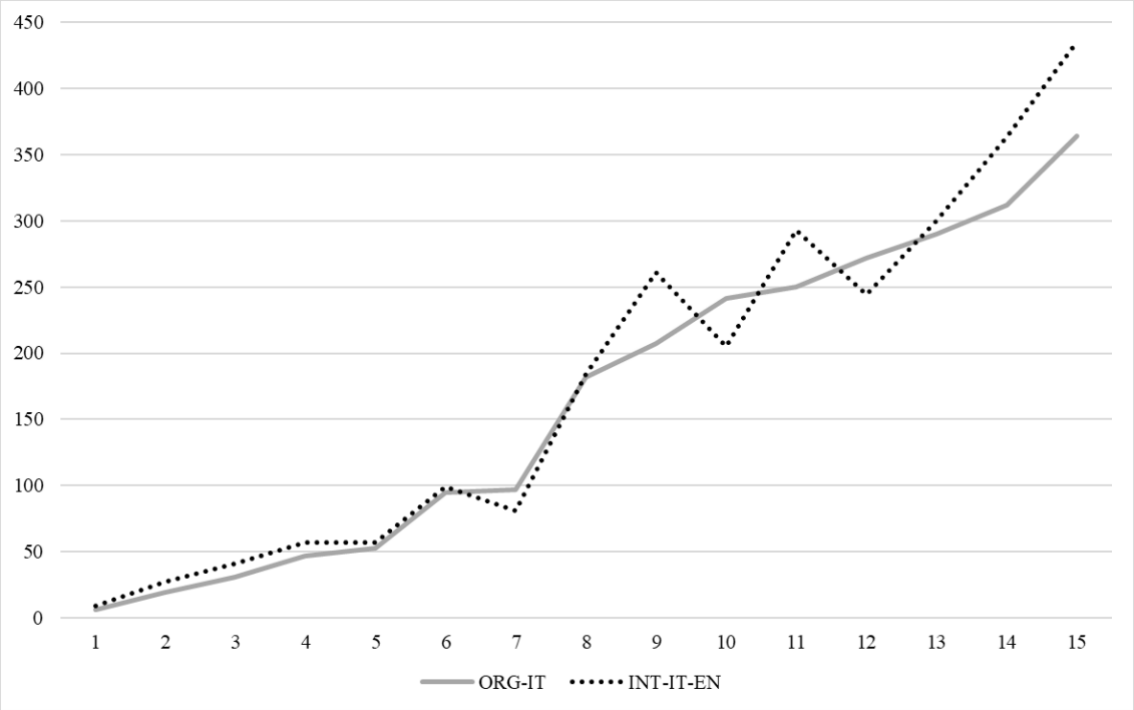


Figure 3: Number of words in Italian ST and English TT in CFF4 (ST < 500 words)

In this group of speech events taken from the CFF4 conference and with a length of less than 500 words (Figure 3), it is possible to note that there is almost always expansion with a higher word output in TT than in ST. We have already pointed out that the directionality factor probably plays a key role in this data set, as one of the interpreters on duty is a native speaker of English (UK-01). Table 10 lists the exact number of words in the interpreter's output for each speech event.

Number of words in ORG-IT (CFF4)	Number of words in INT-IT-EN (CFF4)
6	9
19	27
31	41
47	57
53	57
95	99
97	81
182	184
207	261
241	205
250	293
272	244
290	300
312	363
364	435

Table 10: Number of words in Italian ST and English TT in CFF4 (ST < 500 words)

Table 11 and Figure 4 instead report text compression and expansion in CFF4 speech events with ST larger than 500 words. The trend noted in Figure 3 is also noted in texts longer than 500 words. Again, we should consider what has already been commented about the directionality factor for this particular group of texts, where the working languages of one of the two interpreters are English as language A and Italian as language B.

Number of words in ORG-IT (CFF4)	Number of words in INT-IT-EN (CFF4)
547	721
590	645
1,299	1,214
3,805	4,238

Table 11: Number of words in Italian ST and English TT in CFF4 (ST > 500 words)

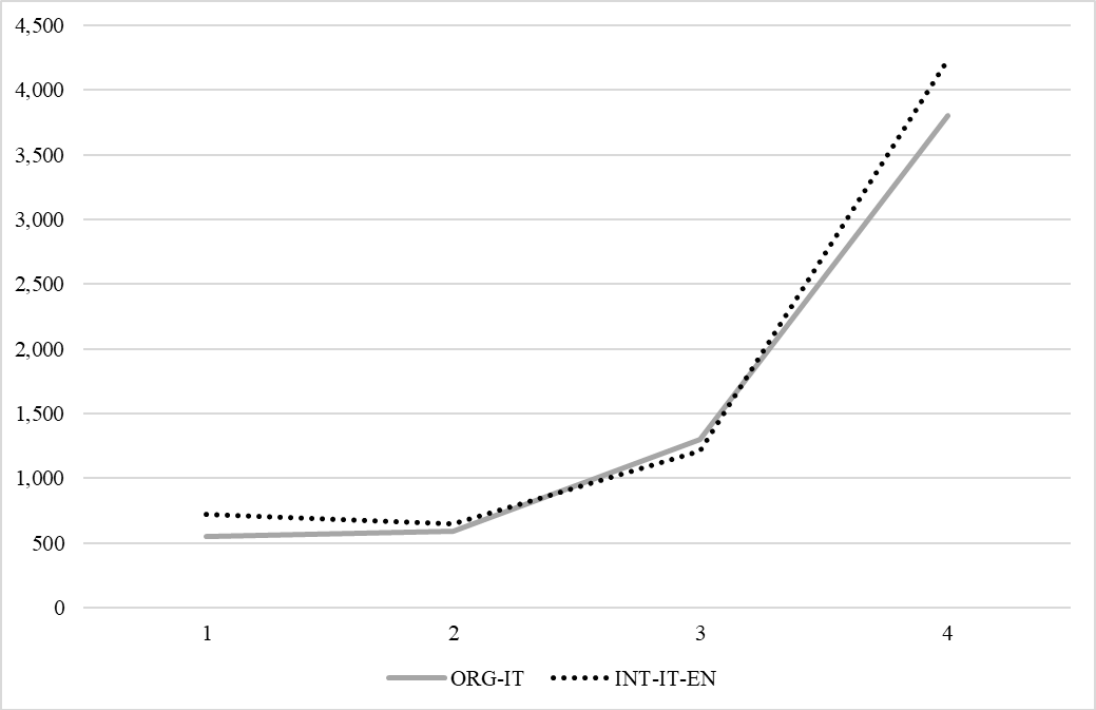


Figure 4: Number of words in Italian ST and English TT in CFF4 (ST > 500 words).

The same analytical procedure was applied to the ELSA conference. Figures 5 and 6 (along with Tables 12 and 13) report the text compression/expansion (between ST with less and more than 500 words respectively) and the corresponding TT.

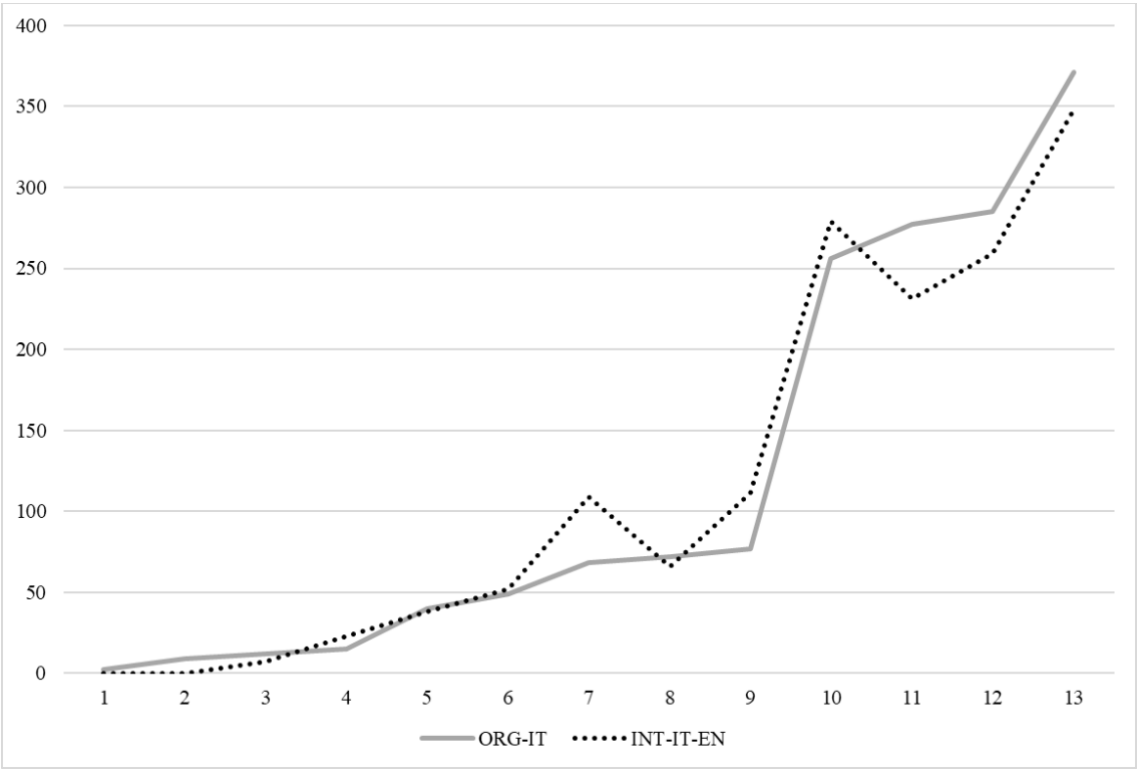


Figure 5: Number of words in Italian ST and English TT in ELSA (ST < 500 words).

In the speech events under 500 words in length selected from the ELSA conference, it is possible to observe that the number of words in the TT generally remains similar or slightly above the number of words in the relevant ST, with the exception of the last three texts (over 270 words) and the first two (extremely short and not translated by the interpreter). This finding challenges the general trend noted from a global observation of the data (it should be specified, however, that the spike in speech event number 7 is due to the failure to record the first few seconds in this speech).

Number of words in ORG-IT (ELSA)	Number of words in INT-IT-EN (ELSA)
2	0
9	0
12	7
15	23
40	38
49	52
68	109
72	66
77	111
256	279
277	231
285	259
371	348

Table 12: Number of words in Italian ST and English TT in ELSA (ST < 500 words)

Let us now consider Italian ST longer than 500 words and their TT into English in the ELSA conference (Table 13 and Figure 6). The two lines in Figure 6 show partial correspondence with the general trend of text compression in TT. However, there are a couple of exceptions where the TT has a slightly higher number of words than the ST. In any case, the other TT follows the general trend and show a lower number of words than the corresponding ST.

Number of words in ORG-IT (ELSA)	Number of words in INT-IT-EN (ELSA)
554	444
821	942
1,187	1,254
2,271	2,117
3,456	2,948

Table 13: Number of words in Italian ST and English TT in ELSA (ST > 500 words)

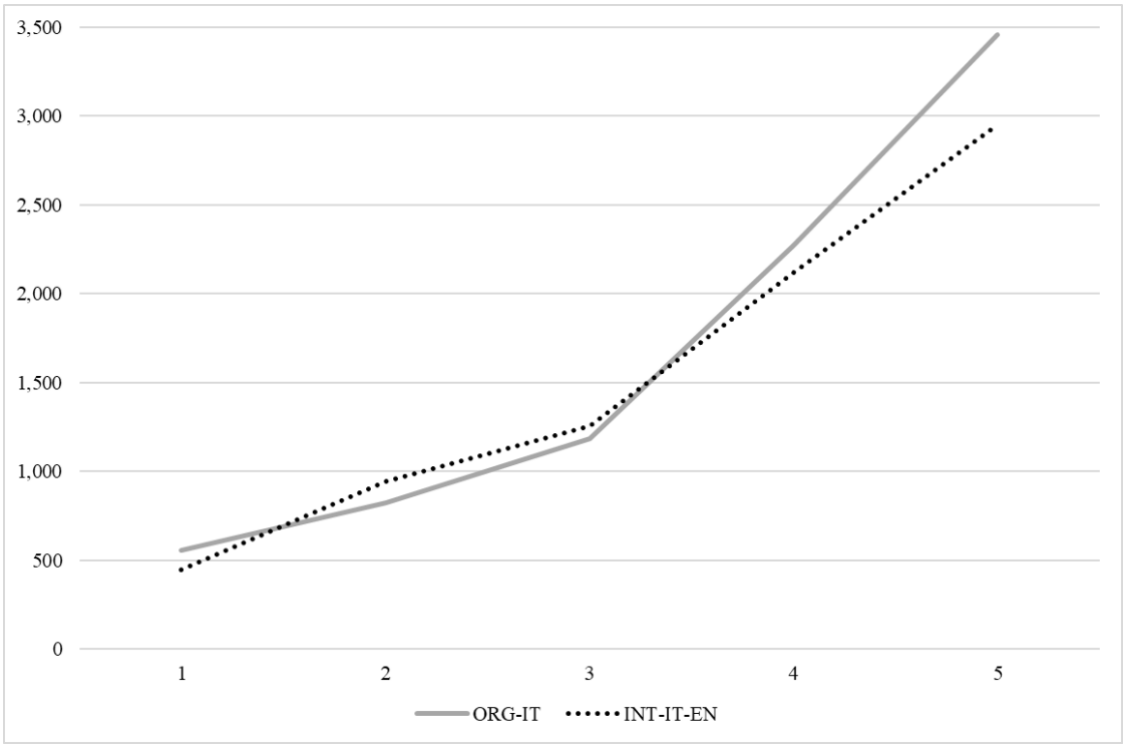


Figure 6: Number of words in Italian ST and English TT in ELSA (ST > 500 words)

Finally, the last conference to be analyzed is CFF5. As above, ST up to 500 words are considered first (Figure 7 and Table 14), and ST with more than 500 words and their TT are considered second (Figure 8 and Table 15).

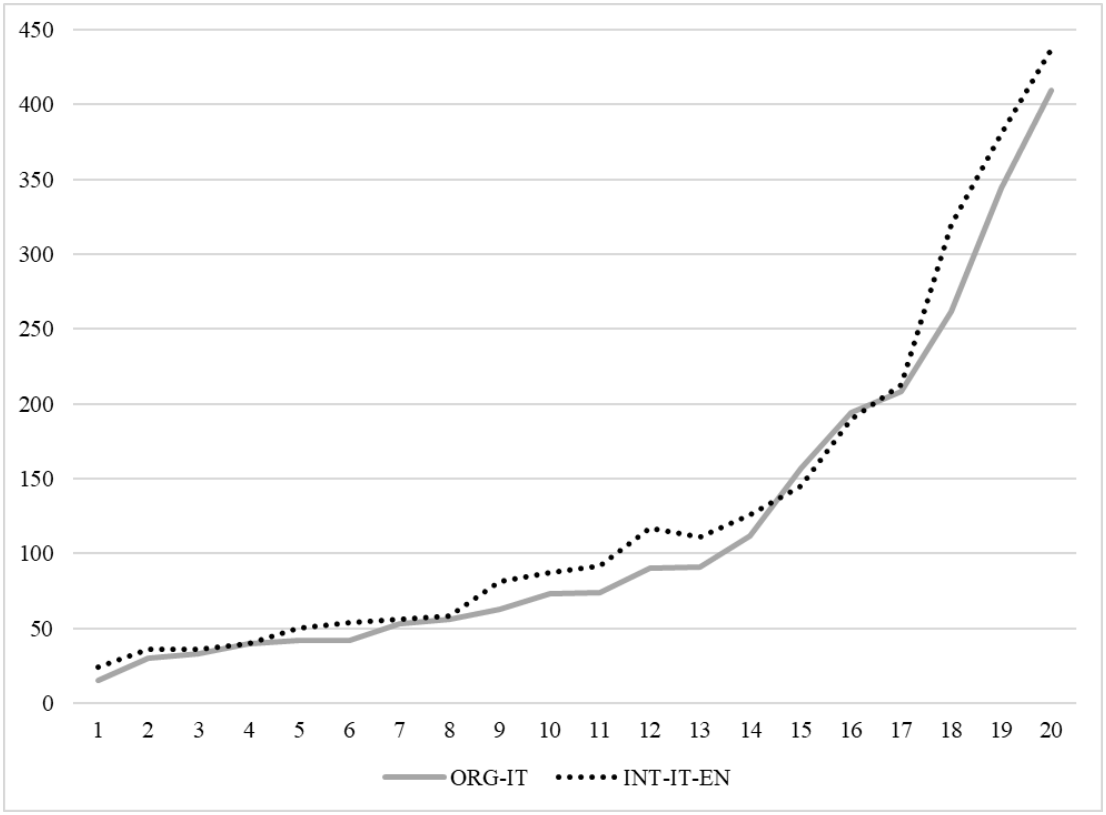


Figure 7: Number of words in Italian ST and English TT in CFF5 (ST < 500 words)

Number of words in ORG-IT (CFF5)	Number of words in INT-IT-EN (CFF5)
15	24
30	36
33	36
40	40
42	50
42	54
53	56
56	58
63	81
73	87
74	92
90	117
91	111
112	126
157	145
194	190
208	213
262	320
344	381
410	437

Table 14: Number of words in Italian ST and English TT in CFF5 (ST < 500 words)

In source speeches with less than 500 words from CFF5, the number of words is most times always higher in TT and lower in the related ST, in total contrast to the general trend whereby TT are always shorter in length than ST.

By contrast, the general trend is confirmed once again in the results shown in Table 15 and Figure 8, with ST larger than 500 words from the CFF5 conference. The number of words in TT is always lower than the number of words in the corresponding ST, with just one exception where the number of words is the same (i.e., 648, the shortest ST from this category), as shown in Table 15.

Number of words in ORG-IT (CFF5)	Number of words in INT-IT-EN (CFF5)
648	648
789	760
1,357	1,254
2,042	1,609
2,168	1,752
5,454	4,131

Table 15: Number of words in Italian ST and English TT in CFF5 (ST > 500 words)

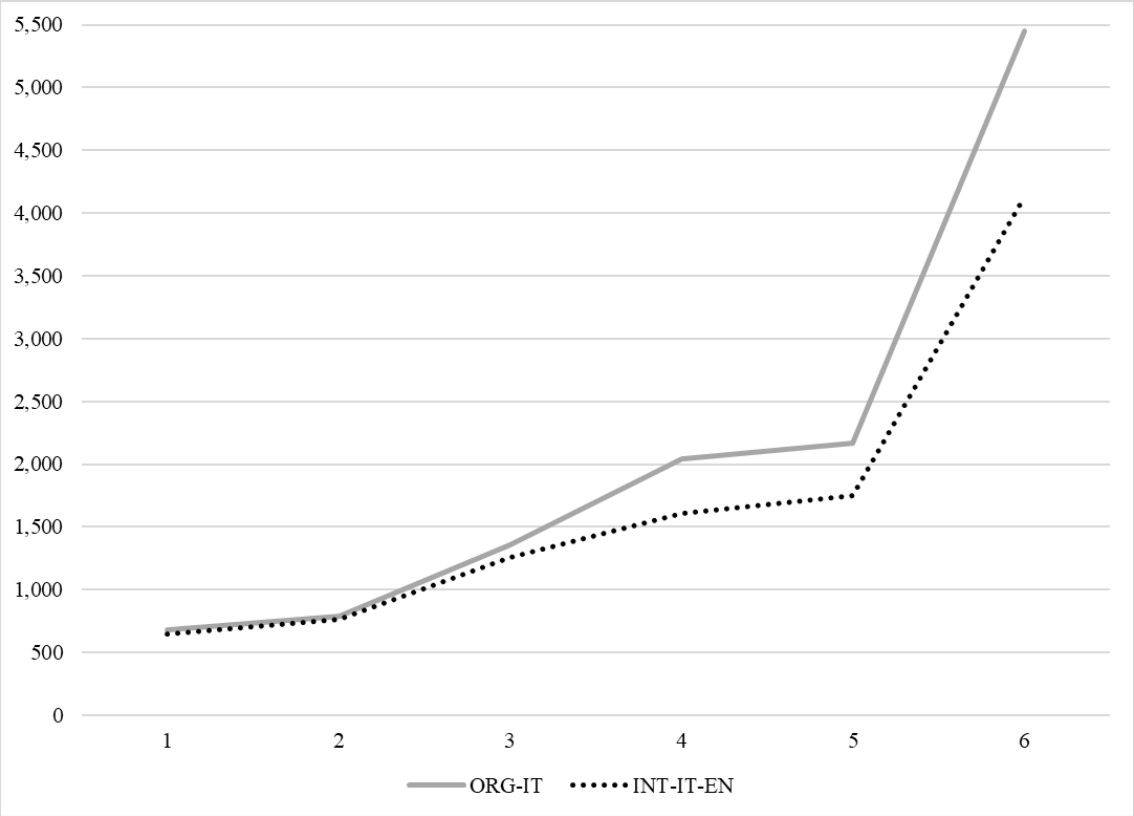


Figure 8: Number of words in Italian ST and English TT in CFF5 (ST > 500 words)

5. Discussion

Some interesting trends could be noticed in the comparative analysis of the number of words present in the output of source speakers and simultaneous interpreters. Overall, TT have a shorter length (in terms of word count) than ST. However, some exceptions could be identified.

A first exception concerns TT in English as language A or language B. In the specific case represented in the DIRSI corpus with the CFF4 conference, the general trend is not followed by the interpreter working with English as language A. His output in (native) English is larger than the word count of the corresponding Italian ST. This might be explained by the greater language availability of interpreters when working towards their A language. However, this did not occur in the opposite directionality, as the output of both interpreters in the same conference is in line with the general trend, that is, they produce fewer words in the Italian TT than the number of words in the English ST. This is also confirmed by all the other interpreters with Italian A serving in the other conferences included in the DIRSI corpus.

In the next step of the analysis, going from the general observation to a more in-depth level of detail, a different state of affairs was attested, depending on the length of the ST: when translating extremely short source speech events (with less than 500 words) interpreters tended to produce more words than there are in the original. This is seemingly related to the type of speech events involved. In fact, such a short duration is detectable in the following types of speech event: opening-closing remarks, floor allocation, procedure, or housekeeping announcements (as well as question, answer, and comment), that is, in all types of text with the exception of those classified as papers or lectures.

With these data in hand, it can be hypothesized that the expansion (or lack of compression) attested in the TT was due to the handling of the particular information contained in the reported speech events, as well as to the need to comply with certain rhetorical and politeness formulas not compulsorily present in Italian ST, but essential in the rendering of TT in English. Therefore, despite the fact that the interpreters worked towards their B language, it seems that the need to convey information explicitly and with appropriate linguistic-communicative choices induced them to expand their output, thus making the non-Italian-speaking participants fully and effectively share the situational context. Differences in the kind of time constraint posed by sequences of these speech events may also play a role. Conversely, when translating longer ST, interpreters tended to perform text reduction, and this may be due not only to possible cases of omission, but also to the streamlining of a sometimes redundant, repetitive if not wordy and poorly structured ST.

Below are some specific examples of speech expansion in TT retrieved from the DIRSI corpus. The transcripts are provided in tabular form, with the ST on the left and the TT on the right. Speech expansion is highlighted in bold. A literal, backtranslation into English of all Italian ST and TT is provided in square brackets.

The first example is from a closing remark by one of the organizers of the ELSA conference. Several elements that explain the larger number of words in the TT produced by the interpreter (IT-04), both obligatory and non-obligatory, can be identified. The largest expansion is due to the part where the interpreter informs listeners that it is not possible to translate what an audience member is saying, as they take the floor without using the microphone, as shown in (1a).

(1a) ELSA-023-ORG-IT	**ELSA-023-INT-IT-EN**
io concludo solo con ringraziamenti a Federica Leonardo Sabrina Francesca di ARCO a Milena // in particolare poi a Barbara che è la vera organizzatrice di tutto questo evento e che ha seguito tutta questa cosa // vedevo una mano alzata laggiù in fondo // **[audience speaks without microphone]** // va bene grazie mille ok // io credo che con con queste parole di augurio e di ringraziamento // di nuovo grazie a tutti per essere stati qua e ci sentiamo perché da lavorare ce n'è tanto	finally I'd like to conclude by thanking Federica Leonardo Sabrina Francesca from ARCO and Milena // in particular I'd like to thank Barbara who's the real organizer of thes- this whole event and for following this event through // I saw someone raises his hands at the towards the end of the conference room **and of course also this speech is delivered without using the microphone sorry for that** // thank you // I think that with these final remarks we can call it a day // I'd like to thank you all for being here with us and I think we shall speak to each other again because there is still a lot to do

Other non-obligatory expansions can be identified where contextual cues are provided more explicitly, as in (1b):

(1b) ELSA-023-ORG-IT	**ELSA-023-INT-IT-EN**
vedevo una mano alzata **laggiù in fondo** [I saw a hand up **back there**]	I saw someone raises his hands **at the towards the end of the conference room**

In addition, managing politeness in English appears to involve the addition of a number of lexical (and grammatical) elements which are not present in the Italian ST, as illustrated in (1c):

(1c) ELSA-023- ORG-IT	**ELSA-023- INT-IT-EN**
io concludo solo con ringraziamenti a [**I conclude** just with thanks to]	**finally I'd like to conclude** by thanking
in particolare **poi a** Barbara [in particular **then to** Barbara]	in particular **I'd like to thank** Barbara
e **ci sentiamo** perché da lavorare ce n'è tanto [and **let's talk** because to work there is much]	and **I think we shall speak to each other again** because there is still a lot to do

An instance of sentence completion in the TT can also be observed. This is done with respect to a unit of meaning in the ST that is not fully completed, as can be seen in (1d). In this example, there is also a case of text compression where a lexical repetition, that is, *parole di augurio e ringraziamento* ('words of wishes and thanks'), is reduced to two words in the English TT ('final remarks'):

(1d) ELSA-023- ORG-IT	ELSA-023- INT-IT-EN
io credo che con con queste parole di augurio e di ringraziamento // di nuovo grazie a tutti per essere stati qua [I believe that with with these words of wish and thanking // again thanks to all for being here]	I think that with these final remarks **we can call it a day** // I'd like to thank you all for being here **with us**

Examples (2a) and (2b) below are taken from an opening remark from the CFF5 conference (DIRSI-2007-05-11-VR-CFF5-001-ORG-IT). The conference started with some delay due to miscommunication about the opening time of the proceedings (a preliminary program had been circulated with a different time). In (2a), the interpreter (IT-01) provides a more extended explanation than the original speaker does, thus making contextual cues more explicit. In (2b), there are also TT expansions that occur in conjunction with several phenomena: different structuring of information that is segmented into several mutually independent utterances, management of politeness, and more explicit contextual cues.

(2a) CFF5-001-ORG-IT	CFF5-001-INT-IT-EN
ci scusiamo ancora per l'equivoco // l'orario // cercheremo di riparare [we apologize again for the misunderstanding // the time // will try to repair]	I do apologise for this problem **we had with the beginning of the conference** // **as some people knew it was at ten thirty for the preliminary programme**

(2b) CFF5-001-ORG-IT	CFF5-001-INT-IT-EN
questo **quinto seminario** riproduce una tradizione di incontro dei più interessati su alcuni aspetti più emergenti della fibrosi cistica **coinvolgendo** alcuni esperti che vengono da varie parti e che ci sembrano quelli che in questo momento possono dare su quei temi un segnale di aggiornamento efficace // s- **sono** i temi che conoscete in programma // [this **fifth seminar** reproduces a tradition of meeting of the most interesting on some aspects most emerging of cystic fibrosis **involving** some experts who come from various parts and who seem to us those who in this moment can give on those themes an effective signal of update // **they are** the themes that you know in the programme]	this **is the fifth edition of our spring seminar** // it is a tradition of meeting on some interesting aspects emerging aspects concerning cystic fibrosis // **we are pleased to involve** great experts coming from all around the world // we consider these experts to be those people who can give a a significant contribution as an an update // **and they will be discussing** the subjects that you have in the programme //

Example (3) has also been retrieved from a speech event classified as an opening remark (as in the case of the previous example) at the CFF4 conference (DIRSI-2006-05-20-VR-CFF4-002-org-it). Unlike examples (1) and (2), the interpreter here works towards his native language, as he is a native English speaker (UK-01). In (3), the expansion made by the interpreter might result from an attempt to make the message (expressed in the ST) less cryptic and more explicit by increasing its anaphoric references to the units of meaning expressed in the previous part:

(3) CFF4-002-ORG-IT	**CFF4-002-INT-IT-EN**
è per questo che a me piace questa giornata e piacciono questi incontri **perché mettono insieme proprio i due poli che si parlano** //	this is why I am very happy to open these proceedings and I'm very happy about these meetings because **they're an opportunity** to bring together two sides **which can exchange and exchange views with regard to solutions and analysis of problems** //
[is for this that I like this seminar and like these meetings because **bring together exactly the two poles that speak to each other**]	

Finally, examples (4a), (4b), and (4c) have been taken from a speech event categorized as a paper presentation in the ELSA conference (DIRSI-2006-10-19-FC-ELSA-012-org-en). In fact, it contains 2,045 words, so it does not belong to the group of very short speech events and, according to the general trend attested in other long speeches, it should be affected by text compression. However, in this case, the TT was expanded by nearly 8.5 percent (2,216 words). The interpreter (IT-04) translated from English into Italian (her A language), and this is sometimes reflected in more elaborate lexical choices and additional options, as shown in (4a). On the other hand, the source speaker is not speaking in her native language and makes use of English as foreign language or as a lingua franca (Bendazzoli 2017). Other instances of expansion in this example are possibly due to the interpreter's attempt to make up for a faulty wording of the ST, which is sometimes confused or expressed with lexical juxtapositions (4b and 4c), and to provide additional information for the sake of clarity, particularly about the English term *carer* which is also kept in Italian (4c). The speaker's difficulty in expressing herself in a foreign language can be perceived quite clearly from the recording of her presentation, though her language weaknesses cannot be perceived when listening to the interpreter.

(4a) ELSA-012-ORG-EN

since two thousand and four we changed our **organisation** in an independent carers' support centre because we **discovered** that sometimes the family carers as we call them

ELSA-012-INT-EN-IT

dal duemilaquattro abbiamo modificato la nostra **struttura organizzativa e siamo diventati** un centro di supporto indipendente ai carer perché **ci siamo resi conto** che a volte i family carer come noi li chiamiamo

[since two thousand and four we have changed our organizati**onal structure and we have become** an independent support center for carers because **we have become aware that** at times family carers as we call them]

(4b) ELSA-012-ORG-EN

the the family members who take care of their parents brothers and sisters partners children disabled children have sometimes trouble with home care organisations //

ELSA-012-INT-EN-IT

cioè i carer che sono membri della famiglia che si prendono cura dei propri cari **che possono essere** figli disabili genitori anziani o comunque familiari malati a volte si trovano in difficoltà nei confronti delle organizzazioni delle cure domiciliari

[**that is the carers who are** members of the family who take care of their dear ones **who can be** disabled children elderly parents or anyway ill relatives at times they find themselves in difficulty with respect to organizations of home care]

(4c) ELSA-012-ORG-EN

and as being a part of such home care organisation **would bring us in a difficult situation** to to help the **family carers**

ELSA-012-INT-EN-IT

e **il fatto che noi facessimo parte** di questa organizzazione di assistenza **di cura** domestica **ci ha fatto trovare nella situazione in cui ci era difficile** poter aiutare i family carer **cioè i familiari che si prendevano cura dei loro cari**

[and **the fact that we were part** of this organization of assistance of home care **has made us find in the situation in which we had difficulty** to help the family carer **that is the relatives who took care of their dear ones**]

6. CONCLUSION

A quantitative analysis of textual output in source speeches and their simultaneous interpretations in the DIRSI corpus (considering the number of words) showed that there is a general trend by which interpreted speeches always contain a lower number of words than their originals, regardless of directionality. The only exception to this trend were the target speeches in English produced by interpreter UK-01, who had English as language A. However, this result differs from what was verified in the Italian target speeches produced by Italian native interpreters, and also from the EPIC corpus, where working conditions might lead to reduced textual output in TT, even when produced by interpreters working from language B (or C) to language A.

In addition to the global observation of the data in the DIRSI corpus, a more in-depth analysis was conducted. This consisted of isolating the data on textual output in ST and TT in each conference making up the corpus. This deeper level of analysis showed that the general trend attested in the overall data does not hold constant for all kinds of speech events: TT produced from extremely short ST (under 500 words, typically opening/closing remarks, floor allocation, announcements) usually contain more words than there are in the corresponding ST. On the other hand, above the 500-words threshold the general trend is confirmed: that is, fewer words in TT than in ST. Among possible motivations for the expansion of TT related to shorter speech events, we noted the addition of more explicit information by the interpreter, the use of formulas for managing politeness in English, and the optimization of the TT when the ST displays incomplete or grammatically deficient sentences.

Overall, there were no instances of very marked expansions of TT compared to ST. However, the maintenance of a similar level of textual output (number of words) between the two types of texts is in sharp contrast with the general picture where the number of words produced in TT is always lower than the number of words produced in ST. Besides the particular features of the very short source speeches where this trend was registered, it is worth emphasizing that every ST was considered individually. Yet, those speech events were actually part of a seamless sequence making up the conference as a communicative event. The deployment of such a sequence may come with pauses that would provide interpreters with more leeway in managing the critical constraint of time in SI. Another important limitation of this study lies in that text compression/expansion

was measured in terms of the number of words, which is a rough indicator of a much more complex linguistic and cultural mediation activity.

REFERENCES

Abbasi, Atefeh and Mansour Koosha. 2016. Exploring expansion and reduction strategies in two English translations of Masnavi. *Advances in Language and Literary Studies* 7/2: 219–225.

Aston, Guy. 2018. Acquiring the language of interpreters: A corpus-based approach. In Mariachiara Russo, Claudio Bendazzoli and Bart Defrancq eds., 83–96.

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis and Elena Tognini-Bonelli eds. *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins, 233–250.

Bartłomiejczyk, Magdalena. 2022. Addressing others through an interpreter: Is the directness reduced across the pragmatic spectrum? *The Interpreters' Newsletter* 22: 1–19.

Bartłomiejczyk, Magdalena, Ewa Gumul and Danijel Koržinek. 2022. *EP-Poland*: Building a bilingual parallel corpus for interpreting research. *GEMA Online Journal of Language Studies* 22/1: 110–126.

Bendazzoli, Claudio. 2010. *Corpora e Interpretazione Simultanea*. Bologna: Asterisco.

Bendazzoli, Claudio. 2012. From international conferences to machine-readable corpora and back: An ethnographic approach to simultaneous interpreter-mediated communicative events. In Francesco Straniero Sergio and Caterina Falbo eds., 91–117.

Bendazzoli, Claudio. 2017. Benefits and drawbacks of English as a Lingua Franca and as a working language: The case of conferences mediated by simultaneous interpreters. In Cecilia Boggio and Alessandra Molino eds. *English in Italy: Linguistic, Educational and Professional Challenges*. Milano: FrancoAngeli, 119–141.

Bendazzoli, Claudio. 2019. Discourse markers in English as a target language: The use of *so* by simultaneous interpreters. *Textus* 32/1: 183–201.

Bendazzoli, Claudio. 2023. Breaching protocol and flouting norms on the European Parliament floor: Reactions from a micro- and macro-context perspective in 22 languages. *Contrastive Pragmatics* 4/1: 64–87.

Chernov, Ghelli V. 2004. *Inference and Anticipation in Simultaneous Interpreting: A Probability-Prediction Model*. Amsterdam: John Benjamins.

Cresswell, Andy. 2018. Looking up phrasal verbs in small corpora of interpreting: An attempt to draw out aspects of interpreted language. *inTRAlinea Special Issue: New Findings in Corpus-based Interpreting Studies*. https://www.intralinea.org/specials/article/2319

Defrancq, Bart, Koen Plevoets and Cédric Magnifico. 2015. Connective items in interpreting and translation: Where do they come from? In Jesús Romero-Trillo ed. *Yearbook of Corpus Linguistics and Pragmatics*. Singapore: Springer, 195–222.

Frankenberg-Garcia, Ana. 2009. Are translations longer than source texts? A corpus-based study of explicitation. In Allison Beeby, Patricia Rodríguez-Inés and Pilar Sánchez-Gijón eds. *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate*. Amsterdam: John Benjamins, 47–58.

Gile, Daniel. 2009. *Basic Concepts and Models for Interpreter and Translator Training: Revised Edition*. Amsterdam: John Benjamins.

Gumul, Ewa. 2017. *Explicitation in Simultaneous Interpreting: A Study into Explicitating Behaviour of Trainee Interpreters*. Katowice: Wydawnictwo Uniwersytetu Śląskiego.

Kajzer-Wietrzny, Marta. 2018. Interpretese vs. non-native language use: The case of optional *that*. In Mariachiara Russo, Claudio Bendazzoli and Bart Defrancq eds., 99–113.

Morselli, Niccolò. 2018. Interpreting universals: A study of explicitness in the intermodal EPTIC corpus. *inTRAlinea Special Issue: New Findings in Corpus-based Interpreting Studies*. https://www.intralinea.org/specials/article/2320

Pöchhacker, Franz. 1994. *Simultandolmetschen als Komplexes Handeln*. Tübingen: Gunter Narr.

Pöchhacker, Franz. 2004. *Introducing Interpreting Studies*. London: Routledge.

Riccardi, Alessandra. 1995. La conferenza quale evento comunicativo ed il ruolo dell'interprete. In Gerald Parks ed. *Miscellanea n. 2*. Trieste: Scuola Superiore di Lingue Moderne per Interpreti e Traduttori, 99–104.

Riccardi, Alessandra. 2005. On the evolution of interpreting strategies in simultaneous interpreting. *Meta* 50/2: 753–767.

Russo, Mariachiara. 1999. La conferenza come evento comunicativo. In Caterina Falbo, Mariachiara Russo and Francesco Straniero Sergio eds. *Interpretazione Simultanea e Consecutiva: Problemi Teorici e Metodologie Didattiche*. Milano: Hoepli, 87–102.

Russo, Mariachiara. 2018. Speaking patterns and gender in the *European Parliament Interpreting Corpus*: A quantitative study as a premise for qualitative investigations. In Mariachiara Russo, Claudio Bendazzoli and Bart Defrancq eds., 115–131.

Russo, Mariachiara, Claudio Bendazzoli and Bart Defrancq eds. 2018. *Making Way in Corpus-Based Interpreting Studies*. Singapore: Springer.

Russo, Mariachiara, Claudio Bendazzoli, Annalisa Sandrelli and Nicoletta Spinolo. 2012. The *European Parliament Interpreting Corpus* (EPIC): Implementation and developments. In Francesco Straniero Sergio and Caterina Falbo eds., 53–90.

Shalom, Celia. 2002. The academic conference: A forum for enacting genre knowledge. In Eija Ventola, Celia Shalom and Susan Thompson eds., 51–68.

Straniero Sergio, Francesco and Caterina Falbo eds. 2012. *Breaking Ground in Corpus-based Interpreting Studies*. Frankfurt: Peter Lang.

Ventola, Eija. 2002. Why and what kind of focus on conference presentations? In Eija Ventola, Celia Shalom and Susan Thompson eds., 15–50.

Ventola, Eija, Celia Shalom and Susan Thompson eds. 2002. *The Language of Conferencing*. Frankfurt: Peter Lang.

Corresponding author
Claudio Bendazzoli
University of Turin
School of Management and Economics
Department of Economics, Social Studies, Applied Mathematics and Statistics
Corso Unione Sovietica 218 Bis
10134 Torino
Italy
E-mail: claudio.bendazzoli@unito.it

received: March 2023
accepted: May 2023

RiCL Research in Corpus Linguistics

Non-fluency and language-pair specificity in Chinese-English consecutive interpreting: A corpus-driven study

Bing Zou – Binhua Wang
Guangdong University of Foreign Studies / China
University of Leeds / United Kingdom

Abstract – Language-pair specificity, which refers to linguistic and cultural differences between the language pair, has been hypothesized as one of the variables shaping the interpreting performance and product. The current study adopts a corpus-driven paralinguistic approach to testifying the language-pair specificity hypothesis. The corpus is a bilingual parallel corpus of *Chinese-English Interpreting for Premier Press Conferences*, which consists of 200,000 words/characters in total. The original and interpreted discourses are aligned at the sentential level and annotated at linguistic, paralinguistic, and extra-linguistic levels. The paralinguistic analysis focuses on non-fluency, specifically the different types of pauses and self-repairs. It is found that a majority of non-fluencies in the interpreted utterances are syntax-driven, which means that most of the pauses and self-repairs in Chinese-English interpreting are related to syntactical structures in the original speeches. The finding implies that language-pair specificity should be considered an important variable in research and training of interpreting between syntactically-contrastive languages.

Keywords – non-fluency; language-pair specificity; consecutive interpreting; Chinese-English interpreting

1. INTRODUCTION[1]

Translation and interpreting are conducted between two languages and cultures, so their products are shaped by the two distinct linguistic and cultural systems. On the one hand, it is this distinctness that endows translation and interpreting with possibility and necessity and, on the other, it is the very distinctness that poses challenges to translators and interpreters, and sometimes even makes them despaired to claim the sad fact of

[1] We would like to express our sincere gratitude to the editors for their efforts in revising the language and format of our paper. We also appreciate the support from the *Research Fund of Center for Translation Studies* (CTS202209), the *Guangdong Five-Year Plan Project on Philosophy and Social Science* (GD22WZX02–04; GD20WZX01–09), and the *Fujian Social Science Fund Youth Project* (FJ2021C111).

untranslatability or uninterpretability. The distinct linguistic (majorly syntactic) and cultural differences between the paired languages in the act of translation and interpreting are labelled as 'language specificity' or 'language-pair specificity' by previous scholars (Wilss 1978; Setton 1993; Gile 2004). Among previous studies, only a few (Setton 1993; Guo 2011; Wang and Gu 2016; Wang and Zou 2018, among others) examined problems in the English/Chinese language pair. The English and Chinese language pair is a typical representative of European and non-European language pairs, and the large linguistic and cultural divergence between them highlights the potential effects of language-pair specificity on English/Chinese interpreting. As empirical studies are still scarce on language-pair specificity in the English/Chinese language pair, issues are still awaited to be explored, such as causes of language-pair-specific problems and effects of them on the interpreting performance and product. The current study seeks to explore the relation between language-pair specificity and the interpreter's performance and product in Chinese-English interpreting by adopting a corpus-driven paralinguistic approach with a focus on whether and how non-fluency relates to language-pair specificity.

2. Language-pair specificity in interpreting

The discussion of language-pair specificity issues in interpreting started with observational studies, which identified language-pair-specific phenomena as problem triggers in interpreting between two languages and cultures that are different or distant from each other in linguistic structures and cultural conceptualization (Wang and Gu 2016; Wang and Zou 2018).

As one of the fundamental conceptualizations in interpreting studies, however, the *théorie du sens* ('The Interpretative Theory of Translation') did not treat language-pair specificity as a problem trigger in interpreting, and posited that interpreters' output "is, in principle, independent of the source language" (Seleskovitch 1978: 98). Although this assumption might represent a worthwhile effort to encourage interpreting practitioners to break away from the bound of the source language and not to be confined by the formal divergences between the source and target languages, other scholars (Moser 1978; Wilss 1978; Uchiyama 1991; Gile 1992, 2005, 2011; Riccardi 1996; Seeber 2007; Al Zahran 2021) have found that problems are triggered by language-pair specificity in interpreting. Among them, Wilss (1978: 343) proposed that "[a]ny transfer" (including translation and interpreting) is to a certain degree affected by the structural asymmetry (on morphemic,

lexemic, syntagmatic and/or syntactic levels) between the two languages involved. Wilss (1978: 350) even pointed out that "[a]ny SI process is language-pair-specific" due to the structural asymmetries or divergences, which is in line with the observations by Moser (1978) and Gile (1992). Uchiyama (1991) and Riccardi (1996) analyzed difficulties in interpreting triggered by syntactic differences in the Japanese-English and German-Italian language pairs, and proposed some coping strategies. Seeber (2007) and Gile (2005, 2011) further examined specific features of the difficulties and cognitive loads caused by language-pair specificity in interpreting. Al Zahran (2021) observed real-life data of English-Arabic simultaneous interpreting and found that the syntactic asymmetry between the English-Arabic language pair led to form-based processing by the interpreters.

As seen from the previous studies summarized above, the effect of language-pair specificity on interpreting cannot be neglected. According to Setton (1993) and Ra and Napier (2013), language-pair-specific problems are more salient between European-Asian language pairs (i.e., English-Chinese or English-Japanese) than European-European language pairs (i.e., English-French). In other words, interpreting between European and Asian languages poses special difficulties for interpreters. Setton (1993: 253–255) further suggested that the very differences between distant language pairs should be treated as "a catalyst" for invigorating future research and encouraged more researchers to investigate the language-pair-specific problems in interpreting between (Indo-)European and non-(Indo-)European language pairs.

Among the very few studies on English/Chinese interpreting (a typical representative of interpreting between the (Indo-)European and non-(Indo-)European languages), Setton (1993) observed the linguistic structural and morphological differences between English and Chinese, and discussed the difficulties caused by these differences in English/Chinese interpreting practice and training. Guo (2011) and Wang and Zou (2018) explored the effect of Chinese-English structural differences on simultaneous interpreting and consecutive interpreting respectively; they revealed that in Chinese-English interpreting the interpreter has to re-order Chinese front-loaded sentence structures into English back-loaded structures, and analyzed how such re-ordering efforts would cause extra difficulties and cognitive overloads to Chinese-English interpreters. Wang and Gu (2016) also observed the effect of language-pair specificity in English-Chinese simultaneous interpreting, and found that right-branching structures in English,

with corresponding information chunks that are syntactically different (left-branching) in Chinese, caused a lot of long pauses, information loss, and errors in the interpreting process and product. These studies suggest that language-pair specificity should be considered as one of the variables shaping the interpreting performance and product especially in the English/Chinese language pair.

It is worth noting that all these studies are based mainly on the analysis of linguistic features while a paralinguistic analysis might produce more evidence. It is also important to note that paralinguistic features (including filled/unfilled pauses, self-repairs, etc.) are typical of interpreting processes and products, and are explicit representations of interpreters' on-site performances, so a paralinguistic analysis of interpreted discourses would probably provide a new window for exploring the relationship between language-pair specificity and interpreters' performance. Therefore, the current study explores non-fluency as a typical paralinguistic feature of interpreting and an indicator for interpreters' on-site performance and discusses how it relates with language-pair specificity in consecutively-interpreted discourses in the Chinese-English language pair.

3. NON-FLUENCY IN INTERPRETING

The assessment of interpreters' performance is different from that of translators' due to the fact that interpreting involves a lot of non-verbal or paralinguistic elements. In other words, assessing interpreters' performance relies on multiple dimensions of evidence including not only linguistic (lexis, syntax, discourse, etc.) and extralinguistic (background information about the interpreter, speaker, audience/user, patron/organizer, etc.) aspects, but also paralinguistic (non-fluency, prosody, body gestures, etc.) aspects of the interpreting process and product (Zou and Wang 2014). As Setton (2011: 35) mentions, it is "pointless to attempt any realistic model of the process [of interpreting] without taking into account factors such as ... features of live speech like prosody," because these paralinguistic factors and features "give us ideas to explain the phenomena [in interpreting] —recasting, anticipation, added cohesive devices and so on" (*ibid.*: 68).

Among previous interpreting studies from the paralinguistic perspective, non-fluency remains a focus of discussion and has been considered an important indicator of interpreters' performance by multiple scholars (Mead 2000; Tissi 2000; Cecot 2001; Ahrens 2005; Pradas Macías 2006, among others). According to Tissi (2000) and Cecot

(2001), non-fluencies in interpreting can be subdivided into two major categories: 1) silent or unfilled pauses —including (non-)communicative pauses, (non-)grammatical pauses, segmentation/(non-)juncture pauses, etc.— and 2) disfluencies, including filled pauses, parenthetical sentences, utterance interruptions like repetitions, restructuring (self-correction/self-repair), false starts, etc. Mead (2000) and Tissi (2000) conducted experiments on student interpreters whose first language is Italian and asked the subjects to simultaneously interpret from German into Italian and consecutively interpret between English and Italian, respectively. Tissi (2000) found that compared to source texts, interpreters' target texts contain fewer but longer silent pauses, more grammatical pauses, and more vowel and consonant lengthening. Mead (2000) concluded that interpreting into the second language is more fluent and involves significantly more total pauses and higher filled pause times, and put forward that the causes of pauses include difficulties with formulation (of lexis/grammar) and notes, as well as logical doubts. Cecot (2001) and Pradas Macías (2006) carried out experiments by inviting professional interpreters to finish simultaneous interpreting tasks in the English-Italian and German-Spanish language pairs, respectively. Cecot (2001) revealed that segmentation pauses are most frequently used by the subjects, while Pradas Macías (2006) discovered that frequent silent pauses negatively influence users' assessment of interpreting quality. Ahrens (2005) examined natural data of English-German simultaneous interpreting by professional conference interpreters and found that, compared to the source-text speaker, interpreters have a lower rate of articulation, and make less but longer pauses.

In the English/Chinese language pair, particularly, a number of scholars have cast light on the issue of (non-)fluency in interpreting practice and training. Fu (2013) and Yuan and Wan (2019) examined the impact of directionality on student interpreters' fluency in consecutive and sight interpreting tasks respectively, and found that directionality significantly correlates with fluency performance. Jiang and Jiang (2019) and Song *et al.* (2021) invited student interpreters to finish sight interpreting and simultaneous interpreting tasks respectively, and concluded that maximum dependency distance and input rate have a significant impact on fluency performance. Tang (2020) proposed a framework of categorizing student interpreters' self-repairs in consecutive interpreting. Xu (2010) and Qi (2019) investigated the causes of professional interpreters' pauses: Xu revealed that the triggers of pauses in consecutive interpreting include organizing information, retrieving target language, and modifying production, while Qi

discovered that the causes include loosening compact structures, segmenting long information units, and explicating logical connectors. Fu (2012) and Wang *et al.* (2019) investigated pauses in student interpreters' consecutive interpreting products: Fu found that directionality has a significant impact on the frequency of silent pauses rather than on that of filled pauses, while Wang *et al.* (2019) revealed that the level of interpreting competence significantly impacts the frequency of silent and filled pauses. Wang and Li (2015) compared the fluency performances of expert and trainee interpreters in a simultaneous interpreting experiment. They discovered that, compared to trainees, experts have more pauses for monitoring production and adopting strategies, fewer pauses for formulating, waiting, conceptualizing and split attention, and more pauses occurring at major syntactic junctures. Shen *et al.* (2019:135) examined the natural data of expert interpreters' consecutive interpreting products and found that experts' pauses are motivated for "retrieving lexical and morphological information, eliminating logical doubt, and explicating cultural connotation."

The review of previous studies above comes up with some common findings: 1) non-fluency has an impact on interpreters' performance and users' evaluation, 2) directionality and levels of competence have an effect on interpreters' fluency, 3) patterns of interpreters' and speakers' fluency are not the same, and 4) possible causes of interpreters' non-fluency include lexical/morphological, syntactic/grammatical, and logical and cultural difficulties. These findings, especially the last one, imply that interpreters' performance, as measured by fluency indicators (like pauses, self-repairs, etc.), is prone to challenges posed by the features of the source-language discourse which are distinct from that of the target-language discourse, or specifically, the language-pair-specific differences. It is a pity that these studies did not move further to explore the link and interaction between language-pair specificity and interpreters' non-fluency, which enlightens and motivates us to conduct the current study for a further investigation.

4. RESEARCH DESIGN

4.1. Research questions

As mentioned above, the current study takes a paralinguistic approach to testifying the language-pair specificity hypothesis in Chinese-English consecutive interpreting. The objective of the study is to investigate the effect of language-pair specificity, as reflected

by linguistic (syntactic) changes or shifts in interpreting products, on interpreters' performance, which is measured by such non-fluency indicators as filled/silent pauses, juncture/non-juncture pauses and self-repairs. Two research questions are examined:

1) What are the patterns of interpreters' non-fluencies in the Chinese-English consecutive interpreting products?
2) What are the causes of interpreters' non-fluencies in the Chinese-English consecutive interpreting products?

4.2. The corpus and the processing of data

The two questions are explored through a corpus-driven approach. The corpus employed is the *Chinese-English Interpreting for Premier Press Conferences* corpus, a self-built corpus consisting of original Chinese political discourses (with 121,877 characters) and corresponding consecutively interpreted English discourses (with 97,239 words). The bilingual corpus materials were collected from the annual 'Premier Meets the Press' conferences during China's 'Two Sessions' of the congress from 1998 to 2012. The speakers involved include Premier Zhu Rongji (from 1998 to 2002), Premier Wen Jiabao (from 2003 to 2012), and journalists from news agencies all around the world. The seven interpreters involved are all from China's Ministry of Foreign Affairs, and their interpreting services adopt a consecutive mode for the press conferences which take the form of questions and answers. The corpus materials are aligned at the sentential level (using the alignment tool *ABBYY Aligner 2.0*)[2] and annotated at linguistic (part-of-speech, using the tagging tools *TreeTagger 2.0*[3] for English and *Yacsi 0.96*[4] for Chinese), manually annotated paralinguistic (pauses, self-repairs, etc.) and manually annotated extra-linguistic (information about involved speakers and interpreters, etc.) levels.

In the current study, two types of non-fluencies are manually annotated, retrieved, and analyzed: pauses and self-repairs. Among them, four types of pauses were detected: 1) silent pauses, namely a period (over 0.25 seconds in the current study) of no articulation by the interpreters, marked with the symbol '…', 2) filled pauses (a vocalized but non-word period, marked with the symbols 'ah, eh, em, er, uh, um'), 3) juncture pauses (the

[2] https://www.abbyy.com/
[3] https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[4] http://corpus.bfsu.edu.cn/TOOLS.htm

filled and filled plus silent pauses that occur at grammatical junctures or between syntactic units, including the phrase and clause, but excluding the sentence), and 4) non-juncture pauses or the silent and filled pauses that do not occur at grammatical junctures.

Besides pauses, three types of self-repairs were detected and manually annotated: 1) repetitions (namely the cases where interpreters repeat his/her preceding discourse), 2) self-corrections (cases where interpreters correct a mistake in their preceding discourse), and 3) reformulations, which are the cases where interpreters reformulate or restructure their preceding discourse.

Table 1 shows the corpus annotation scheme. The set of annotation symbols adopts a simplified method by using initials of the annotated non-fluency phenomenon. For instance, SJPY represents 'Silent Juncture Pauses-Yes', while SJPN stands for 'Silent Juncture Pauses-No' referring to silent non-juncture pauses. SRRF is an abbreviation of 'Self-Repairs Re-Formulations' referring to reformulations, a subtype of self-repairs.

Non-fluency indicators		**Annotation examples**
Major types	**Subtypes**	
Pauses	Silent juncture pauses	<SJPY>...</SJPY>
	Silent non-juncture pauses	<SJPN>...</SJPN>
	Filled juncture pauses	<FJPY>ah</FJPY>
	Filled non-juncture pauses	<FJPN>ah</FJPN>
Self-repairs	Repetitions	<SRRP> </SRRP>
	Self-corrections	<SRCR> </SRCR>
	Reformulations	<SRRF> </SRRF>

Table 1: The corpus annotation scheme

The processing of the data adheres to four principles:

1) Pauses between sentences, including both silent and filled ones, are not counted as juncture pauses in the current study, since the major function of this type of pauses is either striving for breathing or holding the floor. Only pauses at phrasal and clausal junctures are counted as juncture pauses.

2) Silent pauses that co-occur with filled pauses in neighboring positions are counted in the statistics, since they contain information about interpreters' struggle against difficult situations.

3) Self-repairs are counted as an independent indicator of non-fluency, rather than as a subtype of pauses.

4) Only self-repairs that co-occur with pauses are counted in the current study, while those that occur without pauses in their neighboring positions are not included in the statistics, because the self-repairs that appear alone reflect the fact that the repairing process is not a big cognitive trouble for the interpreters.

The process of data retrieval involves: 1) using *ParaConc 296*[5] to withdraw concordances that contain non-fluency indicators, 2) (manually) excluding cases that disobey the principles of data processing, 3) (manually) marking the causes of the non-fluency cases. The cause-marking process adopts a data-driven approach, that is, the categorization of the causes is gradually formulated along the process of marking. The specific categories of the causes are to be further discussed below.

5. RESULTS AND DISCUSSION

5.1. Patterns in the interpreters' non-fluencies

Altogether 643 interpreter pauses were found in the corpus data, including 249 filled pauses (38.72%) and 394 silent pauses (61.28%), as well as 206 juncture pauses (32.04%), and 437 non-juncture pauses (67.96%). Table 2 presents the raw figures of different types of interpreters' pauses.

	Juncture pauses	**Non-juncture pauses**	**Total**
Filled pauses	94	155	249
Silent pauses	112	282	394
Total	206	437	643

Table 2: Interpreters' pauses in the corpus

In Table 2, it is seen that filled/silent pauses, silent juncture pauses (112; 54.37%), and silent non-juncture pauses (282; 64.53%) are more frequently attested than filled juncture pauses (94; 45.63%) and filled non-juncture pauses (155; 35.47%), respectively. This fact reflects that interpreters tended to make silent pauses more often than filled pauses, whether at juncture or non-juncture positions in their utterances. This implies that the professional interpreters might try to minimize the influence of their vocalized filled pauses and show a preference for keeping silent pauses as short as possible rather than producing articulated filled pauses.

[5] https://paraconc.com/

As regards juncture/non-juncture pauses, silent non-juncture pauses (282; 71.57%) are far more than silent juncture ones (112; 28.43%). An explanation for this result might be that the current study excludes those silent juncture pauses that occur alone at junctures and do not co-occur with filled pauses in their immediate adjacent positions. But the fact that filled non-juncture pauses (155; 62.25%) are more than filled juncture ones (94; 37.75%) shows that interpreters came across cognitive difficulties when they had uttered parts of a phrase or clause more often than the occasion when they had uttered a complete phrase or clause and started uttering the next phrase or clause. It is at the point of the filled non-juncture pauses when the interpreters 'suddenly' realized the cognitive challenge and tried to figure out a solution to that challenging situation or made a remedy for a mistake in the preceding discourse. The level of difficulties rises when the interpreters were 'in the process' of organizing a target utterance than when they had already 'finished the process' of uttering the target discourse.

In terms of the interpreters' self-repairs, a total of 225 self-repairs are identified in the corpus, including 41 repetitions, 86 self-corrections and 98 reformulations. Table 3 summarizes the distribution of self-repairs that co-occur with four different types of pauses. It should be noted that, as mentioned above, statistics in the current study merely include those self-repairs that co-occur with pauses, so there are some more cases of stand-alone self-repairs not reported in the statistics. As is shown in Table 3, all the three types of self-repairs tend to co-occur more often with silent pauses (183; 81.33%) than filled pauses (42; 18.67%) and co-occur more frequently with non-juncture (181; 80.44%) pauses than juncture pauses (44; 19.56%).

	Filled/Silent pauses		**Juncture/Non-juncture pauses**		**Total**
	Filled	Silent	Juncture	Non-juncture	
Repetitions	9	32	7	34	41
Self-corrections	9	77	14	72	86
Reformulations	24	74	23	75	98
Total	42	183	44	181	225

Table 3: Interpreters' self-repairs in the corpus

Examples (1) to (3) present a demonstration of the concordance of the annotated three types of self-repairs retrieved from the corpus data. Provided together with the source and target discourses are literal translations of the source discourses. The literal English translations try to deliver word-for-word information about the source Chinese discourse

and maintain the original sequence of the Chinese words and punctuations. Some words are added in the literal translations for the sake of making the clause or sentence complete and comprehensible. The added words in square brackets correspond to what is missing in the source Chinese discourse in the same position. The underlined and double-underlined signs mean that the causes of the non-fluent phenomena in the marked parts are lexis-driven and syntax-driven respectively, which will soon be discussed below.

In Example (1), the interpreter made a repetition when encountering a lexical problem with '吃掉' (*chi diao*: 'have eaten'). The interpreter realized that it would not be appropriate to collocate the noun 'deficit' with the verb 'eat', and during the period of a silent non-juncture pause, he quickly figured out a solution after repeating the words 'it's not'.

1. Source discourse:

   [02-27] 至关重要 的 是 ， 我 这个 赤字 不 是 用在 弥补 经常性 的 预算 方面 ， 没有 把 它 吃掉 ， 是 用在 基础 设施 建设 方面

   Literal translation:

   Most importantly, this deficit is not used on the aspect of making up the regular budget's deficiency, [we] have not eaten it, [but] used [it] on the aspect of infrastructure development.

   Target discourse:

   [02-27] Most importantly, the deficit is not incurred to make up the deficiency in our regular budget. <SRRP>It's not <SJPN>...</SJPN> it's not</SRRP> consumed; rather the deficit is used to develop infrastructure projects.

In (2), the interpreter first came across a null-subject clause and added the subject 'we' after a filled non-juncture pause. Then, the interpreter found that there is no tense marker either, so he decided to add the modal verb 'should'; the uttering of this word is incomplete (as seen in the transcription 'sh-'), possibly due to the fact that he found modal verb 'should' not to be a proper choice in the linguistic context. As a result, the interpreter changed his idea and made a self-correction by using the modal verb 'would'. This is also an example of lexical challenge to interpreters. It reflects that modal auxiliary verbs are often troublemakers for interpreters especially in political settings, which is in line with the findings in previous studies on political interpreting (Li 2018).

2. Source discourse:

   [99-146] **第二 是 开放 中国 的** 电信 **市**场， 让 **外国** 资本 进入 **中国 的** 电信 **市**场

   Literal translation:

   Secondly [it] is [to] open China's telecommunications market, let foreign capitals enter into China's telecommunications market.

   Target discourse:

   [99-146] Secondly, <FJPY>er</FJPY> we <SRCR>sh- <SJPN>...</SJPN> would</SRCR> open the telecommunications market in China to foreign investors.

In (3), the interpreter also ran into a null-subject clause and attempted to adopt a different way of handling the situation. He tried to use a passive sentence by placing the object 'students [and minors]' in subject position, but soon found that the result of such an endeavor would be problematic. The passive sentence, if completed, would be 'no students and minors are allowed to engage in dangerous activities'. In this situation, the problem is that the students and minors would be described as being self-voluntary to take the action, which is illogical and against the fact that they were actually forced or seduced to engage in those dangerous activities. It is based on these considerations during the period of a silent non-juncture pause that the interpreter decided to make a reformulation by restructuring the whole utterance. The process of making such a decision did cause a lot of cognitive overload to the interpreter. The underlying reason behind this is that Chinese is a connotative language which usually conceals the real actor of an action, and that the addressee always has to dig it out after a cognitive process of reasoning. In the example, the actor 'anyone' is not found in the source discourse, so that the interpreter had to reason it out by consuming his own cognitive resources. The very example reveals the possible cognitive overload that might be exerted on interpreters due to the language-pair specificity in the English-Chinese language pair.

3. Source discourse:

   [01-221] **也就是**说， 绝对 **不能 允**许 **学生 和 未成年 的 儿童** 进行 **危**险 **生命 危**险 **的** 劳动

   Literal translation:

   Also [that] is to say, [we] definitely cannot allow students and minors/children to engage in activities dangerous to life.

   Target discourse:

> [01-221] That is, <SRRF>no students <SJPN>...</SJPN> we will never allow anyone to ask the students</SRRF> or minors to engage in activities and work that will pose a danger to their life.

5.2. Causes of interpreters' non-fluencies

In order to find out what caused the above non-fluencies in interpreting, this section examines the structural changes or shifts made by the interpreters in relation to all types of non-fluency.

By means of a data-driven method of cause-marking, as mentioned above, the causes for interpreters' non-fluencies in the current study are finally categorized into three types: lexis-driven, syntax-driven, and other types, which include sensitive topics, cultural factors, etc. Table 4 presents the figures for the three types of causes of the interpreters' seven types of non-fluencies.

		Syntax-driven causes	**Lexis-driven causes**	**Other causes**	**Total**
Pauses	Filled pauses	173	76	0	249
	Silent pauses	221	166	7	394
	Juncture pauses	152	48	6	206
	Non-juncture pauses	242	194	1	437
Self-repairs	Repetitions	22	17	2	41
	Self-corrections	19	67	0	86
	Reformulations	64	34	0	98

Table 4: Totals of interpreters' pauses

As seen in Table 4, syntax-driven causes constitute the largest proportion among all the four types of pauses, that is, 69.48 percent for filled pauses (173), 56.09 percent for silent pauses (221), 73.79 percent for juncture pauses (152), and 55.38 percent for non-juncture pauses (242). Among all the three types of self-repairs, most of the causes of repetitions (22; 53.66%) and reformulations (64; 65.31%) are syntax-driven, while most of the causes of self-corrections are lexis-driven (67; 77.91%). The reason why most self-corrections are caused by lexical problems could be noticed in example (2), in which the self-correction by the interpreter is to solve a simple lexical problem. Actually, most cases of self-corrections found in the corpus data are related to the treatment of lexical problems such as that illustrated in (2).

Overall, except for the cases of self-corrections, the results about the causes of interpreters' non-fluencies are inspiring, for it implies that syntactic problems seem to be the major causes of the interpreters' pauses, repetitions and reformulations as discovered through the corpus data. In other words, syntactic differences between the source language (Chinese) and the target language (English) cause most of the challenges to interpreters. In what follows, the current study will discuss the interactive relationship between language-pair specificity and interpreters' performance as reflected by non-fluency indicators, and some characteristic examples will be provided.

In (4), below, the interpreter first made a silent pause between the words 'we' and 'need to', all of which are words added by the interpreter. The addition of 'we' is made for the consideration that the first part of the source discourse is a null-subject clause, so the interpreter had to add an appropriate subject by referring to the context. The addition of 'need to' is due to the fact that the clause does not have an obvious tense marker, so the interpreter had to figure out a solution and finally chose the modal auxiliary verb 'need to' without clearly mentioning the tense. All these complex thoughts and actions of the interpreter happened in a flash, or exactly during the period of the silent pause, on the very site of the interpreting activity. Actually, null subject and absence of tense markers are typical features of Chinese but are rarely seen in English. Completing the clause or sentence by adding a proper subject and a tense marker (or a substitute for tense markers) did consume a lot of the interpreter's cognitive resources, as is reflected by the silent non-juncture pause that interrupted the interpreter's fluency. It is interesting to compare the second part with the first part of the source discourse. The second part is also a null-subject clause with no specific tense marker, but the interpreter went smoothly, without any pauses, by adding the subject 'we', the modal verb 'need to' (as a substitute for tense markers) and the conjunctive words 'and' and 'also'. The reason for the smoothness of the second part may be that the interpreter had just overcome the cognitive challenge in the first part and immediately drew experiences from it.

4. Source discourse:

   [07-120] **减少** 权力 过分 **集中 的** 现象， **加强 人民** 对 **政府 的** 监督

   Literal translation:

   [We] reduce the phenomenon of over-concentration of power, [and] enhance people's oversight over the government.

   Target discourse:

> [07-120] We <SJPN>...</SJPN> need to reduce the over-concentration of power, and we also need to enhance the oversight <FJPN>er</FJPN> <SJPN>...</SJPN> <SRRF>of <SJPN>...</SJPN> over</SRRF> the government by the people.

The other three non-juncture pauses in the second clause of the target discourse have something to do with a syntactic structure specific to Chinese, that is, the left-branching modifying structure marked with the character '的' (*de*, similar to the possessive *'s* in English). In English, possessive *'s* always goes after a noun while, in Chinese, preceding the character *de* there can be a noun, a phrase, a simple clause, or even a complex clause. This language-specific difference often causes trouble to Chinese-English interpreters (Wang and Zou 2018). As is shown in (4), a word-for-word rendering of the second part of the source discourse is 'enhance people-toward-government *'s* oversight'. The long left-branching structure before the word 'oversight' makes the rendering awkward and grammatically incorrect. It is a common practice for translators and interpreters to either render the *de* structure (like '人民 的 利益', literally translated as 'people *'s* interests') into the possessive *'s* structure (like 'the people's interests'), or render the *de* structure into the possessive *of* structure (like 'interests of the people') if a noun precedes the character *de*. However, if what precedes the character *de* is a phrase or a clause, the situation becomes complicated and has to be treated carefully in a case-by-case manner. In the very example, the interpreter should have easily rendered the second part like 'enhance people's oversight over the government', but since the interpreter had already uttered the word 'oversight' beforehand, the interpreter had no choice but to keep moving and adopting another solution ('enhance the oversight over the government by the people'), possibly for the sake of avoiding a whole restructuring of the already uttered words. It is also noticeable that the interpreter first used the word 'of' after the word 'oversight': probably at that moment in the interpreter's mind came out the expression 'oversight of the people', but he soon realized that this might not be a correct rendering, so he just changed the structure into 'oversight … by the people'. All these complicated thoughts and actions took place within the short period of one filled pause and two silent pauses, just as the annotation in the target discourse shows. The restructuring of the *de* structure, together with the addition of the missing subject and tense marker, helps reveal the cognitive difficulties, underlying the interpreter's performance, that language-pair specificity in the English-Chinese language pair might give rise to.

In (5), below, the interpreter made two reformulations. Both of the self-repairs are accompanied by filled or silent pauses, which serves as triangulated evidence of the cognitive challenge that the interpreter was experiencing. In this example, the first reformulation was made due to the fact that the second clause of the source discourse (literally translated as 'affected some Japan's Banks' debts') does not have a subject, which results in a silent pause during which the interpreter had to add an appropriate subject ('this problem'). It is within the period of this pause that the interpreter might have had a quick search in his memory or a quick check of his notes. Absence of subject is a typical phenomenon in Chinese, but rarely seen in English. Finding out what the missing subject is greatly relies on the reasoning and even guessing of the readers or listeners from the context. In the example, at the moment when the speaker (i.e., the Premier) quoted and responded again to the journalist's question about Japanese banks' debts, it is already 28 sentences away from the moment when the question was proposed by the journalist. So, the interpreter had no choice but to search his memory or check his notes in order to tackle his uncertainty about what the journalist mentioned before, that is, what affected Japanese Banks' debts. Such a process, taking place in a very short period of time though, requires a huge amount of cognitive efforts.

5. Source discourse:

   99-123] 对于 刚才 你的 提问 ， 影响 了 一些 日本 银行 的 这个 债务 ， 我 也 感到遗憾 ， 但是 我 想 ， 今后 可能 也 不会 再有 了吧

   Literal translation:

   Regarding just now your question, [the problem] affected some Japanese banks' debts, I also feel regretful, but I think, in the future maybe [such problem] will not arise anymore.

   Target discourse:

   [99-123] I regret that <SRRF>this question <SJPN>...</SJPN> this problem</SRRF> has somewhat affected <FJPN>er</FJPN> some Japanese banks. <SRRF>And <FJPY>er</FJPY> <SJPY>...</SJPY> but</SRRF> I guess maybe there will not be such situation in the futures.

In (5), after solving the problem with the missing subject, the interpreter came across a syntactic problem which is caused by the *de* structure, as is seen in the second clause of the source discourse. A literal translation of this structure would be 'some Japanese banks' debts'. The interpreter might have checked his memory and notes, and finally was sure that the word 'debts' did not appear in the journalist's question. So, the interpreter

decided to delete it in the target discourse after a period of filled non-juncture pause. This reflects that in political settings, interpreters might play the role of monitoring and checking the authenticity of the source discourse by the speaker, whose role also brings to interpreters lots of cognitive overloads, as is seen from the occurrence of the filled non-juncture pause. In the second clause of the target discourse, the interpreter made the second reformulation, which is caused by a logic doubt. There are two explanations for the interpreter's use of the word 'and': one is that it serves as just a language filler which might help the interpreter gain some seconds for breath, the other is that it is a mistake made by the interpreter. Whatever the truth is, the filled and silent non-juncture pauses after the word do reveal that the interpreter immediately realized that the coordinating conjunction might not be appropriately appearing in the position. So, after a period of two pauses' time, the interpreter reformulated the logical link between the two clauses by using the contrasting conjunction 'but', whose process is in fact also in great demand of the interpreter's cognitive efforts.

The discussion of examples (4)–(5) above might help explain why most of the causes of interpreters' non-fluencies are syntax-driven. The language-pair-specific differences between English and Chinese, especially syntactic differences such as null-subject clauses, absence of tense markers, the *de* structures and invisibility of the actor of an action verb, did cause plenty of trouble to the interpreters who had to stop for a while to seek the most appropriate solutions. The very solution-seeking process demands that interpreters invest a lot of cognitive resources, leading to the fact that the interpreters had no choice but to sacrifice the fluency of their target utterances for the sake of gaining enough time for (re-)thinking, memory-retrieving, notes-checking, information-(re)organizing, repeating, self-correcting, reformulating, or restructuring. That is when the non-fluencies happen. In a word, the journey from language-pair specificity to interpreter's non-fluencies is a time-consuming and cognitive effort-consuming process which should be considered as an important factor in the evaluation and assessment of interpreters' on-site performance.

6. Conclusion

The current study conducted a descriptive corpus-driven investigation to identify the relationship between language-pair specific issues and non-fluency in interpreting. It is found that the language-pair-specific structural differences between English and Chinese

function as a cause for interpreters' non-fluencies including such pauses as filled/silent pauses and juncture/non-juncture pauses, as well as such self-repairs as repetitions and reformulations. The finding implies that, in addition to other major variables such as interpreter competence, on-site cognitive conditions and norms of interpreting (Wang and Gu 2016), language-pair specificity should be considered as one of the variables shaping the interpreting performance and product, especially for language pairs that contrast sharply in syntactic structures.

REFERENCES

Ahrens, Barbara. 2005. Prosodic phenomena in simultaneous interpreting: A conceptual approach and its practical application. *Interpreting* 7/1: 51–76.

Al Zahran, Aladdin. 2021. Structural challenges in English>Arabic simultaneous interpreting. *Translation & Interpreting* 13/1: 51–70.

Cecot, Michela. 2001. Pauses in simultaneous interpretation: A contrastive analysis of professional interpreters' performance. *The Interpreter's Newsletter* 11: 63–85.

Fu, Rongbo. 2012. Pausing in two-way E-C consecutive interpreting: A contrastive analysis of trainee interpreters' performance. *Foreign Language Teaching and Research* 3: 437–447.

Fu, Rongbo. 2013. Directional effects on disfluencies in consecutive interpreting. *Modern Foreign Languages* 2: 198–205.

Gile, Daniel. 1992. Predictable sentence endings in Japanese and conference interpretation. *The Interpreters' Newsletter* Special Issue 1: 12–23.

Gile, Daniel. 2004. Issues in research into conference interpreting. In Harald Kittel, Armin P. Frank, Norbert Greiner, Theo Hermans, Werner Koller, José Lambert and Fritz Paul eds. *An International Encyclopedia of Translation Studies (Vol. I)*. Berlin: Mouton de Gruyter, 767–779.

Gile, Daniel. 2005. Directionality in conference interpreting: A cognitive view. In Rita Godijns and Michaël Hinderdael eds. *Directionality in Interpreting: The 'Retour' or the Native?* Ghent: Communication and Cognition, 9–26.

Gile, Daniel. 2011. Errors, omissions and infelicities in broadcast interpreting: Preliminary findings from a case study. In Cecilia Alvstad, Adelina Hild and Elisabet Tiselius eds. *Methods and Strategies of Process Research: Integrative Approaches in Translation Studies*. Amsterdam: John Benjamins, 201–218.

Guo, Liangliang. 2011. *An Analysis of the Word Order Pattern in the SI Target Language and its Underlying Reasons in the Language Combination of English and Chinese*. Shanghai: Shanghai International Studies University dissertation.

Jiang, Yue and Xinlei Jiang. 2019. Effect of maximum dependency distance of source text on disfluencies in interpreting. *Foreign Languages Research* 1: 81–88.

Li, Xin. 2018. *The Reconstruction of Modality in Chinese-English Government Press Conference Interpreting: A Corpus-based Study*. Singapore: Springer Nature Singapore.

Mead, Peter. 2000. Control of pauses by trainee interpreters in their A and B languages. *The Interpreter's Newsletter* 9: 199–209.

Moser, Barbara. 1978. Simultaneous interpretation: A hypothetical model and its practical application. In David Gerver and H. Wallace Sinaiko eds. *Language Interpretation and Communication*. New York: Plenum, 353–368.

Pradas Macías, Macarena. 2006. Probing quality criteria in simultaneous interpreting: The role of silent pauses in fluency. *Interpreting* 8/1: 25–43.

Qi, Taoyun. 2019. Pause frequency characteristics of professional interpreter's English-Chinese simultaneous interpreting and their cognitive processes: A case study based on a small-sized corpus with dual modalities. *Foreign Languages and Their Teaching* 5: 135–146, 151.

Ra, Sophia and Jemina Napier. 2013. Community interpreting: Asian language interpreters' perspectives. *Translation & Interpreting* 5/2: 45–61.

Riccardi, Alessandra. 1996. Language-specific strategies in simultaneous interpreting. In Cay Dollerup and Vibeke Appel eds. *Teaching Translation and Interpreting 3: New Horizons*. Amsterdam: John Benjamins, 213–221.

Seeber, Kilian. 2007. *Cognitive Load in Simultaneous Interpreting: A Psychophysiological Approach to Identifying Differences between Syntactically Symmetrical and Asymmetrical Language Structures*. Geneva: University of Geneva dissertation.

Seleskovitch, Danica. 1978. *Interpreting for International Conferences: Problems of Language and Communication*. Washington: Pen and Booth.

Setton, Robin. 1993. Is non-intra-IE interpretation different? European models and Chinese-English realities. *Meta* 38/2: 238–256.

Setton, Robin. 2011. Corpus-based interpretation studies (CIS): Overview and prospects. In Alet Kruger, Kim Wallmach and Jeremy Munday eds. *Corpus-based Translation Studies: Research and Applications*. London: Continuum International, 33–75.

Shen, Mingxia, Qianxi Lv and Junying Liang. 2019. A corpus-driven analysis of uncertainty and uncertainty management in Chinese premier press conference interpreting. *Translation and Interpreting Studies* 14/1: 135–158.

Song, Shuxian, Dechao Li and Jianshe Zha. 2021. An empirical and corpus-based study on the effects of input rate on fluency in simultaneous interpreting. *Foreign Language Research* 3: 103–108.

Tang, Fang. 2020. Features of interpreting learners, self-repairs in English-Chinese consecutive interpreting. *Chinese Translators Journal* 3: 67–77, 188.

Tissi, Benedetta. 2000. Silent pauses and disfluency in simultaneous interpreting: A descriptive analysis. *The Interpreters' Newsletter* 10: 103–127.

Uchiyama, Hiromichi. 1991. Problems caused by word order when interpreting/translating from English into Japanese: The effect of the use of inanimate subjects in English. *Meta* 34/2: 404–413.

Wang, Binhua and Tao Li. 2015. An empirical study of pauses in Chinese-English simultaneous interpreting. *Perspectives* 23/1: 124–142.

Wang, Binhua and Yukui Gu. 2016. An evidence-based exploration into the effect of language-pair specificity in English-Chinese simultaneous interpreting. *Asia Pacific Translation and Intercultural Studies* 3/2: 146–160.

Wang, Binhua and Bing Zou. 2018. Exploring language specificity as a variable in Chinese-English interpreting: A corpus-based investigation. In Mariachiara Russo, Claudio Bendazzoli and Bart Defrancq eds. *Making Way in Corpus-based Interpreting Studies*. Singapore: Springer Nature Singapore, 65–82.

Wang, Jiayi, Defeng Li and Liqing Li. 2019. A study of pauses in EFL learner's interpreting based on PACCEL-S corpus. *Foreign Language Education* 5: 78–83.

Wilss, Wolfram. 1978. Syntactic anticipation in German-English simultaneous interpreting. In David Gerver and H. Wallace Sinaiko eds. *Language Interpretation and Communication*. New York: Plenum, 343–352.

Xu, Haiming. 2010. Pauses in conference consecutive interpreting from English into Chinese: An empirical study. *Foreign Languages Research* 1: 64–71.

Yuan, Shuai and Hongyu Wan. 2019. The impacts of directionality on the fluency of sight translation. *Shanghai Journal of Translators* 1: 30–37.

Zou, Bing and Binhua Wang. 2014. Transcription and annotation of paralinguistic information in interpreting corpora: The status quo, problems and solutions. *Shandong Foreign Language Teaching Journal* 4: 17–23.

Corresponding author
Bing Zou
Guangdong University of Foreign Studies
School of Interpreting and Translation Studies
Center for Translation Studies
North Baiyun Avenue 2
510420 Guangzhou
China
Email: zoubing@gdufs.edu.cn

received: November 2022
accepted: February 2023

RiCL Research in Corpus Linguistics

A Dutch discourse marker in interpreter-mediated police interviewing with drafting: A corpus-based approach to dialogue interpreting

Bart Defrancq – Sofie Verliefde
Ghent University / Belgium

Abstract – This study systematically analyses the use of a Dutch discourse marker (*dus*) by nine interpreters assisting in 12 police interviews. It is an attempt to approach dialogue interpreting with the analytical framework of corpus-based linguistics and a data collection that can stand the comparison with existing corpora of mostly simultaneous interpreting. In terms of frequencies, the results show that interpreters do not seem to divert from general usage patterns for spoken Dutch. However, their use of *dus* is mostly disconnected from the speech they are interpreting. While explicitation seems to be at play in a certain number of cases, the bulk of instances serves interaction coordination purposes. A substantial number of instances with a filler function are also found, where interpreters struggle to understand the source speech or to articulate their interpretation. Finally, some interesting cases of so-called discursive control enforced by *dus* are observed, further confirming the special relationship interpreting holds with drafting of written records during the interview.

Keywords – dialogue interpreting; interpreter-mediated police interviewing; discourse marker; turn management; written record

1. INTRODUCTION

Miriam Shlesinger is traditionally credited with initiating the corpus-based turn in interpreting studies. In Shlesinger (1998), she called upon interpreting scholars to start building corpora of interpretation in order to offer interpreting researchers a collection of naturalistic data and an opportunity to perform large-scale empirical research. Shlesinger's call should not be mistaken for the starting point of the collection of interpreted text. Long before 1998, researchers had been collecting interpretations. Lederer (1980), for instance, reports findings based on a collection of simultaneous interpretations carried out at one conference and completed with experimental data. Using Lederer's data for a study of anticipation in interpreting, van Besien (1999) explicitly refers to the data as 'a corpus'. In the field of dialogue interpreting, considerable amounts

of data recorded during court hearings, police interviews, medical consultations, etc., have been compiled, exploited and, in some cases, even made accessible to the research community. Rarely are those called corpora. So, the first question we need to answer in this paper is: when does a text collection qualify as a corpus?

Text collections of different kinds are called corpora and there is no clear cut-off point beyond which a collection cannot be considered a corpus any longer (McEnery and Wilson 1996). However, corpus linguists usually put forward a number of critical features of corpora. They are expected to be:

1. machine-readable (McEnery and Wilson 1996), facilitating consultation and control of the results;
2. representative of the language, including the language varieties (Biber 1993), allowing researchers to draw generalisations and to replicate research;
3. sizable (Crystal 1995), providing enough data to draw reliable conclusions and to investigate low-frequency features;
4. collected and sampled on the basis of language-external factors (Sinclair 2005), providing naturalistic and independent data for multiple research purposes.

Few collections of interpreting data meet these criteria. For instance, although Bakti and Bóna (2014: 34, our emphasis) claim to perform an analysis of "an experimental *corpus* collected for an earlier study," their research data are clearly not collected based on language-external factors, as they were elicited through a linguistic experiment geared towards the study of a particular linguistic feature. Similarly, Davitti (2013), a study analysing performances of public service interpreters in three meetings, does not qualify as a corpus study, because three meetings can hardly be considered representative of the field of public service interpreting. A study based on 65 interpreter-mediated encounters, such as the one reported in Gavioli and Baraldi (2011), has a better chance of representing at least some of the variety, although the authors themselves specifically deny representativeness in a later study (Baraldi and Gavioli 2012).

If text collections do not count as corpora, as is the case of most of the collections of dialogue interpreting, does it make sense to refer to 'corpus-based dialogue interpreting studies'? The question is an important one. Corpus-based interpreting studies has rapidly evolved over the last ten to 15 years. The availability of corpus materials from several international institutions and the significant improvement of automatic transcription tools removed two of the main obstacles to research that Shlesinger (1998) identified.

Important, though comparatively small-scale, corpora have been collected in various research centres. Quite a few research papers applying corpus-based methods have been published over the years, focusing on lexical and pragmatic properties of interpreted texts, translation universals and cognitive load, though in a small number of language pairs. These studies have mainly focused on conference interpreting while dialogue interpreting has mostly been overlooked. It is remarkable, for instance, that the special issue of the *Interpreters' Newsletter* (issue 22), published in 2017 and specifically devoted to *Corpus-based Dialogue Interpreting Studies*, as its title reads, contains not a single contribution showcasing empirical work based on a corpus (Bendazzoli 2017). In the issue, two contributions are theoretical (Angelelli 2017; Gao and Wang 2017), one is pedagogical illustrating the use of corpora in a training programme for dialogue interpreters (Spinzi 2017), one is an empirical study based on one single naturalistic instance of dialogue interpreting over the telephone (Määttä 2017), and another one is based on data collected during a moot court (Liu and Hale 2017).

The dearth of studies has numerous reasons. The limited accessibility of settings where dialogue interpreting takes place and the severe restrictions imposed in terms of data protection prevent many available collections to reach the status of a fully exploitable corpus. Unsurprisingly, most of the larger data collections stem from court hearings (Hale 2004; Mason 2008; Angermeyer 2015), most of which are open to the public and are recorded in written form by the court itself.

The research agenda put forward in the research on dialogue interpreting is also a factor. Focusing on interactional coordination and interpreters' roles in interpreter-mediated communication, research into dialogue interpreting has rarely promoted investigation of consistent linguistic patterns across dialogue interpretations. The analyses of discourse marker used in Hale (2004) and Mason (2008) are exceptional in that respect. The purpose of this study is to analyse, in the same systematic way, discourse marker use by interpreters in the context of police interpreting.

Section 2 will first review the broader literature on discourse marker use in interpreter-mediated dialogues. Subsequently, in Section 3, we will motivate our choice to focus on the police context and put forward the research questions for our study. While Hale (2004) focused on discourse markers held especially relevant of witness examination in court (*well*, *see*, and *now*), we will focus on a particular discourse marker whose use is critical in police interviews, namely *dus*, which is the Dutch equivalent of

English ‘so’. In the same section, we will show why certain uses of *so/dus* are procedurally important. Data and methods are set out in Section 4, while the results, discussion, and conclusions are presented in Sections 5, 6 and 7, respectively.

2. Discourse markers in interpreting

It is widely acknowledged that interpreters have a propensity to shape the discourse quite differently from the source text, especially with regard to the marking of semantic coherence relations between and inside topical units. In consecutive conference interpreting, for instance, there is evidence that interpreters —novices and professionals alike— use cohesive markers that have no equivalent in the source text. In one particular case of consecutive interpreting at a literary conference in Italy, Mead (2012) finds that the interpreter uses *quindi* (‘so’), even though the consecutive relationship is not explicitly marked at that particular point of the English source text. Interestingly, Mead (2012: 176) attributes the addition by the interpreter to a systemic difference between English and Italian, the latter allegedly preferring explicit marking of cause-effect relations. Similarly, Bastin (2003) observes that interpreting trainees add cohesive devices when interpreting consecutively from English into French, emphasising that the additions improve the perceived quality of their performance.

Similarly, in dialogue interpreting the presence of untriggered cohesive markers is widely attested, as well as the absence of markers at points where they should have been triggered by markers in the source text (Berk-Seligson 1990; Hale 2004; Mason 2008; Gallai 2013, 2017; Blakemore and Gallai 2014). The latter four studies analyse naturalistic data and are directly relevant to a corpus-based approach. Based on an analysis of a substantial corpus of court interpretations, Hale (2004) concludes that interpreters often omit discourse markers that underscore the confrontational stance taken by the speaker. As a result, the illocutionary strength of the speech act performed is altered, which may have an effect on how the addressee will respond. She also speculates that omissions may be attributable to two factors: on the one hand, systemic differences between languages that make it difficult to translate discourse markers; on the other hand, omission may also be the result of the interpreter’s focus on the propositional content of the speaker’s utterance. This focus may divert interpreters’ attention away from items that do not contribute directly to the propositional content.

By contrast, Gallai (2013, 2017) and Blakemore and Gallai (2014) mostly discuss cases where interpreters add discourse markers. Blakemore and Gallai (2014) argue that these additions are signposts of the interpreters' understanding of the speaker's utterances, but stress that hearers are unable to recognise them as such, as they have no access to the speaker's utterances. Additions thus have the effect of strengthening mutuality between speaker and hearer, as the hearer of an interpreted utterance is bound to project contextual assumptions triggered by the discourse marker onto the speaker, and not onto the interpreter (see Delizée and Michaux 2019). In Gallai (2013, 2017) the question of the interpreter's visibility is raised in connection with additions of discourse markers. Finally, Mason (2008) attempts to tie particular tendencies in discourse marker use to gender properties of the interpreters involved. These tendencies are both cognitively and socially determined. Male interpreters tend to omit utterance-initial discourse markers more often than female interpreters, because of limited memory resources. Greater awareness of social hierarchies, in turn, makes men omit more politeness items when an addressee is of lower status. Poorer awareness, in contrast, lets women omit more deferential items. However, women tend to add politeness items to their interpretations more than men, which is interpreted by Mason (2008) as an effect of prioritising group solidarity.

It is sometimes hypothesised that interpreters use untriggered markers to better represent the speaker's 'mental model' (Jacobsen 2002), that is, given the context in which they interpret, they assume that leaving the relationship implicit would not convey the speaker's thoughts in the target language accurately enough (Jacobsen 2002; Blakemore and Gallai 2014).

Finally, as far as simultaneous interpreting is concerned, Shlesinger (1995) concludes from an experimental study that the majority of cohesive shifts, namely, differences in the use of cohesive items from source text to target text, consist of omissions of cohesive markers. In a corpus study based on simultaneous interpretations performed during sessions of the European Parliament, however, Amon (2006) observes one untriggered occurrence of French *donc* 'therefore' in a target text, which he analyses as a placeholder for a substantial omission. In a much larger corpus drawn from the same setting as Amon's data, Defrancq *et al.* (2015) observe that the addition of cohesive markers is quite common across two language combinations (French-English and French-Dutch). For some frequent markers, such as *so* and its Dutch equivalent *dus*, additions

account for 40 per cent to 50 per cent of all occurrences. A general tendency to add discourse markers was found for other language combinations in the European Parliament (see Götz 2020 for English-Hungarian, and Gumul and Bartłomiejczyk 2022 for English-Polish), casting doubt on systemic differences as an explanatory factor. Defrancq *et al.* (2015) also point out that additions cannot always be explained in terms of explicitation, that is, they do not always represent the speaker's assumed 'mental model'. Quite a few occurrences of English *so* and Dutch *dus* are used to cover up large omissions and create an illusion of coherence. Similarly, Defrancq (2016) observes that English-speaking simultaneous European Parliament interpreters sometimes use an untriggered *well* when they seem to feel that their interpretation is inaccurate. For instance, *well* frequently appears to occur in self-repairs performed by several interpreters. Clearly, these are not cases in which interpreters endeavour to reflect the speakers 'mental model'; the items rather reflect interpreters' monitoring of their own speech.

The two explanatory dimensions for the addition of discourse markers that we can draw from the literature are thus systemic differences between source and target language, on the one hand, and the tendency to explicitate either implied speakers' intentions or to express a personal assessment on form (or content) of the interpretation by the interpreters themselves. The modal dimension seems to be less relevant, as additions seem to occur across interpreting modes. However, judging by the number of cases different authors discuss, it seems that the simultaneous mode is more affected by additions than consecutive in dialogue. Our study will seek to challenge these findings on the basis of a larger dataset of dialogue interpreting than is used in most other studies.

3. POLICE INTERPRETING

3.1. Research into police interpreting

The police context is underrepresented in interpreting studies (Gamal 2017). The available empirical research on police interpreting is limited both in empirical and in contextual scope. Most of the studies are based on no more than five police interviews: Krouglov (1999), four interviews; Komter (2005), one interview; Gallai (2013), five interviews; Nakane (2014), four interviews; Kredens (2017), one interview; Monteoliva-García (2017), two interviews; Defrancq and Verliefde (2018), one interview; and Tipton (2021), two interviews. Verliefde and Defrancq (2022) draw on ten interviews. Only

Russell (2001) and Filipović (2022) stand out with 28 and 100 police interviews, respectively. However, their interviews were very fragmentarily transcribed.

Studies focus almost exclusively on police interpreting in legal systems that belong to common law: Australia in Nakane (2014); the United Kingdom in Krouglov (1999), Russell (2001), Gallai (2013, 2017), Blakemore and Gallai (2014), Kredens (2017), Monteoliva-García (2017) and Tipton (2021); and the United Kingdom and the United States in Filipović (2022). Only Komter (2005), Defrancq and Verliefde (2018), and Verliefde and Defrancq (2022) deal with police interpreting in continental Europe, the Netherlands and Belgium, respectively.

The continental inquisitorial legal system is particularly interesting as police officers are required to conduct oral police interviews and to (simultaneously) draft a written record of those interviews. According to Komter (2006), the drafting phase has turn-like status in the interaction. In such a context, interpreter-mediated interviews include not only the spoken interaction of three participants, but also the entextualisation process (Park and Bucholtz 2009). This process is a polyphonic representation in itself: police officers record interviewees' statements as rendered by interpreters. Interpreters are seen to actively engage in a variety of ways with this entextualisation process (Defrancq and Verliefde 2018; Verliefde and Defrancq 2022).

This is where the interest of discourse marker use in police interpreting lies: multiple discourses intersect during the interview, which are meant to be conflated into one single written discourse at the end. In an inquisitorial legal system such as the Belgian one, interviewees have the right to request that their statements be taken down verbatim (Smets and Ponsaers 2011). However, they rarely are, as police officers tend to enhance the logical and chronological coherence of the interviewees' accounts, while focusing on cause-effect relationships to establish interviewees' involvement in criminal offenses (Smets and Ponsaers 2011). That sort of enhancement is partly achieved through discourse markers. Interpreters, in turn, are known to assist with the drafting process, adapting answers to the question format, pausing, spelling names (Pöchhacker and Kolb 2009), and upgrading the interviewee's register (Defrancq and Verliefde 2018). The latter pattern is likely to affect discourse features of the interpretation, including the use of discourse markers. The role of markers of consequence is paramount in this respect as those are instrumental to making cause-effect relationships explicit.

There is another reason why the European continental police context is especially interesting for the analysis of markers of consequence. In English, the most frequent of these markers, *so*, is reported to be used as "an agent of discursive control" (Ainsworth 2018: 36), meaning that a reformulation or summary introduced by *so* is difficult to challenge for the interlocutor. Interestingly, police officers are reported to use these kinds of reformulations frequently before taking down interviewees' statements (Komter 2022). For instance, in Excerpt (1), taken from Komter (2022),[1] the first quoted utterance starts with *so* and is followed by the information that is about to be typed up (albeit from a different deictic framework).[2]

Excerpt (1)

1. P: So yesterday you went to the market with your children.
2. S: Yes.
3. P: ((types, 6 s)
Yesterday,
4. P: To the market, then we're talking about Waterlooplein I assume.
5. S: What do you say, yes.
6. P: Yes.
7. ((types, 17 s:))
I went to the Waterlooplein, together with my children.
8. P: Uh (4) have you uh been to the stalls

Pre-drafting reformulations invite interviewees to agree with the wording the police officer proposes to use in the written record. They underscore that the police officer is in charge of the written discourse that results from the recent exchange. Given the interpreters' role in the drafting process, interpretations might also present evidence of discursive control.

The data used for this paper was collected in an area where only Dutch can be used in legal proceedings. We will therefore first review the literature on the most common Dutch marker of consequence: *dus*.

3.2. Dus

Compared to its English's and French counterparts (*so* and *donc*, respectively), *dus* has attracted little research. As our purpose is empirical, we will focus in this section on the

[1] https://www.frontiersin.org/articles/10.3389/fcomm.2022.797145/full
[2] In fact, Komter's (2022) examples are translations of transcriptions based on interviews conducted in Dutch. The agent of discursive control is *dus*, rather than English *so*.

different uses of *dus* which are described in the literature, giving special attention to the literature on spoken Dutch.

For spoken registers, the absolute frequency of *dus* reported in Oostdijk (2000) on a sample of 615,000 tokens of the *Corpus Gesproken Nederlands* (CGN)[3] is 3,895, or a relative frequency of 6.3 occurrences per 1,000 words. Degand (2011) reports a relative frequency of 4.3 per 1,000 words in a larger sample of CGN (1.7 million tokens), consisting exclusively of spoken data from the Netherlands. Finally, Degand and van Bergen (2018) report a frequency of 7.2 occurrences per 1,000 words in a CGN subcorpus comprising only face-to-face interactions. Face-to-face interactions are directly relevant to the interpreting data to be analysed, as those were collected in dialogue settings. Higher frequencies in face-to-face interactions are likely due to the floor management functions for which *dus* is recruited. It is to be expected that interpreting data collected in dialogue settings show similarly high frequencies of *dus*. As interpreters are known to take charge of turn coordination in dialogue interpreting (Wadensjö 1998), one may wonder whether interpreters use *dus* to render turn management organisation by the participant or rather their own turn management.

Dus is traditionally described as a connective (Pander Maat and Degand 2001; Stukker *et al.* 2009) or a discourse marker (Evers-Vermeul 2010; Degand 2011; Buysse 2017; Degand and van Bergen 2018). Most authors attribute functions to *dus* in three widely accepted domains of discourse: the ideational, the interpersonal, and the textual domain (Halliday 1985). In the ideational domain, *dus* connects states of affairs that are in a causal relationship; its use foregrounds subjective features of that relationship (Pander Maat and Degand 2001; Stukker *et al.* 2009). In the interpersonal domain, *dus* may signal inferences connecting illocutionary meanings with locutionary meanings (Degand 2001) and turn management functions, such as turn uptake and turn yielding (Degand and van Bergen 2018). In the textual domain, *dus* enables reactivation of previously uttered information (Evers-Vermeul 2010), including concluding, rephrasing and repetition. In this particular function, *dus* may be used as an agent of discursive control in the sense of Ainsworth (2018). As pointed out above, Komter (2022) quotes several examples of *dus* introducing pre-drafting reformulations by police officers, which typically ensure discursive control. Finally, Defrancq *et al.* (2015) show that simultaneous interpreters frequently use *dus* to create an illusion of cohesion during and after a period

[3] https://taalmaterialen.ivdnt.org/download/tstc-corpus-gesproken-nederlands

of inadequate rendition. It often occurs in those cases in combination with hesitation markers, such as *uh*, and usually has no equivalent in the source text.

Unfortunately, there are few frequency data on the individual functions of *dus*. Buysse (2017) provides a functional breakdown of occurrences in both source texts and translations showing that inferential uses and concluding uses are most frequent and make up slightly more than half of the occurrences.

3.3. Research questions

There seem to be significant research gaps in the field and in particular as regards dialogue interpreting in a police context. There is evidence that simultaneous interpreters add substantial numbers of discourse markers to their interpretations. There is also some evidence to that effect in dialogue interpreting but, due to the small collections of dialogue data, it is unclear how strong this tendency is. For lack of sufficient instances, it is unknown which factors could explain additions in dialogue interpreting. A thorough functional analysis is therefore needed to find out to which functional categories additions mostly belong. Therefore, the research questions of this study are as follows:

1. How frequent is *dus* and how frequent are untriggered (that is, added) instances of *dus* in dialogue police interpreting?
2. What are the functions of *dus* used by dialogue interpreters in police contexts?

It is important to mention that this study will only focus on the use of *dus* in Dutch interpretations. Discourse marker use in other languages is not taken into account. As the corpus contains seven languages other than Dutch, including them would require a detailed review of discourse markers in all the languages, which is beyond the scope of this study.

4. DATA AND METHODOLOGY

The data used in this study are drawn from a collection of 12 police interviews conducted in Belgium between 2014 and 2019. According to Belgian law, these interviews are conducted in Dutch and, if the interviewee has not got sufficient knowledge of the language, interpreted by a sworn interpreter. Recordings of these interviews were authorised under the Court of Ghent's Prosecutor General's authorisation and stored with

password protection on local servers at Ghent University. Besides the recordings, the written records of the interviews were obtained and stored. The recordings were transcribed using the Jefferson (2004) conventions and pseudonymised. Transcriptions were made by different legal interpreters or, when those were not available, by people with a language degree in the non-institutional language. Turns in the non-institutional language were back-translated to Dutch. Table 1 provides an overview of the interviews and their main features.

Interview	**Language**	**Interpreter**		**Topic**	**Duration**	**Recording**
		Dutch	Gender			
1	French	A	F	Threats and assault with a knife	2h 55m	Audio
2	English	A	F	Sham relationship	3h 00m	Video
3	Turkish	B	M	Sham marriage	1h 45m	Video
4	Romanian	B	F	Human trafficking and forced prostitution	4h 15m	Audio
5	Arabic	B	M	Sham marriage	2h 45m	Audio + video
6	Arabic	B	M	Sham marriage	1h 25m	Audio + video
7	Pashto	B	M	Drug trafficking	1h 55m	Audio
8	Pashto	B	M	Possession of prohibited weapon	0h 30m	Audio
9	Romanian	B	F	Human trafficking and forced prostitution	3h 40m	Audio + video
10	Romanian	B	F	Human trafficking and forced prostitution	4h 15m	Audio + video
11	Greek	A	M	Sham relationship	2h 30m	Video
12	Greek	A	M	Sham relationship	2h 20m	Video
Total					**31h 15m**	

Table 1: Overview and main features of the interpreter-mediated interviews

In all, the collection contains approximately 31 hours of interpreter-mediated police interviewing involving nine different interpreters (the same interpreters are active in Interviews 5 and 6, 9 and 10, and 11 and 12). Half of the recorded interviews deal with sham relationship procedures. Police officers are able to plan these interviews in advance, which makes it easier for them to coordinate with the researchers in these cases.

The nine interpreters involved are all sworn interpreters according to the pre-2016 requirements. In a nutshell, this means that they provided proof of their knowledge of two languages, including Dutch, and had no criminal record prior to the oath they were invited to take at the court to become sworn interpreters. The Belgian law on sworn interpreters

and translators was overhauled in 2014 as a result of the implementation of Directive 64/2010/EU.[4] It now imposes a legal training programme of, at least, 32 hours. The interpreters' level of experience was not queried at the time of the interviews they interpreted.

Interview	Language	Tokens	Interpreter tokens
1	French	32,000	15,000
2	English	13,000	7,000
3	Turkish	11,000	4,000
4	Romanian	42,000	18,000
5	Arabic	16,000	8,000
6	Arabic	9,000	5,000
7	Pashto	12,000	5,000
8	Pashto	6,000	2,000
9	Romanian	39,000	19,000
10	Romanian	30,000	14,000
11	Greek	21,000	10,000
12	Greek	14,000	7,000
Total		**245,000**	**114,000**

Table 2: Numbers of tokens in the different sub-corpora

As shown in Table 2, the corpus contains roughly 245,000 tokens. Slightly less than half of those tokens (114,000) are attributable to the interpreters. Figures vary considerably across encounters depending on the features of the interrogation. In half of the encounters the written record drafted during the interrogation is sight-translated by the interpreter at the very end of the interview. This accounts for higher shares of interpreting in the total token count of the encounter. The interviews in Pashto were partly conducted in the regionally dominant language, as the Pashto suspect had some knowledge of it, which accounts for the lower share of interpreter tokens. Most instances of interpretation are carried out in consecutive mode. However, in Belgian police interviews it is not uncommon to see interpreters use the simultaneous mode and even switch multiple times from one mode to another. Interviews 1, 9 and 10 contain stretches of simultaneous interpreting. Simultaneous interpreting is mainly used when the interviewee is speaking.

[4] Law of 10 April 2014 modifying several provisions regarding the creation of a national register for legal experts and with a view to create a national register of sworn translators, interpreters, and translators-interpreters. *Belgian Official Journal* of 19 December 2014, p. 104479.

By most standards, the collection is small; in the area of interpreting research, however, its size is fairly average. The problems and usefulness of interpreting research based on nano-corpora, such as the *Interpreter-mediated Police Interviewing with Drafting* corpus (IMPID; Verliefde 2022), have been reviewed in the literature (Defrancq and Collard 2019). The data are not publicly available as they contain sensitive personal information that needs to be protected. It is therefore debatable whether the collection represents a real corpus of interpreting. We will nevertheless apply regular corpus-based methods to query the corpus and extract both quantitative and qualitative data.

All instances of *dus* were extracted using *AntConc* 3.4.4 (Anthony 2014) and placed in a wide context window. A considerable number of utterance-initial occurrences was expected, compelling us to take into account a substantial piece of previous context to be able to identify the function of the connective.

In assigning functions, we privileged a manual close reading approach, using the different semantic and interactional functions listed previously as a frame of reference. In doing that, we have applied the following annotation principles:

1. only assign a consequential function if the relationship between two phrasal or clause units can be interpreted as a cause-effect relationship between states of affairs;
2. only assign an inferential function if the relationship between two phrasal or clause units can be interpreted as an inferential relationship;
3. only assign a rephrasing function if the unit or clause following *dus* contains information already communicated in the same language. The latter requirement is important in the context of interpreting, as interpreting itself is inherently an act of rephrasing which includes self-repairs and conclusive statements at the end of turns that summarise the content of the turn;
4. only assign a turn-management function if turns are effectively transferred;
5. only assign a filler function if *dus* occurs in combination with hesitations, pauses and substantial omissions, while its function cannot be accounted for by any of the other functions;
6. label all cases that could not be assigned to one of the previous categories to a category ‘unassigned’.

The annotation principles were applied in the described order. This implies that ambiguous cases are attributed to the higher category. For instance, if an instance of *dus*

occurs at the beginning of a successfully transferred turn in combination with hesitations, it is analysed as a turn-taking device rather than as a filler.

5. RESULTS

5.1. Quantitative analysis

Table 3 shows the absolute and normalised frequencies of *dus* in the interpreters' turns in each interview and the number and share of occurrences that are elicited, that is, that can be considered to be triggered by a discourse marker in the source speech.

Interview	**#**	**/1,000w**	**Elicited**	**Non elicited**	**Percentage elicited**
1	29	4.38	3	26	10.3
2	9	3.10	0	9	0.0
3	14	8.24	1	13	7.7
4	79	10.97	11	68	13.9
5	1	0.03	0	1	0.0
6	0	0	N/A	N/A	N/A
7	5	2.50	0	5	0.0
8	2	2.50	0	2	0.0
9	25	3.47	2	23	0.8
10	54	9.64	3	51	5.6
11	32	8.02	3	29	9.4
12	15	5.17	6	9	40.0
Total	**265**	**5.88**	**29**	**236**	**10.9**

Table 3: Frequencies of *dus* per interpretation

The overall relative frequency of *dus* in interpretations seems to be in line with the frequency reported in Degand and van Bergen (2018) for monolingual face-to-face interactions, which was 7.2 per 1,000 words. Compared with simultaneous interpreting data from the European Parliament, the dialogue interpreters in our sample appear to use *dus* slightly more. In the data presented by Defrancq *et al.* (2015), the frequency of *dus* in simultaneous interpretation performed in the European Parliament was 3.8 occurrences per 1,000 words (98 occurrences in a corpus of roughly 26,000 words).[5]

Variation across interpreters is high. There seems to be no plausible explanation for the variation other than individual usages. There is no observable relation with A-

[5] Defrancq *et al.* (2015) aggregate data for several discourse markers. The data given here report on the subset of occurrences of *dus*.

interpretation or B-interpretation into Dutch, as both highest and lowest frequencies are found in the group of B-interpreters. Usage does not seem to be gender-related either.

What all interpretations have in common is a low elicitation rate. Overall, only one in nine occurrences can be ascribed to the presence of a marker in the source speech. One fifth of the elicited instances occur in one single interpretation (Interview 12). These figures are all the more striking as the elicitation rate observed in the simultaneous data used by Defrancq *et al. (*2015) was 57.1 per cent (56 out of 98 cases). Dialogue interpreting appears to incentivise interpreters more to add the connective *dus* to their interpretations. In this regard, it should be noted that a substantial number of untriggered instances of *dus* occur in non-renditions (Wadensjö 1998), that is, in interpreter utterances that cannot be analysed as interpretations. In non-renditions, interpreters address one of the primary participants directly. In our corpus, 43 examples of this type were found. Excerpt (2), drawn from Interview 1, illustrates such a case (S = interviewee and I = interpreter).[6]

Excerpt 2

1. S <EN HOEVEEL KEER (.) PENDANT DEUX MOIS> (.) COMBIEN DE
   [in Dutch] how many times [in French] during two months how many
2. S FOIS IL EST VENU CHEZ VOUS <person 1>
   times did he come here <person 1>
3. I hoeveel keer dat <person 1> hier in de afgelopen twee maanden bij jullie
   how many times <person 1> has been here with you in the last two months
4. I geweest is↓ **dus** dat zou hij wel eens willen weten↓
   so that's what he would like to know

In line 4, the interpreter adds an utterance referring to the interviewee in third person: *hij* ('he'), while explicitating the illocutionary force of the interviewee's turn. An instance of *dus* is used to introduce the addition.

Occurrences in non-renditions account for almost a fifth of the non-elicited cases. A functional analysis was carried out to find out in what circumstances interpreters add the other 80 per cent of non-elicited cases.

[6] In Interview 1, the interviewee has some knowledge of Dutch and uses it occasionally throughout the interview.

5.2. Functional analysis

Nearly all instances of *dus* can be straightforwardly categorised using the annotation criteria put forward in Section 3. Table 4 presents the breakdown of the observed cases.

Function	**Total**	**Non elicited**	**Non renditon**	**Elicited**	**Percentage elicited**
Consequence	26	20	7	6	23.1
Inference	27	22	6	5	18.5
Rephrasing	52	49	9	3	5.8
Turn taking	98	87	11	11	11.2
Turn yielding	13	11	3	2	15.4
Filler	42	42	3	0	0.0
Unassigned	7	5	4	2	28.6
Total	**265**	**222**	**43**	29	**10.9**

Table 4: Functions of *dus* in interpreting

All categories are well represented in the corpus. Turn management functions (turn taking and turn yielding) prevail, totalling almost 40 per cent of the cases. Rephrasing and filler functions jointly account for a third of the occurrences. Consequential and inferential *dus* amounts to almost a fifth. In total, seven cases could not straightforwardly be assigned to one of the categories.

The share of elicited instances is highest in the consequential uses. This is expected, as adding a marker in the ideational plane contributes to the meaning of the utterances and may distort the participants' message. In other uses, the risk of distortion is smaller, as well as the resulting ethical pressure to avoid additions.

In what follows, we will discuss a number of illustrative cases of untriggered uses, placing them in a wider context of interpreter strategies.

5.2.1. Explicitation of a cause-effect relationship

Excerpt (3), drawn from Interview 11, shows a consequential use of *dus* in line 14, which has no equivalent in the Greek source. The interviewee refers to the flight tickets the couple might have used to return to Greece, which would be evidence of their initial intention not to stay in Belgium. However, as he made the journey back in a lorry, he never used the tickets.

Excerpt (3)

1. S μπορέσουμε κάνουμε κάτι μείναμε αν δεν μπορέσουμε είχαμε
   if we can find anything we'll stay, if we cannot find anything
2. S κόψει τα εισιτήρια επιστροφής και θα φυ=θα ξανά γυρνάγαμε
   there is still the return ticket and then we'll go back again
3. S πάλι πίσω[(.hhh) μετά η (person17) έμεινε εγώ έφυγα πιο νωρίς
   together later (person 17) stayed and I left earlier
4. I [χμ
   (uhum)
5. S με:: φορτηγό και τα εισιτήρια πήγανε χαμένα δεν ξέρεις τα
   by lorry and the tickets got lost I mean we never
6. S χρησιμοποιήσαμε ποτέ τα επιστροφής
   used the tickets for the return flight

[7-9]

10. I met 't idee van te komen en te zien:: (.) of:: we:: (.) we hier konden
    with the idea to come and to see whether we would be able
11. I blijven of nie (.hh) en: als het zou tegenslagen dan hadden we ons
    to stay here or not and if we failed then we'd still have our
12. I euhm (.) terugticket onze terugvlucht al (.) ma dan uiteindelijk is
    return ticket our flight back but in the end
13. I (person17) gebleven (.) en ik ben me een vrachtwagen teruggegaan
    (person 17) stayed and I went back by lorry
14. I **dus** de (.h) de retourtickets die::: die zijn gewoon verloren gegaan
    so the return tickets they they basically got lost

In Greek the clauses in line 5 are linked up with the coordinate conjunction και ('and'). The interpreter, however, chooses to foreground the cause-effect relationship between the journey in the lorry and the failure to use the flight tickets. The use of consequential *dus* may be regarded as a typical explicitation of a clausal relationship. Explicitation might be an attempt to downplay the loss of evidence, as is further evidenced by the addition of a trivialising *gewoon* ('just/basically').

5.2.2. Rephrasing, repeating, and marking the most suitable segment as an answer

Rephrasing is obviously nearly always the result of an addition by the interpreter: interpreters usually do not copy rephrasings or repetitions by the primary participants. Excerpt 4, drawn from Interview 4, shows how the interpreter rephrases a previous segment of the interpretation, while no rephrasing takes place in the source utterance. The rephrasing is signalled by *dus*.

Excerpt (4)

1. S deci dacă mă lasă să termin povestea de la ce-am ajuns cu
   so if he allows me to finish my story about what happened with the fight
2. S ↑cearta pot să-i spun eu când am ajuns aici el mi-a arătat site-ul lui
   then I can tell him that when I arrived here he showed me his website
3. I >hij zeg< als je [m::ij (2) het verh:aal
4. S [dar (.) >după aceia am aflat mai multe și de aia am și plecat<
   but after that there was more that I found out and I left
5. I dat eu::h eh van begin tot einde **dus** tot de ruzie (.) dan gaat u m:e=euh
   that uh uh from beginning to end so until the fight then you'll
6. I misschien begrijpen van waar dat hij geld had↑
   perhaps understand where he got the money

Often these instances of *dus* occur when interpreters summarise the content of a long turn by the interviewee, emphasising the information unit most likely to be a suitable answer by repeating it near the end of their turn and by marking it with *dus*. In Excerpt 5, from the same interview, questions are asked about the whereabouts of a particular person suspected of being the ringleader of a human trafficking network. The interviewee starts describing a bar where members of the ring met. He is interrupted by the interpreter in line 7, who starts rendering his turn. She concludes by repeating a clause from the beginning of her turn, introducing it with *dus*.

Excerpt (5)

1. P [(xxxx) in die café he↑ (xxxxxx)
   xxxx in that bar right xxxxx
2. I În acea cafenea
   in that bar
3. S >În acea cafenea< (1) era el (person7) (3) (person20) (2) (person22)
   in that bar he was there (person7) (person20) (person22)
4. I °ja°
   yes
5. P ja
   yes
6. S Și mai erau:: dar acum nu știu dacă erau cu el era mai
   and there were more people but I don't recall if they were with hem there
7. S multă lume jucau un biliard înăuntru [dar nu știu::: eu am stat afară (xxx)
   were many more they played pool inside but I don't know I was outside
8. I [er wa- (.) er waren nog mensen **nu weet ik nie** of die
   there we- were more people now I don't recall if
9. I andere ware::n (.) euh samen met hem of nie (.) 't was een eu::h
   those others were with him or not it was a uh
10. I biljarttafel ze waren eu:h (a.) aan 't spelen en ik zat buiten stond
    pool table they were uh playing and I was outside was
11. I buiten te roken↓ °**dus 'k weet het nie**
    outside smoking so I don't know

Excerpt (6), drawn from Interview 12, shows a case where the interpreter repeats a segment from one of his previous turns, following turns by both participants. The segment is the most suitable element to form a question-answer pair with the question asked by the police officer in line 1 whether the couple is considering getting engaged. The use of *dus* seems to single out that particular element to fit the adjacency pair.

Excerpt (6)

1. P oe=iz=eu::h is er al sprake van een verloving↑
   How is uh can we say that you are already engaged
2. I (.hhh) αραββώνας υπάρχει κιόλας↑ έχετε αραββωνιαστεί↑
   are you already engaged are you engaged
3. S αραββώνας↑ τι εννοείτε↑
   engaged what do you mean
4. I ε::: αν έχετε αραββωνιαστεί (.) δηλαδή <επίσημα> [είπατε ότι
   uh if you're already engaged so actually if you already officially registred
5. S [°να° (3) επίσημα↑ (2) όχι
   officially no
6. I θα παντρευτείτε
   that you'd marry
7. I neen
   no
8. S δηλαδή να έρθουν οι γονείς μου και οι γονείς του και να:::↓
   you mean my parents and his parents come and
9. I ge bedoelt da::: (.) zijn ouders en mijn ouders samenzitten en
   you mean my parents and his parents meeting and
10. I bespreken da we kunnen trouwen **dat is** [(.) **nog nie gebeurd**
    discussing marriage that has not been the case yet
11. P [ja: da=in iedere cultuur
    yeah in every culture
12. P eu::h gaat het er wat anders aan toe
    uh people go about it somewhat differently
13. I ανάλογα με τον πολιτισμό λέει μπορεί να είναι διαφορετικό
    depending on the culture he says it may differ
14. I αλλά::: δεν έχετε::: [(.) βάλει
    but you haven't yet
15. S [όχι γιατί είναι λίγο δύσκολο ((laughs)) (xxx)
    [no cos it is a bit complicated
16. S αλλά αν γίνει κάτι (.) θα γίνει και θα ρθουν όλοι μαζί (.) έτσι
    but when we get to that point it is obvious that we'll all meet
17. I sowieso als als er iets gebeurt als we zouden trouwen of zo (.hh)
    in any case if something happens if we'd want to get married or so .hh
18. I dan:: m=moeten de ouders eu:h allemaal samenkomen **dus**
    then our parents uh should all meet so
19. I **dat is nog nie gebeurd**
    that has not been the case yet

Excerpt (7), drawn from Interview 11, illustrates a similar case. The most suitable answer to the police officer's question is the reference to *Netflix* in the interviewee's statement.

That segment is again introduced by means of *dus*. Interestingly, the police officer had already identified the reference to *Netflix* directly from the source in line 5. In other words, the interpreter singles out the segment which is already available in the common ground between the police officer and himself.

Excerpt (7)

1. P ok e=een favoriet euh tv-programma van eu::h hem en haar
   okay a favourite tv programme of uh hers and his
2. I ε:: κάποιος ε::: <αγαπημένο> πρόγραμμα στη τηλεόραση (.)
   does either of you have a favourite tv programme
3. I δικό σου της (person17)
   you or (person17)
4. S εδώ δεν ξέρω (.) εδώ κάποια στιγμή βλέπαμε νετφλιξ δεν έχει κάτι
   well I don't know there are times that we watch Netflix but
   there isn't much
5. P [**Netflix**↑
6. S [ναι ναι (.) ναι κάποια στιγμή βλέπαμε το νετφλιξ εδώ γιατί όλες
   yeah yeah at some we watched Netflix cos
7. S τις άλλες >τα κανάλια δεν τα καταλαβαίνουμε<
   we couln't understand any of the other channels
8. I [hm ge-
9. S [(.hhh) αλλά συνήθως >πιο πολύ τηλεόραση που έχουμε στο
   .hhh but more often than not we don't switch the tv on
11. S σπίτι είναι μονίμως κλειστή< (.) περισσότερες φορές γιατί (xxx)
    very often cos for one you've got to reach out for it
12. S ένα να τεντώσεις το χέρι και δεν (.) πολύ ασχολιούμαστε με την
    and we don't want to spend time watching
13. S τηλεόραση
    television
14. I το τελευταίο ↑στο σπίτι↑
    the last one home
15. S τς πιο πολύ ασχολ=την τηλεόραση είναι κλειστή [(.) δεν (.) μόνο
    well we more often d the tv is not on only at night when we
16. I [ναι οκει
    [yeah ok
17. S τα βράδια άμα αν έχουμε όρεξη >θα δούμε καμιά ταινία στο
    feel like it then we watch some movie on
18. S νετφλιξ< (.hh) και τίποτα άλλο
    Netflix that's it no more than that
19. I echt tv kijken doen we nie omda::: ja alle kanalen die hier gegeven
    we don't actually watch tv cos the channels that are available here
20. I worden daar begrijpen we niks van (.hh) **dus** het enige da we af en
    we cannot understand any of it .hh so the only thing we occasionally
21. I toe::: euh zien is **een film op Netflix** (.hh) ma m:eestal staat de tv
    uh watch is a movie on Netflix .hh but mostly the tv is switched off
22. I eigenlijk uit
    actually

Cases like (4), (5) and (6) are most frequent in the Romanian and Greek interviews and seem to be characteristic of interpreters who grant primary participants long turns in the conversation and are also given the opportunity by one of the participants to engage in asides with the other primary participant. Most interestingly, they seem to prompt police officers to focus on the item singled out while drafting the written record.

5.2.3. Fillers

As far as the use of fillers is concerned, it is noticeable that *dus* is especially frequent in stretches of simultaneous interpreting, as illustrated in Excerpt (8), drawn from Interview 10. Unsurprisingly, these instances appear to be due to comprehension or production issues. Witness the many filled pauses that co-occur with *dus*.

Excerpt (8)

1.	S	eu am avut ocazia să-l cunosc de abia în 2018 (.) I had the opportunity to get to know him only in 2018
2.	I	ik ik heb hem gekend [in 2018 I I got to know him in 2018
3.	S	[prima dată pe (person5) (2) si stând la mine în oraș↑ (xxx) the first time (person5) while he was staying in my town
4.	I	[de eerste keer op=(person5) (.) en (person5) woont **eu::h dus eu::h** the first time on (person5) and (person5) lives u:h so u:h
5.	I	in dezelfde stad me mij al twintig jaar↓ in the same town as I for twenty years

In line 2 the interpreter starts rendering the interviewee's turn in line 1, but she is interrupted halfway through the first sentence. She chooses not to yield the turn, interpreting simultaneously until the point where background noises make the recording inaudible (indicated with xxx in line 3). These noises seem to distract her causing her to hesitate (line 4) and to use *dus*. These uses come very close to the ones reported in Defrancq *et al.* (2015), where *dus* is shown to fill up gaps in the interpretation.

5.2.4. Turn management

Most of the observed instances of *dus* (104 out of 265) are turn management devices. In slightly less than 90 per cent of the cases, interpreters signal their own turn management. The remaining cases are ambiguous. In Excerpt (9), for instance, *dus* in line 2 may either

signal the interpreter's own turn management or render the participant's turn taking device *bun* 'good', 'right' in line 1.

Excerpt 9

1. S **bun** întrebarea este în felul următor [(.) cum s-a petrecut am înțeles↓
   good the question is the following: how did it happen I got that
2. I [**dus** de vraag (1) is
   so the question is
3. S [ce exact să îmi spuna exact (.) ce vor să știe cum s-a petrecut ce
   what exactly they tell me what they want to know how did what happen
4. I [hoe de zaak in elkaar zit (1) <WAT precies> (1)
   what the case looks like what exactly
5. I over wat wil je spreken
   what you want to talk about

Excerpt (9) also illustrates another feature of *dus* as a turn taking device: frequently it introduces overlapping speech and, occasionally, a stretch of simultaneous interpreting. The presence of *dus* seems to indicate that the interpreter's intent may not be to interpret simultaneously, but rather to force a turn transition from the interviewee to herself. The interviewee's failure to yield the turn leads to the overlapping speech.

5.2.5. Unassigned cases

Five of the seven unassigned instances seem to share a distinctive feature, namely a sudden change of perspective or even rendition mode. This is illustrated in Excerpt (10), taken from Interview 1. In line 2 the interpreter first reports in third person what the interviewee said in line 1. Then she produces a short sigh (.hhh) and starts interpreting in first person what the interviewee had previously said. The transition between the renditions modes is marked with an instance of *dus*.

Excerpt (10)

1. S c'est assassiner comme tuer quelqu'un ça↓
   that's to assassinate like killing someone
2. I voor hem is't een is dat een moord (.hhh) **dus** ik wil duidelijk maken dat
   to him that's murder so I want to clarify that
3. I <het gaat (.) om ongeboren kinderen (.hhh) die geaborteerd worden>
   this is about unborn children that are being aborted

The use of *dus* in (10) is likely associated with the written record that is being drafted: it signals the start of the segment that is to be recorded by the police officer in the appropriate first-person style used in written records. This is further evidenced by the

reporting phrase *duidelijk maken* ('clarify'), which is typical of the discourse of the written register and was never used by the interviewee. Examples such as (10) seem to be connected to the use of *dus* as a turn taking device: instead of signalling the start of her turn, the interpreter seems to signal the start of the segment to be recorded. Figure 1, taken from the written record of the interview, confirms that the police officer started taking down the interpreter's turn from the moment she uttered *dus* onward. There is no reference to *murder* in the written record.[7]

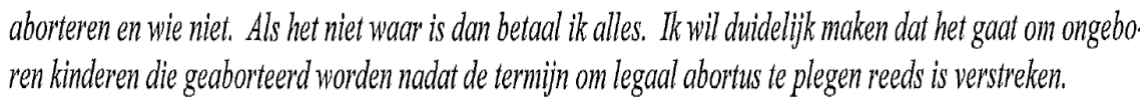

aborteren en wie niet. Als het niet waar is dan betaal ik alles. Ik wil duidelijk maken dat het gaat om ongeboren kinderen die geaborteerd worden nadat de termijn om legaal abortus te plegen reeds is verstreken.

Figure 1: Extract from the written record of Interview 1

6. DISCUSSION

First, it is important to underline that the overall frequency of *dus* in interpreters' utterances is comparable to its frequency in dialogic spoken Dutch registers. However, it is also quite clear that the discourse marker's use is mostly disconnected from the primary participants' discourse. This strongly suggests that interpreters re-shape the original discourse to a significant extent in dialogue contexts. Of course only one discourse marker (*dus*) was analysed and it was used by only nine interpreters. Therefore, the results definitely need to be confirmed for a larger set of markers and a larger population of interpreters. With regard to the latter aspect, the data clearly point to considerable variation across interpreters.

The interaction format seems to induce interpreters to use instances of *dus* that are unrelated to the primary participants' discourse. An essential part of the dialogue interpreters' task is to coordinate talk (Wadensjö 1998), which is reflected in a large number of instances where *dus* is used to manage turn taking. This raises an interesting question regarding coordination in interviews where interpreters do not use *dus* or only use it very parsimoniously, namely, Interviews 5 to 8. There are indications that coordination is indeed weaker in Interviews 7 and 8. These are conducted in Dutch and Pashto with the same participants and the same interpreter. At several points the interpreter's competence is called into question and he is sidelined for a significant part

[7] Translation: 'abort and who not. If it is not true, I'll pay for everything. I want to clarify that this is about unborn children that are being aborted after the deadline for a legal abortion is passed'.

of the interview, as the interviewee has knowledge of Dutch. This clearly makes him a secondary participant in the interaction. As for Interviews 5 and 6, interpreted by the same interpreter with different interviewees, turn management lies firmly in the hands of the interpreter. However, he mostly relies on *ok* as a turn management device. Finally, interpreters who resort to simultaneous interpreting in the course of the interview are the ones who use *dus* most frequently. At first sight, this seems surprising, because simultaneous interpreting requires less coordination. However, simultaneous interpreting could be a side-effect rather than a strategy: a participant who is unwilling to yield the floor can force an interpreter into simultaneous interpreting by simply ignoring the signal for turn transition.

A second point that needs to be raised is how interpreters seem to use *dus* in relation to the written record. Several examples show interpreters singling out bits of information by repeating them and signposting them by means of *dus*. Often these segments constitute the most suitable answer to a question previously asked by the police officer. We hypothesised that the role of *dus* is to draw the police officer's attention to the signposted segment in order for it to be taken down in the written record. This use comes close to the discursive control function discussed in Section 3.1: *dus* introduces a reformulation of the previous discourse in a version that is difficult to challenge for the police officer as it suits the required features of the written record. Excerpt (10) is particularly illustrative of this as *dus* is used at the transition point between two very different representations of the previous discourse: one meant for the police officer (third person) and one specifically designed for the written record (first person and register update). It remains to be seen whether this pattern can also be found in other contexts of dialogue interpreting with drafting, but it certainly adds to other research showing that interpreting for the written record prompts particular discursive strategies in interpreters (Defrancq and Verliefde 2018; Verliefde and Defrancq 2022).

7. Conclusions

The motivation for this study was the observation that text collections of dialogue interpreting rarely meet the criteria put forward by corpus linguists to qualify as a corpus. The criterion of representativeness is especially problematic as most text collections are small and only include interpretations of a limited number of interpreters. That does not

disqualify the research carried out on them, which can yield valuable insights in terms of the interpreter's role, responsibility, interaction patterns, etc.

Our study set out to systematically analyse the use of one particular Dutch discourse marker by nine interpreters recorded in 12 police interviews. The data collection used for this is comparable in size to most available interpreting corpora. The main results can be summarised as follows: in terms of frequencies, interpreters do not divert from general usage patterns for spoken Dutch. However, their use of *dus* is mostly disconnected from the speech they are interpreting. Nearly 90 per cent of the occurrences have no equivalent in the corresponding source utterances. While explicitation seems to be at play in a certain number of cases, the bulk of instances serves interaction coordination purposes. Given the cognitive challenges interpreting poses it is not surprising to also find a substantial number of filler instances where interpreters struggle to understand the source speech or to articulate their interpretation. Some interesting cases of so-called discursive control enforced by *dus* were also observed, further confirming the special relationship interpreting holds with drafting of written records during the interview.

REFERENCES

Ainsworth, Janet. 2018. Anatomy of a false confession: The linguistic and pyschological characteristics of a false confession. In Girolamo Tessuto, Vhijay Bhatia and Jan Engberg eds. *Frameworks for Discursive Actions and Practices of the Law*. Newcastle upon Tyne: Cambridge Scholars, 23–39.

Amon, Marri. 2006. Cohérence dans le discours: Quelques remarques sur les difficultés et les stratégies des interprètes. In Kjersti Fløttum ed. *Phénomènes Linguistiques et Genres Discursifs*. Rodskild University: Rodskild University Digital Archive. http://ojs.ruc.dk/index.php/congreso/article/download/5265/2868

Angelelli, Claudia. 2017. Can ethnographic findings become corpus-studies data? A researcher's ethical, practical and scientific dilemmas. *The Interpreters' Newsletter* 22: 1–16.

Angermeyer, Philipp S. 2015. *'Speak English or what?': Codeswitching and Interpreter Use in New York City Small Claims Court*. Oxford: Oxford University Press.

Anthony, Laurence. 2014. *AntConc* (Version 3.2.4). Tokyo: Waseda University. https://www.laurenceanthony.net

Bakti, Mária and Judit Bóna. 2014. Source language-related erroneous stress placement in the target language output of simultaneous interpreters. *Interpreting* 16/1: 34–48.

Baraldi, Claudio and Laura Gavioli. 2012. Introduction: Understanding coordination in interpreter-mediated interaction. In Claudio Baraldi and Laura Gavioli eds. *Coordinating Participation in Dialogue Interpreting*. Amsterdam: John Benjamins, 1–22.

Bastin, Georges. 2003. Les marqueurs de cohérence en interprétation consecutive. *The Interpreters' Newsletter* 12: 176–187.

Bendazzoli, Claudio. 2017. Editorial: A dialogue on dialogue interpreting (DI) corpora. *The Interpreters' Newsletter* 22: VII–XVII.

Berk-Seligson, Susan. 1990. *The Bilingual Courtroom*. Chicago: University of Chicago Press.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 243–257.

Blakemore, Diane and Fabrizio Gallai. 2014. Discourse markers in free indirect style and interpreting. *Journal of Pragmatics* 60: 106–120.

Buysse, Lieven. 2017. English *so* and Dutch *dus* in a parallel corpus: An investigation into their mutual translatability. In Karin Aijmer and Diana Lewis eds. *Contrastive Analysis of Discourse-pragmatic Aspects of Linguistic Genres*. Cham: Springer, 33–60.

Crystal, David. 1995 *The Cambridge Encyclopaedia of the English Language*. Cambridge: Cambridge University Press.

Davitti, Elena. 2013. Dialogue interpreting as intercultural mediation: Interpreters' use of upgrading moves in parent-teacher meetings. *Interpreting* 15/2: 168–199.

Defrancq, Bart. 2016. Well, interpreters… A corpus-based study of a pragmatic particle used by simultaneous interpreters. In Gloria Corpas Pastor and Miriam Seghiri Dominguez eds. *Corpus-based Approaches to Translation and Interpreting: From Theory to Applications.* Bern: Peter Lang, 105–128.

Defrancq, Bart and Sofie Verliefde. 2018. Interpreter-mediated drafting of written records in police interviews: A case study. *Target* 30/2: 212–239.

Defrancq, Bart and Camille Collard. 2019. Using data from simultaneous interpreting in contrastive linguistics. In Renata Enghels, Bart Defrancq and Marlies Jansegers eds. *New Approaches to Contrastive Linguistics: Empirical and Methodological Challenges*. Berlin: Mouton de Gruyter, 159–182.

Defrancq, Bart, Koen Plevoets and Cédric Magnifico. 2015. Connective markers in interpreting and translation: Where do they come from? In Jesus Romero Trillo ed. *Corpus Pragmatics in Translation and Contrastive Studies*. Singapore: Springer, 195–222.

Degand, Liesbeth. 2001. *Form and Function of Causation: A Theoretical and Empirical Investigation of Causal Constructions in Dutch*. Leuven: Uitgeverij Peeters.

Degand, Liesbeth. 2011. Connectieven in de rechterperiferie: Een contrastieve analyse van *dus* en *donc* in gesproken taal. *Nederlandse Taalkunde* 16/3: 333–348.

Degand, Liesbeth and Geertje van Bergen. 2018. Discourse markers as turn transition devices: Evidence from speech and instant messaging. *Discourse Processes* 55/1: 47–71.

Delizée, Anne and Christine Michaux. 2019. The negotiation of meaning in dialogue interpreting: On the effects of the verbalization of interpreters' inferences. *Translation, Cognition and Behavior* 2/2: 263–382.

Evers-Vermeul, Jacqueline. 2010. 'Dus' vooraan of in het midden? Over vorm-functierelaties in het gebruik van connectieven. *Nederlandse Taalkunde* 15: 149–175.

Filipović, Luna. 2022. The tale of two countries: Police interpreting in the UK vs. in the US. *Interpreting* 24/2: 254–278.

Gallai, Fabrizio. 2013. *Understanding Discourse Markers in Interpreter-mediated Police Interviews*. Salford: University of Salford dissertation.

Gallai, Fabrizio. 2017. Pragmatic competence and interpreter-mediated police investigative interviews. *The Translator* 23/2: 177–196.
Gamal, Muhammad. 2017. Police interpreting: The facts sheet. *Semiotica* 216: 297–316.
Gao, Fei and Binhua Wang. 2017. A multimodal corpus approach to dialogue interpreting studies in the Chinese context: Towards a multi-layer analytic framework. *The Interpreters' Newsletter* 22: 17–38.
Gavioli, Laura and Claudio Baraldi. 2011. Interpreter-mediated interaction in healthcare and legal settings: Talk organization, context and the achievement of intercultural communication. *Interpreting* 13/2: 205–233.
Götz, Andrea. 2020. Discourse markers and connectives in interpreted Hungarian discourse: A corpus-based investigation of discourse properties and their interdependence. *Beszédtudomány – Speech Science* 2020: 259–284.
Gumul, Ewa and Magdalena Bartłomiejczyk. 2022. Interpreters' explicitating styles: A corpus study of material from the European Parliament. *Interpreting* 24/2: 164–191.
Hale, Sandra. 2004. *The Discourse of Court Interpreting*. Amsterdam: John Benjamins.
Halliday, Michael. 1985. *An Introduction to Functional Grammar*. London: Arnold.
Jacobsen, Bente. 2002. *Pragmatic Meaning in Court Interpreting: An Empirical Study of Additions in Consecutively Interpreted Question-Answer Dialogues*. Aarhus: Aarhus School of Business dissertation.
Jefferson, Gail. 2004. A glossary of transcript symbols. In Gene Lerner ed. *Conversation Analysis: Studies from the First Generation*. Amsterdam: John Benjamins, 13–31.
Komter, Martha. 2005. Understanding problems in an interpreter-mediated police interrogation. In Stacy Burns ed. *Ethnographies of Law and Social Control*. Bingley: Emerald, 203–224.
Komter, Martha. 2006. From talk to text: The interactional construction of a police record. *Research on Language and Social Interaction* 39/3: 201–228.
Komter, Martha. 2022. Institutional and academic transcripts of police interrogations. *Frontiers in Communication* 7. https://doi.org/10.3389/fcomm.2022.797145
Kredens, Krzysztof. 2017. Making sense of adversarial interpreting. *Language and Law/Linguagem e Direito* 4/1: 17–33.
Krouglov, Alex. 1999. Police interpreting. Politeness and sociocultural context. *The Translator* 5/2: 285–302.
Lederer, Marianne. 1980. *La Traduction Simultanée: Fondement Théoriques.* Lille: Université de Lille.
Liu, Xin and Sandra Hale. 2017. Facework strategies in interpreter-mediated cross-examinations: A corpus-assisted approach. *The Interpreters' Newsletter* 22: 57–78.
Määttä, Simo. 2017. English as a Lingua Franca in telephone interpreting: Representations and linguistic justice. *The Interpreters' Newsletter* 22: 39–56.
Mason, Marianne. 2008. *Courtroom Interpreting*. Lanham: University Press of America.
McEnery, Tony and Andrew Wilson. 1996. *Corpus Linguistics*. Edinburgh: Edinburgh University Press.
Mead, Peter. 2012. Consecutive interpreting at a literature festival. In Cynthia Kellett-Bidoli ed. *Interpreting across Genres: Multiple Research Perspectives*. Trieste: University of Trieste, 171–183.
Monteoliva-García, Eloisa. 2017. *The Collaborative Construction of the Stand-by Mode of Interpreting in Police Interviews with Suspects*. Edinburgh: Heriot-Watt University dissertation.

Nakane, Ikuko. 2014. *Interpreter-mediated Police Interviews: A Discourse-pragmatic Approach.* Basingstoke: Palgrave Macmillan.

Oostdijk, Nelleke. 2000. The *Spoken Dutch Corpus*: Overview and first evaluation. In Maria Gravilidou, George Carayanni, Stella Markantonatou, Stelios Piperidis and Gregory Stainhaouer eds. *Proceedings of the Second International Conference on Language Resources and Evaluation*. Paris: ELRA Language Resources Assocation, 887–893.

Pander Maat, Henk and Liesbeth Degand. 2001. Scaling causal relations and connectives in terms of speaker involvement. *Cognitive Linguistics* 12/3: 211–245.

Park, Joseph and Mary Bucholtz. 2009. Public transcripts: Entextualisation and linguistic representation in institutional contexts. *Text & Talk* 29/5: 485–502.

Pöchhacker, Franz and Waltraub Kolb. 2009. Interpreting for the record: A case study of asylum review hearings. In Sandra Hale, Uldis Ozolins and Ludmila Stern eds. *Quality in Interpreting – A Shared Responsibility*. Amsterdam: John Benjamins, 119–134.

RusseLl, Sonia. 2001. 'Let me put it simply…': The case for a standard translation of the police caution and its explanation. *Forensic Linguistics* 7/1: 26–48.

Shlesinger, Miriam. 1995. Shifts in cohesion in simultaneous interpreting. *The Translator* 1/2: 193–214.

Shlesinger, Miriam. 1998. Corpus-based interpreting studies as an offshoot of corpus-based translation studies. *Meta* 43/4: 486–493.

Sinclair, John. 2005. Corpus and text – Basic principles. In Martin Wynne ed. *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 1–16.

Smets, Lotte and Paul Ponsaers. 2011. Het proces-verbaal van een verdachtenverhoor: een bron van informatie? Diverse formats van geschreven communicatie tussen politie en parket. C*ahiers Politiestudies* 21/4: 123–144.

Spinzi, Cinzia. 2017. Using corpus linguistics as a research and training tool for Public Service Interpreting (PSI) in the legal sector. *The Interpreters' Newsletter* 22: 79–100.

Stukker, Ninne, Ted Sanders and Arie Verhagen. 2009. Categories of subjectivity in Dutch causal connectives: A usage-based analysis. In Ted Sanders and Eve Sweetser eds. *Causal Categories in Discourse and Cognition*. Berlin: Mouton de Gruyter, 119–171.

Tipton, Rebecca. 2021. 'Yes I understand': Language choice, question formation and code-switching in interpreter-mediated police interviews with victim-survivors of domestic abuse. *Police Practice and Research* 22/1: 1058–1076.

Van Besien, Fred. 1999. Anticipation in simultaneous interpreting. *Meta* 44/2: 250–259.

Verliefde, Sofie. 2022. *Interpreting-mediated Police Interviewing cum Drafting*. Ghent: Ghent University dissertation.

Verliefde, Sofie and Bart Defrancq. 2022. Interpreter-mediated access to the written record in police interviews. *Perspectives* 31: 519–547

Wadensjö, Cecilia. 1998. *Interpreting as Interaction*. London: Longman.

Corresponding author
Bart Defrancq
Ghent University
Faculty of Arts and Philosophy
Department of Translation, Interpreting and Communication
Groot-Brittanniëlaan 45
9000 Ghent
Belgium
E-mail: bart.defrancq@ugent.be

received: January 2023
accepted: March 2023

RiCL Research in Corpus Linguistics

Sketching the changing patterns in kaleidoscopes: New developments in corpus-based studies of translation features (2001–2021)

Shuangzi Pang – Kefei Wang
Shanghai Jiao Tong University / China
Beijing Foreign Studies University / China

Abstract – Corpus-based Translation Studies (CTS) have developed and advanced substantially since its emergence in the 1990s. This article provides an overview of the evolution of CTS from 2001 to 2021, identifying new challenges and research opportunities. The evolution of CTS is presented into two stages: the establishment of the subject matter and the expansion of research, respectively. We argue that CTS may enter the stage when the traditional specialties, such as using corpora in contrastive linguistics and translation, continue to advance, while a variety of new research points emerge and expand. After outlining current problems and unresolved issues, the analysis presents newly emerged research areas, assumptions, perspectives, and cross-fertilization with neighboring disciplines as the new developments in CTS. Four possible trends in CTS are framed and presented accordingly. The analysis highlights the significant advancements made in CTS over the past two decades and provides a valuable resource for researchers and practitioners interested in understanding the current state of CTS, and the directions it may take in the future.

Keywords – corpus-based translation studies; the third code; translation features; expansion; socio-cognitive constructs; interdisciplinary

1. INTRODUCTION[1]

Corpus-based Translation Studies (henceforth CTS) have developed and advanced substantially since they emerged in the 1990s. As Kenny (2001: Introduction) states:

> Kaleidoscopes allow us to view patterns, and to change those patterns at will. In corpus linguistics, the words and characters of electronic texts act like pieces of coloured glass and paper, constantly forming new patterns, which then recede as others take their place.

[1] This research has been funded by *The National Social Science Fund of China* (20BYY020). The grant was awarded to the project *Influence of Language Contact through Translation on the Register Features of Vernacular Chinese after May Fourth Movement in China*. We would also like to express our sincere gratitude to Sara Laviosa for her feedback and support throughout the course of the research.

Over the past three decades, the kaleidoscope of CTS has continuously witnessed the emergence and evolution of new patterns, driven by the creation and utilization of novel types of corpora. These corpus-based approaches have yielded valuable insights into translation theories and practices, encompassing diverse languages and disciplines. The notion of the 'third code' —the fact that "translation is essentially a third code which arises out of the bilateral consideration of the matrix and target codes" (Frawley 1984: 257)— has been extensively studied as an independent variety, shedding light on the underlying motivations behind this particular phenomenon. Moreover, the exploration of translation universals has led to the creation and investigation of new technical terms within this domain, such as 'interpretese' (Kajzer-Wietrzny 2018: 93), 'foreignese' (Kajzer-Wietrzny 2018: 93), and 'varioversals' (Szmrecsanyl and Kortmann 2009: 33). Through the lens of corpora, previous studies have successfully explored the universality and diversity of translation, reflecting the evolving perception of this field. However, the research focus concerning the features found in translated texts has expanded beyond the mere verification or falsification of alleged translation universals. It is now established that these alleged universals are influenced by various factors, including register, interference from the source language, social norms, socio-contextual factors, language contact, and cognition (Halverson 2003; Kruger 2018a). As of today, CTS has evolved into a distinct field, incorporating increasingly sophisticated methodologies that draw from and contribute to contrastive linguistics, translation features, translation norms, translators' style, language contact, discourse analysis, and cognitive translation studies. Notably, the study of the third code has expanded its scope and depth, encompassing a broader range of observations.

This article presents a global picture of the recent trends and cutting-edge avenues CTS has experienced in the last two decades in the domain of translation features. Section 2 provides an overview of what CTS has experienced and a range of topics and methodologies which have been used in this area for the past two decades. The evolution of CTS is divided into two stages —divergence and convergence— due to the extent of its dependency on other disciplines. Section 3 discusses some unresolved issues and current problems with CTS. Section 4 concentrates on the newly emerging areas, perspectives, and assumptions in CTS in recent years. Finally, four possible trends are framed accordingly in Section 5 and some final remarks are provided in Section 6.

2. Divergence and convergence: What CTS has experienced in the past decades

By the end of the twentieth century, the notion of 'translation universals' had gained prominence following the publication of Baker (1993). In the previous decades, the study of translation universals had produced a fruitful line of research in CTS, which may be divided into two stages.

2.1. Stage I: Establishment of the traditional research object in CTS

According to Shneider (2009), the evolution of a scientific discipline can normally be divided into four stages. Chen (2017: 5) further specifies that the object of the research is established in the first stage. The second stage is characterized by the development of research instruments, or tools, which enable researchers to investigate underlying phenomena. The third stage is a prolific stage, because many results are produced, and the understanding of the research problems is substantially advanced. Finally, in the fourth stage, the original specialty may continue to be investigated along the original research agenda, with the tools developed by the original specialty contributing to the development of other subject domains (Chen 2017: 5).

Applying the division mentioned above to the development of CTS, the scrutinization of the research object (i.e., translation universals) has had parallels with the development of electronic texts and text analysis tools in CTS. This can be exemplified by the fact that the *Translational English Corpus* (TEC)[2] was developed in the mid-1990s when Baker (1993) brought together two strands, namely Descriptive Translation Studies (DTS) and Corpus Linguistics (CL) and proposed the notion of translation universals. In other words, the evolution of CTS cannot be separated from technological tools in the beginning, as evidenced by what Zanettin (2012: 12) argues:

> a comparison of linguistic data in comparable monolingual corpus of translated and non-translated texts could unveil some regular patterns of behavior common to all translated texts.

With the above in mind, the evolution of CTS can be categorized into two stages. In the first stage (1993 to 2010), the research focus was established, and numerous studies emerged to either verify or challenge it. During this period, CTS aimed to set itself apart from other disciplines (De Sutter and Lefer 2019: 55), granting translated texts their

[2] https://genealogiesofknowledge.net/translational-english-corpus-tec/

own rightful place. Concurrently, corpora (such as TEC), research instruments, and tools were developed to investigate underlying phenomena. Subsequently, several parallel and comparable corpora, such as the *English-Norwegian Parallel Corpus* (ENCP)[3] and the *English-Portuguese Bi-directional Corpus* (COMPARA),[4] were compiled to explore various aspects and advance research in the field.

As De Sutter and Lefer (2019) argue, in its early years, CTS strongly focused on delineating the unique characteristics of translation in order to distinguish it from other disciplines. From 1993 to 2010, the primary areas of investigation included verification and falsification of translational features, contrastive studies between different languages, translators' style, and translation norms, among others. Building upon the assumption of translation universals, Klaudy and Károly (2005) proposed the 'asymmetry hypothesis', which assuming that the notions of 'explicitation' and 'implication' are not symmetric translation strategies, refined and expanded upon the existing 'explicitation hypothesis' by considering factors such as implication and translation direction. This hypothesis aimed to bridge gaps in previous studies and provide a comprehensive framework. During this stage, the research scope gradually solidified, and a significant number of studies were conducted to examine the translation features under scrutiny. Numerous scholars conducted innovative studies on translation features from an English-Chinese perspective during this stage. For example, Wang and Qin (2009) analyzed the general features of translated Chinese by comparing translated and non-translated Chinese texts. Their study revealed that translated Chinese exhibited a higher type-token ratio and employed longer sentence segments. However, their results did not fully support the hypothesis of translation universals. Additionally, Xiao (2010) explored the potential lexical and syntactic features of translational Chinese using two comparable corpora —the *ZJU Corpus of Translational Chinese* (ZCTC)[5] and the *Lancaster Corpus of Mandarin Chinese* (LCMC)[6]— and found that translated Chinese utilizes conjunctions more frequently, suggesting a tendency toward explicitation.

[3] https://varieng.helsinki.fi/CoRD/corpora/ENPC/
[4] https://lindat.mff.cuni.cz/repository/xmlui/handle/11372/LRT-866
[5] https://www.lancaster.ac.uk/fass/projects/corpus/ZCTC/
[6] https://www.lancaster.ac.uk/fass/projects/corpus/LCMC/

2.2. Stage II: Expansion of the research issues

The second stage in Shneider's (2009) model ranges from the timeframe when Becher's (2010) influential work was published to the present time, when CTS intends to expand and converge with other neighboring subjects (see Section 2.1 for the first stage). During this stage, the existence of a third code has become more doubtful, and its study has been substantially expanded.

As a breakthrough in CTS, which gained no less attention than Baker's (1993) work, Becher (2010: 1) claims that the dogma of translation-inherent features rests on fallacious theoretical considerations and premature interpretations of empirical data. According to a science mapping (see Section 4), it has been proven that Becher (2010) becomes the most cited reference in the literature on the topic during the following decade in CTS, which indicates that scholars realize that the hypothesis of translation universals is rather problematic. In 2010, there was a consensus that certain factors, such as source language interference and conservatism, were not adequately addressed in Olohan and Bake (2010) on the analysis of the translation of *that*.

Kruger (2012: 335) innovatively shows that the recurrent features, or universals, of translated language are primarily the result of a "mediation process" rather than the particularities of bilingual language processing. More recently, as De Sutter and Kruger (2018: 55) point out, most studies have witnessed a process of removing the interdisciplinary walls. In other words, CTS is being converged with neighboring disciplines, such as interpreting studies, contrastive linguistics, second language acquisition, variational linguistics, sociology, psycholinguistics, and contact linguistics, among others. Studies dealing with the third code have also concomitantly profited from a variety of new research agendas, paradigms, and settings.

As far as a paradigm is concerned, the use of solely parallel or comparable corpora has been evolving in combination. The emerging composite corpus paradigm, which combines both parallel and comparable corpus, has shown to be an innovative approach to studying the relationship between translation and language change. Through the kaleidoscope of composite corpora, we can see source language shining-through effects and the effects of translated languages on the target languages. The indirect influence of English-Chinese translation on the original in Chinese can thus be empirically measured on the basis of the new type of corpus, which incorporates a parallel and comparable corpus.

Regarding new settings, specialties —such as studies of translational features and contrastive studies between languages— have been intersected with the areas of education, language contact, translation evaluation, translator style, translators' expertise, editor's invention, legal translation, news translation and medical translation, etc. These have provided strong evidence for the notion that translated language is a third code shaped by the socio-cognitive constraints that operates in mediating process.

In this study it is proposed that CTS may enter the stage when traditional specialties —such as using corpora in contrastive linguistics and translation— continue to advance, while a variety of new research agendas emerge and are expanded at the same time.

3. Current problems in corpus-based translation studies

Although significant developments have been made in the field of CTS in recent years, there are still unresolved issues. Kruger (2018b) identifies three blind spots in CTS.

The first blind spot is the underestimation of the complexity of translated features. As Kruger (2018a: 9) claims, the hypothesis of translation universals downplays the complexity of translation, which has proven to be constrained by register, source language interference, translators' expertise, translation methods, translation direction, cross-linguistic differences, and socio-cognitive factors in recent years. For example, Redelinghuys and Kruger (2015) agree with the hypothesis that linguistic operationalizations of the translated features demonstrate significant differences in the work of experienced and inexperienced translators. Their study unveils that the translators' expertise may constrain translated features.

The second blind spot concerns the vagueness of conditioning forces, which are difficult to pinpoint. The conclusions yielded in different studies (Øverås 1998; Olohan and Baker 2000; Dimitrova 2005) are difficult to compare and replicate because the indicators are selected and operated differently. According to Kruger (2018b), what is challenging is the disentanglement of the various explanatory hypotheses proposed for the features of translated language. Currently, the causes for translation-inherent features are normally attributed to cross-linguistic influence from source languages, risk-avoidance, conservative overadjustment to the norms for formal writing, and cognitive complexity, such as chunking and entrenchment, which underlie these

differences. De Sutter and Kruger (2018) have adopted a multifactorial approach to address this issue in a corpus study. The study aims at separating the three proposed explanations for explicitation, leading ultimately to a better understanding of what translation is, how it is affected by different circumstances, and how it relates to other types of constrained communication.

Even though the existing studies pointed out above have provided an innovative and in-depth perspective of the forces motivating translation universals, what has garnered great popularity is still confined to synchronic research. There is a paucity of attention given to the investigations from a diachronic dimension. Although Hermans (1999: 155) has been keen to raise the importance of ideology and socio-historical context, previous studies have been unable to model social causation and too much emphasis is put on stability rather than on change. In the Chinese scholarship, Ke (2005) also argues that explicitation and implicitation are the key issues not only discussed in linguistic levels but also embedded in socio-historical contexts.

The third blind spot is related to the fact that, although there is currently a new consensus on the combination of theoretical frameworks and CTS research paradigms, the methods on how to integrate them are seldom investigated. Most studies have explored the assumption of translation universals, but how to combine CTS and other disciplines to gain a profound understanding of the nature of translation needs further exploration.

Methodologically, previous studies can be framed as three strands: 1) traditional research based on parallel corpora; 2) new large-scale quantitative research based on comparable corpora; and 3) composite research based on a combination of parallel and comparable corpora. There is a deficiency of internally coherent multi-lingual composite corpora incorporating parallel and comparable corpus covering different registers. While composite corpora have been employed in some studies (Bisiada 2013; Malamatidou 2018), knowing how to exploit them best remains a challenge. It is also difficult to annotate a large corpus with accuracy and precision on semantic, syntactic, and pragmatic levels, and spoken corpora are even more challenging to compile due to difficulties in data collection and transcription. Furthermore, although more complicated predictive models are designed, how to utilize them to tackle key issues in CTS remains a challenge. Particularly, using corpora properly and effectively to detect conceptual problems is still one of the difficulties in translation studies.

4. New developments in CTS in recent years

In this section, we present new research issues and the progress of methodologies in recent years regarding CTS. New research issues will partly be based on a visual analysis of science mapping. To achieve a global picture of the new development of CTS, we visualize and analyze a dataset from *Web of Science* (WOS) ranging from 2001 to 2021. To do so, we make use of a new version of *CiteSpace 5.0*.[7] The input data of our review is generated by the results from search queries in WOS. WOS is an interdisciplinary database with records from several bibliographic databases, among them *Science Citation Index Expanded* (SCI-EXPANDED) and *Social Sciences Citation Index* (SSCI). WOS contains records of publications from 1900 to the present. The topic terms 'corpus (corpora)' and 'translation' are used in this step; then 734 records, which range from 2001 to 2021, are generated.

Figure 1 shows the landscape view which is generated on the basis of the publications from 2001–2021. Areas in orange are generated earlier than those in yellow. In what follows, we will focus on the large clusters.

Figure 1. A landscape view of the co-citation network in CTS

As Figure 1 shows, the largest cluster is labeled 'mediation effect', and the most actively cited scholar in the cluster is Kruger (2012). The second largest cluster is labeled 'lexical diversity', and the academic work which is most frequently cited in the

[7] https://citespace.podia.com/

cluster is Kajzer-Wietrzny (2020). The third largest cluster is labeled ‘Spanish translation’, and the most cited author is Bisiada (2018).

The data show the first three clusters converge to one primary focus in CTS, namely that translated language is a type of constrained language influenced by a mediated effect rather than a bilingual process alone. Based on the colors, the newly activated areas are mediation effect, lexical diversity (or constrained language), theoretical functions, Spanish translation (or editor’s intervention), legal translation, English-German language contact, and medical translation. Among them, Spanish translation (cluster 2), legal translation (cluster 3), medical translation (cluster 7), and English-German language contact (cluster 14) suggest that CTS is cross-fertilizing with neighboring disciplines.

4.1. New emerging research specialties

Based on the visualization and the articles published in the linguistic and translation journals collected from WOS, the new developments that most CTS demonstrate can be briefly interpreted as the main issues which include mediation effect, contact linguistics, and cognitive translation studies.

4.1.1. Mediation effect/constrained language

For a long time, the notion of ‘translation universals’ has been the primary concern in CTS. In line with this notion, studies concerning the mediation effect, constrained language, and editors’ intervention have been recently explored further and developed into the main issues of CTS. Currently, it has been proven that the translated language is primarily the result of a mediation process that is shared among different kinds of mediated language, such as editor’s intervention rather than bilingual processing (Kruger 2012; Kruger and Van Rooy 2016).

Much of what is considered to be universal features of translation are, in fact, features constrained psycho-linguistically and socio-cognitively. This issue is exemplified in Kruger and Van Rooy (2016), who adopt a multi-dimensional approach to analyze a translation corpus and a parallel set of texts from the *International Corpus*

of English (ICE-East Africa)[8] to determine whether translated and non-native indigenized varieties of English resemble each other. The results demonstrate a shared sets of features between translated and non-native indigenized varieties of English. The study shows that such similarities are the consequence of similar constraints emanating from the cognitive and social environment in which these texts are produced (Kruger and Van Rooy 2016: 27). As a result, translated language is considered as some particular constrained communication. Kaijzer-Wietrzny (2020) further investigates cohesion in the spoken and written registers of constrained language varieties to highlight the similarities and differences in the cohesion patterns of mediated (i.e., interpreted and translated) and non-native texts with respect to original texts produced by native speakers. The study concludes that non-native and mediated texts diverge from native production using cohesive devices in different ways.

4.1.2. Contact linguistics and variational linguistics

Theories related to contact linguistics and variational linguistics have been instrumental in uncovering the motivations behind translated features in CTS. One key finding is that translated language can be seen as a contact variety influenced by language contact. For example, De Sutter and Kruger (2018) have examined the degree of lexico-grammatical explicitness in translated language and compared it to non-contact varieties to determine the factors governing *that*-omission in different types of contact varieties. This approach helps re-evaluate the explanations for the increased explicitness of translated language in the context of language contact. Additionally, CTS can help explore the influence of translation on changes in target languages. In CTS, Pang and Wang (2020) have proposed studying the diachronic correspondence between translated and non-translated texts to uncover the effect of translation on language change. Contact linguistics is a well-established field that employs an electronic methodology and draws on various approaches to predict typical interference from the socio-linguistic and structural description of bilingual communities (Weinreich 1953: 86). The field of contact linguistics covers all linguistic phenomena, such as simplification and other kinds of restructuring, that characterize the outcomes of contact (Winford 2003: 10). Nevertheless, in contact linguistics, there is a paucity of attention to typical properties of

[8] https://www.ice-corpora.uzh.ch/en.html

translations as opposed to non-translations, and such alleged universal properties of translation features are seldom considered.

In recent years, several studies have demonstrated the influence of language contact on translated language and non-translated language. One of them is Kranich *et al.* (2011), who propose 'language contact through translation' (LCTT) as a general technical term and present a stepwise approach to studying it by means of corpus methods. More recently, Malamatidou (2017) has further contributed to providing theoretical and empirical perspectives on complex mechanisms that govern the relationship between translation and language change. Likewise, Malamatidou (2018) has also innovatively developed the traditional language contact theories and attempted to integrate the two strands of translation and contact linguistics.

4.1.3. Cognitive translation studies

The combination of Translation and Cognitive Research (TCR) and CTS has yielded some robust findings, with rigorous experimental designs and sophisticated procedures, and has led to several plausible theories in recent years (Alves and Gonçalves 2007; Halverson 2003, 2017). In terms of cognitive explanations, there are two main strands in CTS: the relevance-theoretical account postulated by Alves and Gonçalves (2007) and the cognitive grammatical account proposed in Halverson (2003). Halverson (2003: 50) concludes that many patterns that are proposed to be unique to translation are most likely to be natural effects of bilingual language production, rather than universal characteristics in the translation process.

Similarly, recent studies have harnessed cognitive theories to explain the notion of translation universals, such as the 'gravitational pull hypothesis', 'magnetism', 'connectivity', and 'general chunking hypothesis' (Halverson 2017). The gravitational pull hypothesis states that translation characteristics, such as under-representation, can be explained by the structure of semantic networks and prototypes, i.e., the distance between the activated concepts in the semantic network of the bilingual or multilingual translators (De Sutter *et al.* 2017: 2). In line with the study of cognitive translation, Szymor (2018) concludes that the cognitive process underlying the human linguistic system may explain the differences between translated and non-translated texts by comparing the use of core modals in Polish legal texts (originally written in Polish) to

legal texts translated into Polish from English. To be precise, Szymor (2018) claims that it is chunking that has influenced the translators' choice of perfective forms and the authors' choice of imperfective forms.

In translated Chinese, Hu (2019) also demonstrates that the *Ba* structure is more frequently used than in non-translated texts due to the gravitation pull or the third pull in the cognitive system. The study is in line with Ke (2003) and shows that the *Ba* structure is more frequently attested in literary than that in non-literary texts.

4.2. Emergence of new assumptions and perspectives

4.2.1. Re-evaluation and re-analysis of the classic topics in CTS

The most classic topic in CTS is syntactic alteration between *that* and zero in English complement clauses (Olohan and Baker 2000). Kruger (2018b) provides a corpus-based multifactorial analysis of the alteration between the retention or omission of *that* in four register-controlled corpora to disentangle the explanations that have been proposed for the increased explicitness of translated English compared to non-translated English texts. Indicators are designed and selected to measure cognitive complexity, pragmatic risk-aversion, and source-language transfer assumptions. Her findings provide strong evidence against the transfer hypotheses and advocate the risk-aversion hypothesis, although the cognitive complexity hypothesis cannot be ruled out.

Similarly, De Sutter and Kruger (2018) analyze English translations from Dutch in Dutch-English parallel corpora and that from Afrikaans, and two self-compiled corpora of written L2 English as the basis to examine the retention and omission of *that*. What makes their study highly interesting is their adoption of the multi-factorial statistical analysis, which incorporates the analysis of the register (which is seen to measure risk-avoidance in respect of formality), the source language structure (to test cross-linguistic influence), and the distance between matrix verb of the main clause and the onset of the complement clause (to measure cognitive effort). The findings show that the choices made in contact varieties are not different from those made in the central variety. The case studies mentioned above represent a methodological advance in CTS and also demonstrate how CTS can benefit methodologically from neighboring disciplines.

4.2.2. Verification of the semantic stability hypothesis in translated texts

For a long time, CTS has been mainly focused on shallow structures in linguistics since such structures are notably marked and easily located, and the research from semantic and pragmatic levels are limited because of the difficulty to operationalize indicators empirically.

As a breakthrough on a semantic level, Vandevoorde (2016) investigates the features of translated texts mainly through the differences between the semantic elements of the verb *start* in translated and non-translated texts by adopting the 'semantic mirror' method. The author argues that the semantic meaning in translated text tends to be flattened when compared to non-translated texts. Similarly, De Baets *et al.* (2018) also investigate the semantic features of translated texts through the 'behavioral profile'. Their study aims to verify the semantic stability hypothesis, namely, whether the semantic structure of an element in translated texts is equal to that in non-translated texts and concludes that the differences between the semantic features in translated texts are lower than those in non-translated texts.

4.2.3. Research on the influence of translation on target language

The study of the relationship between translation and the change of the target language is a key issue in the field of CTS in China and the European countries. In a diachronic study (House 2011), the team led by Juliane House initiated a series of groundwork based on corpora through a 'covert translation' project at Hamburg University. It is concluded that text types and social-historical contexts are the main causes for the language changes. Similarly, Kranich *et al.* (2011) shed light on how to combine historical linguistics and CTS.

Pang and Wang (2020) study the translation and change of modern vernacular Chinese based on a new type of corpus, the *Chinese Diachronic Composite Corpus*, which is still under compilation and incorporates a parallel corpus and a comparable corpus in three sampling periods in the twentieth century and a reference corpus as a starting point in the timeframe. They examine whether explicitness in English–Chinese translations has exerted an impact on the target language, focusing on adversative conjunctions as a measure of explicitness. It is proven that the alleged translation universal of explicitation is difficult to substantiate because implicitation and

explicitation both clearly occur in the data in the three sampling periods. This also offers some support for the argument that translated language is constrained by the social context in which the translation takes place. Besides, the combination between translation features and language contact has gained more attention recently. The effect of translation features, such as explicitation/explicitness, shining-through effects, normalization, and under-representation exerted on the change of target languages are examined in the study, which also brings a new perspective to CTS.

4.3. Compilation of multilingual composite corpus and use of statistical techniques

In terms of corpus compilation, while a parallel corpus prevails in cross-lingual variation, a comparable analysis is frequently used to describe translated text features as opposed to non-translated texts. Nevertheless, due to the integration of contrastive linguistics and translation studies, CTS show a trend of the combination of a parallel corpus and comparable corpus. A type of composite corpus, combining both a parallel corpus and a comparable corpus, provides a new insight in cross-linguistic and translation studies —both synchronically and diachronically— by offering a new rationale and methodology to CTS.

In addition, CTS can also be applied to translator and interpreter training, professional practice, translation quality assessment, and machine translation. For instance, the research team at the University of Leuven, led by Sylviane Granger (Granger *et al.* 2018), has made continuous progress in the improvement of bilingual dictionaries. They have also attempted to integrate teaching research and use of corpora to examine the diversity of translators.

CTS has also benefited from advanced research tools, such as *Python* and *R* software. Annotation, statistics, and visualization derived from these tools have provided scholars with more in-depth and insightful findings that could not be noticed and operationalized in early studies. In terms of corpus annotation, the compilation of small corpora with annotation sophisticatedly designed for special uses is also in focus nowadays.

In recent years, corpus-driven and corpus-based studies have paralleled and advanced together. The statistical techniques presented to date may be used in the exploration of the internal and external motivations for the use of the translated features

through text-mining. Cluster analysis, regression analysis, conditional inference tree, and other statistical techniques have been increasingly used.

4.4. Cross-fertilization with other neighboring disciplines

In recent years, CTS has been growing and blending with other disciplines. In addition to contact linguistics and cognitive translation studies (see section 4.1 above), other disciplines, such as discourse analysis and conceptual history, have also contributed substantially to the recent theoretical successes in CTS. Beyond this, the texts examined for CTS are also broadened to medical translation, news translation, legal translation, literary translation, and religious translation.

With regard to discourse analysis, Munday (2012) conducts several case studies on the re-instantiation of appraisal meanings in different genres, based on a corpus-based approach. Similarly, Hu and Meng (2017) coined the term 'Critical Translation Studies' and argue that it grows out of the marriage between descriptive translation and critical discourse analysis. Their research aims at investigating ideological factors that operate behind the choice of texts to be translated, the use of translation strategies and methods, features of translated texts, reception of translated texts, and the impact of translation on ideology. In the same vein, Pan and Li (2021) examine the retranslation of political texts —specifically work reports by the Communist Party of China— as a special genre in its own right and contributes to the exploration of the relationship between translation and ideology. By focusing on the retranslation of a recurring set of Chinese political concepts, culture-specific items, and preferred usages into English from the early 1990s to the late 2010s, Pan and Li (2021) showcase how and why the retranslations have been carried out as motivated by the evolving ideologies of the authors.

In the examination of conceptual history, Jones (2020) explores how the shifts in attitudes towards the proper aims and methods of history writing might have shaped the interpretation and translation into English of *Thucydides' History of the Peloponnesian War*, a work originally written in classical Greek in the fifth century BCE. It initiated the groundwork of the concept studies in the field of translation through the lens of corpora.

It is also worth pointing out that corpus-based methodologies have proven particularly fruitful in the investigation of legal translation. For instance, results from corpus-based analyses can support the description of terminological and phraseological features of legal genres, as well as the acceptability required to make translation decisions and elaborate lexicographical resources in line with legal and institutional translators' needs (Ramos Prieto 2020). Similarly, it is also worth noting that, in recent years, CTS is expanding to Second Language Acquisition and the study of English as a Lingua Franca because both areas focus on bilingual processing.

5. Possible trends and future directions in CTS

Over the past three decades, CTS has been gradually tearing down the walls between different linguistic disciplines (De Sutter and Kruger 2018) and cross-fertilizing with neighboring disciplines. This has led to equal attention being given to both synchronic and diachronic studies. As De Sutter and Lefer (2019: 2) assert, what is truly essential in CTS is the exploration of corpus-linguistic methods of scrutinizing translational products in order to find the "principles that govern translational behaviour and the constraints under which it operates" (Baker 1993: 235). Motivated by the investigation of the third code, CTS aims to move beyond the traditional examination of indicators, and more theoretical frameworks and disciplines need to be involved. In brief, the possible trends can be framed as shown below.

5.1. Research aims: From empirical research to theoretical constructs

As De Sutter and Lefer (2019: 4) suggest, the preoccupation with finding linguistic differences between translated and non-translated texts has left the explanatory framework postulated by Baker —or any other theoretical framework— underdeveloped. Empirical research is conducted from the outset in CTS, but it needs to be guided by theoretical principles to advance when it develops further. Furthermore, since fundamental questions remain largely unanswered (De Sutter and Lefer 2019: 2), theoretical investigations from social, pragmatic, and cognitive mechanisms are expected to provide nutrition for the discipline. In brief, only through theoretical guidelines can we fully understand the phenomenon and unveil the truth of translated features.

Currently, the theoretical construction of CTS can be framed within three directions. These directions encompass cross-linguistic, social, and cognitive explanations of the conditioning forces at play. Exploring neighboring disciplines such as corpus linguistics, linguistic typology, contact linguistics, sociology, Second Language Acquisition, and psycholinguistics can contribute to a more comprehensive understanding of the fundamental principles underlying the nature of translation and the factors that influence it. As CTS advances and enters a new stage of development, new theoretical models are likely to emerge. For instance, Kotze and Havelson (2021) have introduced socio-cognitive constructs in CTS, based on the concepts that linguistic knowledge represents the cognitive organization of an individual's language experience, and normativity involves a feedback loop of conventionalization and legitimization. Their proposed translation model, which connects these two visions, offers an innovative perspective for CTS to explain translation phenomena and understand translators' behavior. This model provides a fresh and insightful framework that enhances the explanatory power of CTS.

5.2. Research objects: From universality to more constraining factors

The research on the constraining factors of translation texts needs to involve more variables, such as language contact, cross-lingual comparison, register and translators' expertise, and cognitive factors. The exploration of the universality of translated texts as a third code only serves as a starting point. The goal of this discipline is not restricted to seeking common features in varieties. In contrast, the significance of the third code consists of finding similarities and differences across different varieties to unveil the nature of translation *per se* and all the possible factors shaping constrained communication.

Currently, in line with the traditional topics in CTS, the investigation of the relationship between translation and language change remains an issue of interest in this discipline. The studies of translation texts from the perspective of language contact have yielded a substantial body of research. These studies are in fact proceeded in the 'covert translation' project (House 2011), and in the 'cross-linguistic corpora' (Cro Co) project, led by Silvia Hansen-Schirra, Stella Neumann, and Erich Steiner in Germany (Alves *et al.* 2010). At present, the scope of the studies is gradually expanded, covering linguistic typology, social-linguistics, and cognition.

In terms of constraining factors in language contact situations, Kotze (2020: 346) lists five "overarching and interacting constraint dimensions" that may affect language production: 1) language activation (monolingual–bilingual); 2) modality and register (spoken–written–multimodal); 3) text production (independent/unmediated dependent/mediated); 4) proficiency (proficient–learner); and 5) task expertise (expert–non-expert). Apart from these dimensions, it is expected that more constraining factors will be involved and detected in future studies.

5.3. Research perspectives: From pure linguistics to interdisciplinarity

As noted above, corpus-based translation and interpreting studies are experiencing an upward trend. It is this very interdisciplinary approach that continuously defines CTS and gives it strength to develop and continue to advance. As Halverson (2018) claims, collision of multi-disciplinaries and multi-methods will lead to integration in the next generation of CTS scholar' research.

For instance, as one important shaping factor, language contact needs more in-depth investigation and verification. The influence of the source language on the target language will help us unveil the nature of translation, as well as the mechanism of code-switching in language change. From the perspective of research methods, the measurement of the similarities and differences between translated and non-translated texts needs more research models and to involve indicators.

In the present time, conjunctions and pronouns are still the main foci of research in CTS, complemented by content words, such as nouns, adjectives, adverbs, cognate words, n-grams, and phraseologies. Besides, new assumptions have also emerged and promoted the study of translation universals. The research perspectives as starting points include the analysis of word chunks, noun phrases, semantic priming, frequency effect, grammatical metaphors, etc., which are not easy to notice. For example, the comparison of the use of verbless sentences in English and Russian and the contrast of linguistic features between directly translated texts and indirectly translated texts also provide insights and highly enrich CTS. In general, the continuous exploration of traditional topics and investigation of new assumptions and perspectives are expected to bring CTS into a new stage.

5.4. Research procedures: From simplification to sophistication

Some important challenges remain in regard to corpus compilation, which is key to CTS. New design of corpora is needed in corpus building, such as multilingual and diachronic composite corpora with more precise semantic and pragmatic annotation. Not only the compilation of large-scale corpora but also of small corpora should be given more attention. In addition, the comparison between written and spoken corpora is also believed to yield new insights into CTS in the future, because it is effective to expand the scope of analysis to the spoken mediated variety.

In terms of statistical analysis, it is expected that frequency-based methods may be replaced by multidimensional, multifactorial, and multivariate statistics. Corpus building and statistical methods serve as one of the driving forces to advance the research in CTS. In the future, it is predicted that statistical methods may be more diversified and sophisticated, such as the use of language modeling statistics by loading and using *R* language packages. Triangulation methods, which combine process and product with empirical data, will be fully exploited and helpful in the future.

Compared to the research conducted in the European countries over the past 20 years, quantitively speaking, Chinese scholars have provided a large bulk of studies in CTS with many influential pieces of research, such as those by Wang and Qin (2009), Xiao (2010), Hu (2012), and Wang (2012). Although there is always room for Chinese research to improve in theoretical constructs, it has contributed substantially to the compilation of tailor-made corpora, such as the *China English-Chinese Parallel Corpus*[9] compiled by Beijing Foreign Studies University and the *Political Discourse Corpus*[10] compiled by Shanghai International Studies University, both of which expand the domain of CTS.

6. FINAL REMARKS

This article has provided an overview of the progress that CTS has undergone over the past few decades since its inception in the UK. The development of CTS can be identified into two stages: the establishment of traditional research object in CTS and the expansion of new research issues. Here we have highlighted the current problems

[9] http://114.251.154.212/cqp/
[10] http://imate.cascorpus.com/

that persist in the field, including the underestimated complexity of translated features, unresolved conditioning forces, underdeveloped theoretical constructs, and deficient internally coherent corpora with deep annotation. In response to overcoming these problems, the article also presents new developing areas in CTS, such as mediation effects, contact varieties, and cognitive translation studies.

Furthermore, we have identified four trends that have emerged in the field of CTS. First, there has been a shift from empirical research to the development of theoretical constructs in CTS. Second, translation texts are being viewed as contact varieties that are influenced by a range of constraining factors. Third, CTS has adopted an interdisciplinary approach, expanding research perspectives beyond traditional linguistic and translation studies. Finally, there is a growing interest in creating multilingual and diachronic composite corpora and conducting multivariate statistical analyses.

When we look back to what Laviosa (2004: 22) envisaged in CTS at the beginning of this century, it can be said that CTS has not disappointed us in that the potential of corpora in translation studies has been exploited to a large extent. It seems fair to state that, three decades after the publication of Baker's seminal papers, CTS has advanced substantially in the process of converging with neighboring disciplines. At the same time, it has contributed considerably to translation studies, deepening our perception of the nature of translation. We believe what the future holds for CTS is the promotion of interdisciplinary work leading the way towards hybridity and polysemy.

REFERENCES

Alves, Fabio and José L. Gonçalves. 2007. Modelling translator's competence: Relevance and expertise under scrutiny. In Yves Gambier, Miriam Shlesinger and Radegundis Stolze eds. *Doubts and Directions in Translation Studies*. Amsterdam: John Benjamins, 41–45.

Alves, Fabio, Adriana Pagano, Stella Neumann, Eric Steiner and Silvia Hansen-Schirra. 2010. Translation units and grammatical shifts: Towards an integration of product and process-based translation research. In Gregory M. Shreve and Eric Angelone eds. *Translation and Cognition.* Amsterdam: John Benjamins, 109–142.

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis and Elena Tognini Bonelli eds. *Text and Technology: In Honour of John Sinclair.* Amsterdam: John Benjamins, 233–252.

Becher, Viktor. 2010. Abandoning the notion of 'translation-inherent' explicitation: Against a dogma of translation studies. *Across Languages and Cultures* 11/1: 1–28.

Bisiada, Mario. 2013. *From Hypotaxis to Parataxis: An investigation of English-German Syntactic Convergence in Translation*. Manchester: The University of Manchester Dissertation.

Bisiada, Mario. 2018. The editor's invisibility: Analyzing editorial intervention in translation. *Target* 30/2: 288–309.

Chen, Chaomei. 2017. Science mapping: A systematic review of the literature. *Journal of Data and Information Science* 2/2: 1–40.

De Baets, Pauline, Lore Vandevoordee and Gert De Sutter. 2018. Meaning shifts in translation: A corpus-based behavioural profile approach. In Sylviane Granger, Marie-Aude Lefer and Laura Aguiar de Souza Penha-Marion eds., 47–49.

De Sutter, Gert and Haidee Kruger. 2018. Disentangling the motivations underlying syntactic explicitation in contact varieties: A MuPDAR analysis of *that* vs. zero complementation. In Sylviane Granger, Marie-Aude Lefer and Laura Aguiar de Souza Penha-Marion eds., 55–57.

De Sutter, Gert and Marie-Aude Lefer. 2019. 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.

De Sutter, Gert, Bert Cappelle, Orphée De Clercq, Loock Rudy and Koen Plevoets. 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.

Dimitrova, Englund. 2005. *Expertise and Explicitation in the Translation Process*. Amsterdam: John Benjamins.

Frawley, William. 1984. Prolegomenon to a theory of translation. In Lawrence Venuti ed. *The Translation Studies Reader*. London: Routledge, 250–263.

Granger, Sylviane, Marie-Aude Lefer and Laura Aguiar de Souza Penha-Marion eds. 2018. *Book of Abstracts: Using Corpora in Contrastive and Translation Studies*. Louvain-la-Neuve: Centre for English Corpus Linguistics and Université Catholique de Louvain.

Halverson, Sandra. 2003. The cognitive basis of translation universals. *Target* 15/2: 197–241.

Halverson, Sandra. 2017. Developing a cognitive semantic model: Magnetism, gravitational pull and questions of data and method. In Gert De Sutter, Marie-Aude Lefer and Isabelle Delaere eds. *Empirical Translation Studies: New Methodological and Theoretical Traditions*. Berlin: Mouton de Gruyter, 9–45.

Halverson, Sandra. 2018. Cognitive translation studies and the combination of data types and methods. In Sylviane Granger, Marie-Aude Lefer and Laura Aguiar de Souza Penha-Marion eds., 3–5.

Hermans, Theo. 1999. *Translation in Systems: Descriptive and System-oriented Approaches Explained.* Manchester: St Jerome.

House, Juliane. 2011. Using translation and parallel text corpora to investigate the influence of global English on textual norms in other languages. In Alet Kruger, Kim Wallmach and Jeremy Munday eds. *Corpus-based Translation Studies: Research and Applications*. London: Continuum, 187–208.

Hu, Kaibao. 2012. *Introduction to Corpus-based Translation Studies.* Shanghai: Shanghai Jiao Tong University.

Hu, Xianyao. 2019. Re-examining *ba* structure in translated Chinese. Presentation at the *Fifth Conference of Corpus-based Translation Studies*. November 2019 Chongqing: Chongqing university.

Hu, Kaibao and Lingzi Meng. 2017. Critical translation studies: New development in translation studies. *Journal of Foreign Languages* 40/6: 57–68.

Jones, Henry. 2020. Retranslating *Thucydides* as a scientific historian: A corpus-based analysis. *Target* 32/1: 59–82.

Kajzer-Wiietrzny, Marta. 2018. Translationese, interpretese and foreignese: What do they have in common? In Sylviane Granger, Marie-Aude Lefer and Laura Penha-Marion eds. *Book of Abstracts. Using Corpora in Contrastive and Translation Studies Conference*. Louvain-la-Neuve: Centre for English Corpus Linguistics/Université catholique de Louvain, 93–94.

Kajzer-Wietrzny, Marta. 2020. A multivariate approach to lexical diversity in constrained language. *Across Languages and Cultures* 21/2:169–194.

Ke, Fei. 2003. Features and dispersion of *ba* structure in Chinese texts and its Chinese-English translation. *Foreign Languages and their Teaching* 177/12:1–5.

Ke, Fei. 2005. Implicitation and explicitation in translation. *Foreign Language Teaching and Research* 37/4: 303–307.

Kenny, Dorothy. 2001. *Lexis and Creativity in Translation: A Corpus-based Study.* Manchester: St. Jerome.

Klaudy, Kinga and Krisztina Károly. 2005. Implicitation in translation: Empirical evidence for operational asymmetry in translation. *Across Languages and Cultures* 6/1: 13–28.

Kotze, Haidee. 2020. Translation, contact linguistics and cognition. In Fabio Alves and Arnt Lykke Jakobsen eds. *The Routledge Handbook of Translation and Cognition.* Abingdon: Routledge, 113–132.

Kotze, Haidee and Sandra Havelson. 2021. Norms, constraints, risks: A usage-based perspective on sociocognitive constructs in corpus-based translation studies (and beyond). In Sara Castagnoli, Silvia Bernardini, Adriano Ferraresi and Maja Miličević Petrović eds. *Book of Abstracts: Using Corpora in Contrastive and Translation Studies Conference*. Bertinoro: University of Bologna, 189–190.

Kranich, Svenja, Viktor Becher, Steffen Hoder and Juliane House eds. 2011. *Multilingual Discourse Production*. Amsterdam: John Benjamins.

Kruger, Haidee. 2012. A corpus-based study of the mediation effect in translated and edited language. *Target* 24/2: 355–388.

Kruger, Haidee. 2018a. *That* again: A multivariate analysis of the factors conditioning syntactic explicitness in translated English. *Across Languages and Cultures* 20/1:1–33.

Kruger, Haidee. 2018b. Expanding the third code: Corpus-based studies of constrained communication and language mediation. In Sylviane Granger, Marie-Aude Lefer and Laura Aguiar de Souza Penha-Marion eds., 9–12.

Kruger, Haidee and Bertus Van Rooy. 2016. Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide* 37/1: 26–57.

Laviosa, Sara. 2004. Corpus-based translation studies: Where does it come from? Where is it going? *Language Matters* 35/1: 6–27.

Malamatidou, Sofia. 2017. Creativity in translation through the lens of language contact: A multilingual corpus of *A Clockwork Orange*. *The Translator* 23/3: 1–18.

Malamatidou, Sofia. 2018. *Corpus Triangulation: Combining Data and Methods in Corpus-based Translation Studies*. London: Routledge.

Munday, Jeremy. 2012. New directions in discourse analysis for translation: A study of decision-making in crowdsourced subtitles of Obama's 2012 state of the Union speech. *Language and Intercultural Communication* 12/4: 321–334.

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

Øverås, Linn. 1998. In search of the third code: An investigation of norms in literary translation. *Meta* 43/4: 557–570.

Pan, Feng and Tao Li. 2021. The retranslation of Chinese political texts: Ideology, norms, and evolution. *Target* 33/3: 381–409.

Pang, Shuangzi and Kefei Wang. 2020. Language contact through translation: The influence of explicitness in English–Chinese translation on language change in vernacular Chinese. *Target* 42/3: 420–455.

Ramos Prieto, Fernando. 2020. Translating legal terminology and phraseology: Between inter-systemic incongruity and multilingual harmonization. *Perspectives* 29/2: 175–183.

Redelinghuys, Karien and Haidee Kruger. 2015. Using the features of translated language to investigate translation expertise: A corpus-based study. *International Journal of Corpus Linguistics* 20/3: 293–325.

Shneider, Alexander M. 2009. Four stages of a scientific discipline: Four types of scientists. *Trends in Biochemical Sciences* 34/5: 217–223.

Szmrecsanyl, Benedikt and Bernd Kortmann. 2009. Vernacular universals and angloversals in typological perspective. In Markku Filppula, Juhani Klemola and Heli Paulasto eds. *Vernacular Universals and Language Contact: Evidence from Varieties of English and Beyond*. New York: Routledge, 33–53.

Szymor, Nina. 2018. Translation: Universals or cognition? A usage-based perspective. *Target* 30/1: 53–86.

Vandevoorde, Lore. 2016. *On Semantic Differences: A Multivariate Corpus-based Study of the Semantic Field of Inchoativity in Translated and Non-translated Dutch.* Ghent: Ghent University dissertation.

Wang, Kefei. 2012. *Exploring Corpus-based Translation Studies.* Shanghai: Shanghai Jiao Tong University Press.

Wang, Kefei and Hongwu Qin. 2009. A parallel corpus-based study of general features of translated Chinese. *Foreign Language Research* 146: 102–105.

Weinreich, Uriel. 1953. *Language in Contact: Findings and Problems*. The Hague: Mouton de Gruyter.

Winford, Donald. 2003. *Contact Linguistics*. Oxford: Blackwell.

Xiao, Richard. 2010. How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics* 15/1: 5–35.

Zanettin, Federico. 2012. *Translation-Driven Corpora*. Manchester: St. Jerome.

Corresponding author
Kefei Wang
Beijing Foreign Studies University
National Research Center for Foreign Language Education
North Xisanhuan Avenue
PO Box 89-45
100089 Beijing
China
Email: kfwang@bfsu.edu.cn

received: November 2022
accepted: February 2023

RiCL Research in Corpus Linguistics

Lexical simplification in learner translation: A corpus-based approach

Ho Ling Kwok[a] – Sara Laviosa[b] – Kanglong Liu[a]
The Hong Kong Polytechnic University[a] / China
University of Bari Aldo Moro[b] / Italy

Abstract –The advance of corpus-based methodology in translation studies has greatly enhanced our understanding of the nature of translational language. While most research efforts have focused on identifying the unique features of translations carried out by professionals, comparatively fewer studies have investigated the linguistic features of student translations. In this corpus-based study, we examine if learner translations carried out by Hong Kong students exhibit lexical simplification features *vis-à-vis* comparable written texts. The study is based on two comparable corpora: the *International Corpus of English in Hong Kong* (ICE-HK) and the *Parallel Learner Translation Corpus* (PLTC) compiled at The Hong Kong Polytechnic University. Following Laviosa (1998), we compare four main lexical features (lexical density, type-token ratio, core vocabulary coverage, and list head coverage) to investigate if student translations show a simplification trend. The results demonstrate that Chinese-to-English translation is not lexically simpler than English as a Second Language (ESL) writing. Furthermore, it is lexically denser than ESL writing. Our study aims to provide new insights into learner translation as a form of constrained communication.

Keywords – lexical simplification; learner translation; corpus-based approach; students' translations

1. INTRODUCTION[1]

Translational language is often regarded as a 'third language' (Duff 1981) or 'third code' (Frawley 1984) since it involves the bilateral consideration and accommodation of at least two different codes. In this regard, Baker (1993: 243) proposed the hypothesis of translation universals, referring to them as "universal features of translation […] which typically occur in translated text." Baker (1996) put forward four translation universals: 1) simplification (tendency to simplify language subconsciously), 2) explicitation (tendency to make information clearer), 3) normalization or conservatism (tendency to

[1] This research is sponsored by a General Research Fund (GRF) grant from the *Research Grants Council of Hong Kong* (Ref: 15605520).

conform to typical patterns of the target language), and 4) levelling out (tendency to be more homogeneous than the original texts). Many translation scholars have argued that the term 'universals' is not scientifically sound (e.g., Tymoczko 1998; Pym 2008; Saldanha 2011). House (2015: 62) even suggested that "the quest for translation universals is in essence futile." In her opinion, the absence of careful comparative analyses is an inadequacy in most existing studies, and the terms used to denote them — 'simplification' and 'normalization'— are overly general and lack a clear operational definition. Besides, House (2015: 62–63) argued that universality in translation is questionable as some translation universals are subject to the variables of translation directions and genres. This empirical evidence challenges the claim of translation universals.

Despite these controversies, Baker's initial proposal suggested several research directions that have yielded new insights into the features of translational language (e.g., Olohan and Baker 2000; Xia 2014; Liu and Afzaal 2021). Baker (1993, 1995, 1996) also pioneered the application of corpus methods to identify features of translated texts, especially the use of comparable corpora to compare translated texts with non-translated ones. Over the years, the quest for translation features has been spurred by advances in corpus-based translation methods and the availability of large-scale computerized corpora. These developments have made it possible to study translation phenomena systematically instead of relying on the researchers' own experience and subjective evaluation, thus enhancing our understanding of the nature and role of translational language.

In the past three decades, although the features of professional translations have been examined extensively, comparatively little effort has been made to investigate the linguistic features of student translations. As corpus research into learner translation can reveal potential learner problems, a systematic investigation of some central issues in this area has pedagogical implications. This line of research was pioneered by Bowker and Bennison (2003), who described the construction of a student translation archive. Since then, more scholars have devoted themselves to this research area. A recent effort is the *Multilingual Student Translation Project* (MUST; Granger and Lefer 2020), which aims at compiling a student translation corpus covering different language pairs. Overall, we have witnessed an increase in scholarly interest in the field of learner translation in recent years.

The current study is based in Hong Kong, which has been active in learner corpus research over the past two decades due to its bilingual environment. However, these learner corpus studies are mainly related to Second Language Acquisition (SLA) rather than to translation studies (Liu *et al.* 2022). As L1-L2 translation and L2 writing share common challenges and constraints in terms of L2 language production, we used various lexical simplification indicators to identify the extent to which L2 learner translation differs from L2 writing and uncover their possible relationship from a constrained language perspective.

1.1. Constrained communication

Lanstyák and Heltai (2012: 100) suggested the term 'constrained communication' to indicate "communication taking place under conditions where one or several of the potential limiting factors play a greater than average role." This framework implies that all communicative events are influenced by different types and degrees of constraints. However, some have exceptionally prominent constraints, such as language contact situations, including translation and bilingual communication (Lanstyák and Heltai 2012: 100). Based on this framework, Kruger and van Rooy (2016: 27) proposed the term 'constrained language' to denote the language produced under apparent constraints. They also pointed out that both translation and L2 language varieties share the same constraints in the form of bilingual activation and language contact.

These constraints are exhibited in two ways, namely psycholinguistic and social. From a psycholinguistic perspective, constraints are associated with language processing. Bilinguals activate languages along the continuum from a monolingual mode to a bilingual mode in different contexts (Grosjean 2013: 15). In this regard, translation is always operated in the continuous bilingual activation mode (Kruger and van Rooy 2016: 29). In addition, translation is restricted by the pre-existing source text, which can interfere with target language production (Toury 2012: 310–311). The constraints of L2 production are associated with the cognitively demanding environment experienced by non-native speakers in contact situations (Kruger and van Rooy 2016: 31).

From a social perspective, the constraints are also related to language and translation norms. Translators must ensure that the translation is faithful to the source language culture as well as acceptable in the target language culture. Similarly, L2 writers

also need to conform to the perceived norms of L2 (Kruger and van Rooy 2016: 27). Against this background, L2 translation, especially the one done by student translators, can be a unique constrained language output that combines features of both translation and L2 production.

In translation, the rapid bi-directional switching involved in the translation process increases the demand for working memory. The lack of a common communication context due to linguistic differences can also lead to malcommunication or non-communication. Therefore, strategies for cognitive load reduction (Carl and Dragsted 2012) and risk minimization (Pym 2015), such as literal translation, explication, and simplification, are often applied by translators. Similarly, L2 writing also involves cognitive and communicative difficulties on the part of non-native speakers who need to use and process an additional language. Some scholars argue that simplification is one of the strategies to deal with these challenges (e.g., Kortmann and Szmrecsanyi 2009; McWhorter 2011). Under these claims, simplification is believed to be one of the features related to translational language and L2 varieties.

1.2. Lexical simplification

Simplification is defined as "the tendency to simplify the language used in translation" (Baker 1996: 181). Over the years, simplification has been studied at different levels, such as lexical (Laviosa 1998; Ferraresi *et al.* 2018; Nasseri and Thompson 2021) and syntactic (McWhorter 2011; Liu and Afzaal 2021). In the current research, we aim to investigate whether L2 translation and L2 writing are (dis)similar in terms of lexical simplification. Lexical simplification can be described as "making do with less words" (Blum-Kulka and Levenston 1983: 119). Lexical simplification is usually operationalized through indicators such as lexical density, the use of frequent words, type-token ratio, and mean sentence length, amongst others (Hu 2016: 101).

The lexical simplification hypothesis has been widely discussed in translation studies. Chesterman (2004) regarded simplification as a potential T-universal, which concerns the translation features in relation to non-translation in the target language. The simplification hypothesis thus assumes that translated texts are simpler than comparable non-translated native texts in the target language. Laviosa (1998) reported evidence that supports the lexical simplification hypothesis with certain parameters. She found that

translated narrative prose has lower lexical density, a higher proportion of high-frequency words, and more repetition of list head words than original narrative works in English. Hu (2016: 12, 22) reviewed a few studies focusing on translated Chinese that confirm Laviosa's (1998) findings. For instance, translated fiction is found to have lower lexical variety, lower lexical density, and a higher percentage of high-frequency words than non-translated fiction (Hu 2007). Similarly, Wen (2009) showed that translated detective fiction has lower type-token ratio, lower lexical density, and lower mean sentence length than non-translated detective fiction.

A number of studies, however, have not confirmed the simplification hypothesis. These studies reported contradictory findings with some parameters, such as higher mean sentence length (Laviosa 1998), untypical lexical patterning (Mauranen 2000), and overuse of degree modifiers (Jantunen 2004) in translated compared with non-translated texts. The study by Ferraresi *et al.* (2018) even rejected Laviosa's (1998) findings. For example, they found that French-English translated texts are not simpler than non-translated texts and that Italian-English translated texts are more complex than the non-translated ones in that they contain fewer common words and are also lexically denser. Lexical simplification is thus a controversial translation hypothesis.

Lexical simplification (or complexity) is also a research topic in SLA. It is considered an indicator of lexical proficiency and language production quality of L2 users (Bulté and Housen 2012; Lu 2012). Generally speaking, texts that are lexically more complex are associated with higher L2 proficiency (Laufer and Nation 1995; Jarvis 2002; Crossley and McNamara 2012). Controversial results were obtained when comparing the lexical complexity of texts produced by L2 writers with those by native speakers. Gonzalez (2013) found that native texts show significantly greater lexical diversity and a higher proportion of low-frequency words than non-native texts. Jarvis (2002) also suggested that native texts generally have higher lexical diversity than non-native texts. By contrast, Nasseri and Thompson (2021) compared academic writing produced by English native, ESL, and English as Foreign Language (EFL) students, and showed that the texts produced by EFL students have the lowest lexical density and diversity despite the fact that the English native and ESL groups produced texts with similar lexical density and diversity. These findings seem to suggest that other factors, such as L1 background, L2 instruction, and L2 proficiency, probably have an influence on the lexical complexity of L2 writing (Jarvis 2002: 57).

1.3. Research questions

Translation and L2 writing are forms of constrained communication. Simplification, a strategy dealing with constraints, is conceivably a feature that characterizes both of them. However, translation and L2 communication are different in that the former is "dependent text production" while the latter is "independent text production" (Lanstyák and Heltai 2012: 101). As student translators usually have sufficient L2 proficiency but are not fully competent in translation, they might experience different degrees of constraints which affect their use of translation strategies during text production.

The review of the literature above makes it clear that there is a gap in corpus research into L2 learner translation and L2 writing. To bridge this gap, the present study examines how L2 learner translation and L2 writing might converge or diverge in lexical simplification. In this study, we address two research questions:

RQ 1: How is L2 learner translation (dis)similar to L2 writing in the four lexical simplification parameters?

RQ 2: What are the possible factors that account for the (dis)similarities?

The findings are expected to provide a better understanding on how L2 learner translation is possibly influenced by translation and L2 (interlanguage) factors. As prospective professional translators, student translators are "major stakeholders in translator training" (Li 2002: 513). The current study is thus important for the learners' development of L2 translation competence.

2. METHODOLOGY

2.1. Corpora

This study adopts a corpus-based methodology to investigate the lexical simplification of L2 learner translation and L2 writing. The investigation is based on two comparable corpora: the *International Corpus of English in Hong Kong* (ICE-HK; Bolt and Bolton 1996; Nelson 2006) and the *Parallel Learner Translation Corpus* (PLTC) compiled at The Hong Kong Polytechnic University.

ICE-HK is an existing corpus initiated by Bolt and Bolton (1996) in the early 1990s. It is part of the *International Corpus of English* project (ICE) initiated by Greenbaum (1988). The project aimed to collect comparative English data representing different

regional varieties of English. ICE-HK follows the general structure of ICE[2] worldwide to collect English data from the Chinese population in Hong Kong, whose first language is Cantonese and whose primary and secondary education is in Hong Kong. ICE-HK contains a wide range of text categories, including different communication modes (i.e., spoken and written) and registers (e.g., direct conversations, broadcast news, business letters, academic writing, and novels, among others). ICE-HK thus represents English as a second language (ESL) in Hong Kong (Nelson 2006).

PLTC is a learner translation corpus being compiled at The Hong Kong Polytechnic University.[3] It is constructed to match the composition in text categories and proportion in size as the written component (printed subcategory) of ICE-HK. The aim of PLTC is to document learner translated English in different written registers (e.g., academic writing, novels, etc.) in Hong Kong. The compilation of PLTC consists of a two-stage procedure: preparation of Chinese textual materials and collection of learner translations done from Chinese to English. In the first stage, qualified translators first translate ICE-HK texts from English into Chinese. Copyeditors then check and edit the texts for quality control. The edited Chinese texts are then further checked and approved by the translator and copyeditors together, and the approved versions are used as source texts for translation in the next stage. In the second stage, second-year to fourth-year undergraduate students majoring in translation at The Hong Kong Polytechnic University are invited to participate in the current study via mass email. Interested students sign the consent form to indicate their intention to participate in the study. The researchers then select the participants by taking into consideration their language and educational background. The eligible participants must speak Cantonese proficiently and receive secondary education in Hong Kong. So far, a total of 28 eligible students have been recruited as participants, as shown in Table 1.

[2] https://www.ice-corpora.uzh.ch/en.html

[3] PLTC is still under compilation. More participants will be recruited to produce additional translated texts for the corpus. It is anticipated that PLTC will have a more balanced distribution of participants and registers in its completed version. Further information about the corpus project may be found at https://cerg1.ugc.edu.hk.

	Students (*n* = 28)
Age (years)	20.3 (*SD* = 1.7)
Gender	
Female	27 (96.4%)
Male	1 (3.6%)
Education level	
Year 2	15 (53.6%)
Year 3	8 (28.6%)
Year 4	5 (17.9%)

Table 1: Demographic data of the participants in PLTC

1. Please, translate the essay from Chinese into English. The target audience are native English speakers who are interested in learning more about this essay topic.

2. There is no time or word limit.

3. You can use different tools, including books, dictionaries, and internet resources, to help you complete the translation, but you cannot consult others for any translation solutions.

4. Please, record the approximate time and all translation tools you use to complete the translation.

5. Please, use proper wording and grammar. Make sure that the translation is complete and appropriate.

6. The register of the essay is [*register and sub-register are provided*].

Registers and sub-register	
Academic writing. (Humanities)	Non-academic writing. (Humanities)
Academic writing. (Social sciences)	Non-academic writing. (Social sciences)
Academic writing. (Natural sciences)	Non-academic writing. (Natural sciences)
Academic writing. (Technology)	Non-academic writing. (Technology)
Reportage. (Press news reports)	Instructional writing. (Administrative writing)
	Instructional writing. (Skills and hobbies)
Persuasive writing. (Press editorials)	Creative writing. (Novels and stories)

Table 2: Translation brief (translated from Chinese)

Participants are provided with a written translation brief in Chinese. The brief, which is shown in Table 2, above, states that their task is to translate a Chinese text into English for native English speakers who are interested in learning more about the essay topic. It also stated what the register of the source text belongs to. Participants are also instructed to use any resources and tools they think they are useful to assist them with their translation without time and word limits. However, they cannot consult other people about translation solutions. Each participant translates one to four texts, depending on the willingness to continue with the study. In order to ensure the representativeness of the translated texts for each register, no participant is allowed to translate more than one text

for each register. Besides, no participants are allowed to translate more than four texts to ensure participant/subject representativeness.

At the time of writing, 53 texts have been collected for PLTC. This study is based on these 53 text samples to represent the L2 learner translation corpus (henceforth L2T) and 53 corresponding text samples extracted from ICE-HK to represent the L2 writing corpus (henceforth L2W). L2T and L2W cover six major registers: academic writing, popular writing, reportage, instructional writing, persuasive writing, and creative writing. Each text contains around 2,000 words. L2T has a total of 125,178 tokens (total number of items) and 11,880 types (number of unique items), while L2W has 127,835 tokens and 12,704 types, as shown in Table 3.

Corpora	**Label**	**Nature**	**Files**	**Tokens**	**Types**	**Type-token ratio (TTR)**	**Standardized type-token ratio (STTR)**
PLTC	L2T	L2 learner translation	53	125,178	11,880	9.49	41.05
ICE-HK	L2W	L2 writing	53	127,835	12,704	9.94	41.47

Table 3: Composition of the corpora

2.2. Parameters and analysis

Following Laviosa (1998) and Ferraresi *et al.* (2018), we examined four parameters of lexical simplification: lexical density, standardized type-token ratio, core vocabulary coverage, and list head coverage. The operational definition of each parameter is stated in Table 4, below. In this study, several tools were employed to obtain the quantitative data. Both corpora were annotated using *Stanford CoreNLP* (Manning *et al.* 2014) to retrieve a part-of-speech tag for each word. It helped distinguish lexical words from running words. *WordSmith 8.0* (Scott 2021) automatically calculated the standardized type-token ratio (STTR) of each text, generated a list head word list for each corpus, and counted the number of core vocabulary and list head words of each text.

Parameters	Operational definitions (see Ferraresi *et al.* 2018)
Lexical density	It is used to evaluate the information load of a text. It is calculated by dividing the number of lexical words by the number of running words.[4] $= \frac{no.of\ lexical\ words}{no.of\ running\ words}$
Standardized type-token ratio	It is used to measure the lexical diversity of a text. It is obtained by calculating the ratio of the number of unique words to the number of running words on the basis of 1,000 words. $= \frac{no.of\ unique\ words\ (types)}{no.of\ running\ words\ (tokens)}$
Core vocabulary coverage	It is used to measure lexical diversity by exploring patterns of frequent word use of a text in comparison to an external reference. It is obtained in two steps: 1) by establishing a list of 200 most frequent words (core vocabularies) from a reference corpus —the written component of the *British National Corpus* (BNC; Leech *et al.* 2001) was selected (see Appendix 1), and 2) by calculating the proportion of core vocabularies to the number of running words. $= \frac{no.of\ core\ vocabularies}{no.of\ running\ words}$
List head coverage	It is used to measure lexical diversity by exploring patterns of frequent word use from an angle of internal corpus measure. Unlike Ferraresi *et al.*'s (2018) study, which performed the analysis at a sub-corpus level, this study measures frequent word use at the text level. This is achieved in two steps: 1) by creating a list with the 100 most frequent words (list head words) from each corpus examined in the study, namely L2T (See Appendix 2) and L2W (Appendix 3), respectively, and 2) by dividing the number head words of a text by the number of running words. $= \frac{no.of\ list\ head\ words}{no.of\ running\ words}$

Table 4: Operational definition of each parameter

Preliminary checks on the normality showed that the data of the core vocabulary coverage was normally distributed, so a paired t-test was run to examine if significant differences exist between the two corpora. However, the data of the remaining three parameters were not normally distributed, so Wilcoxon tests were used. Paired t-test and Wilcoxon tests were useful because L2T and L2W are related to each other. The source texts of L2T are the Chinese translation of L2W, that is, the translated texts of L2T are back translations of L2W. To illustrate the quantitative findings, some examples were extracted to supplement the quantitative results.

3. RESULTS

The descriptive data and differences between the two corpora in terms of the four parameters are summarized in Table 5. The data show that the lexical density of L2T (*M*

[4] Running words represent the total numbers of items. They are based on the unit of part-of-speech tagging, that is, each tagged word is regarded as one running word (excluding symbols, digits, and punctuations). Lexical words are nouns, verbs, adjectives, and open-class adverbs (those that end in *-ly*, except *only*).

= 57.77, *SD* = 4.51) is significantly higher than that of L2W (*M* = 57.27, *SD* = 4.47), *V* = 1014, *p* = .008. However, L2T is not significantly different from L2W in terms of standardized type-token ratio, core vocabulary coverage, and list head coverage. Standardized type-token ratio of L2T (*M* = 40.99, *SD* = 4.46) is slightly lower than that of L2W (*M* = 41.33, *SD* = 4.94). The core vocabulary coverage of L2T (*M* = 51.48, *SD* = 5.66) is slightly lower than L2W (*M* = 51.58, *SD* = 5.22). L2T (*M* = 47.21, *SD* = 3.86) also shows a greater but not significant list head coverage than L2W (*M* = 46.98, *SD* = 3.75).

Parameters	**L2T** (*n* = 53)		**L2W** (*n* = 53)		**Wilcoxon test**		**Paired t-test**	
	***M* (%)**	***SD***	***M* (%)**	***SD***	***V***	***p***	***t***	***p***
Lexical density	57.77	4.51	57.27	4.47	1014	.008*	/	/
Standardized type-token ratio	40.99	4.46	41.33	4.94	526.5	>.05	/	/
Core vocabulary coverage	51.48	5.66	51.58	5.22	/	/	-0.46	>.05
List head coverage	47.21	3.86	46.98	3.75	841	>.05	/	/

Table 5: A comparative analysis of L2T and L2W

The similarities and differences between L2T and L2W are illustrated in (1)–(3) below. Lexical density denotes information loaded on a text. Unlike function words, which mainly perform grammatical functions, lexical/content words carry semantic information. A high proportion of lexical words (emphasis added in the example) indicates that the text is packed with dense information. In example (1), L2T is lexically denser than L2W. In the example, L2T contains participial phrases, while L2W uses apposition in the second half of the sentence, resulting in lexical density differences between the two text varieties. A noun phrase in the apposition often needs a determiner (function word) to mark the noun, like '*an* insistence' and '*an* emphasis' (L2W), but a participial phrase does not. Also, when a noun phrase is modified by another noun phrase, a preposition (function word) is needed to express the modification, such as 'an emphasis *on* moral sensitivity' (L2W). For the participial phrase, a preposition is added after the participle only when the participle is an intransitive verb. In (L2T), 'emphasizing' is a present participle with a transitive nature which can be followed by a direct object, 'moral sensitivity,' without a preposition. Due to the above two reasons, the sentence in L2T is lexically more packed than in L2W.

(1) **L2T**	**L2W**
In both ethics, there seems to be a lack of universal rules or general principles, insisting that rules are not absolute and overriding everything, and emphasizing moral sensitivity above principles and moral reasoning.	In both ethics, there seems to be an absence of universal rules or general principles, an insistence that rules are not absolute and overriding, and an emphasis on moral sensitivity over principles and moral reasoning.

Standardized type-token ratio, core vocabulary coverage, and list head coverage are the measures of lexical diversity or variation. A high proportion of unique words and low repetition of common words indicate that the text is composed of a wide variety of vocabulary. Examples (2) and (3) show similar lexical diversity in all three parameters. This indicates that discrepancies between two sets of comparable sentences are not great enough to result in a significant difference in lexical diversity.

(2) **L2T**	**L2W**
In fact, she **was** just **trying** to **have a joke** on **the animal**, the most **innocent** kangaroo she had ever seen. She knew how fragile her life was, and she understood the rules of the forest. **With any** luck, if she **was not** eaten **today**, she **might** be eaten the next **day** by **some** careless animal or big bird. Anyway, she **thought**, **looking** up **at** the sky.	In fact, she just **wanted** to **play a trick** on **this creature**, the most **naive** kangaroo she had ever seen. She knew **exactly** how fragile her life was and she understood the rules of the forest. **By** luck if she **avoided being** eaten **one day**, she **could** be eaten the next, by **another** careless animal, or a big bird. Anyway, **as** she **was thinking**, **she looked** up **to** the sky.

(3) **L2T**	**L2W**
The **plight of** many of Hong Kong's elderly is a **worrying reflection of** a society that has traditionally **given** great care to the **elderly**. A study commissioned by the government on the **condition** of **older** citizens **produced disheartening results**. In Hong Kong, 30 percent of suicides involve the **elderly**, even though they make up only 14 percent of the population. On average, one elderly **person** is reported to **have committed** suicide every 1.5 days.	The **unhappy conditions in which** many of Hong Kong's elderly **live** is **cause for concern, and reflects poorly on** a society that has traditionally **taken** great care of the **aged**. A study commissioned by the Government on the **state** of **our senior** citizens **makes depressing reading**. Thirty percent of the suicides in Hong Kong involve the **aged**, even though they make up only 14 percent of the population. On **an** average, one **case of** elderly suicide is reported every 1.5 days.

By examining the examples closely, we find that major differences between them (emphasis added in the examples) can be categorized into grammatical factors (e.g., verb

tense and the use of function words) and non-grammatical factors (e.g., lexical word choice). Non-grammatical issues are likely to be the determining factors of lexical diversity. From a grammatical perspective, the expressions of verb tense and function words (e.g., prepositions and determiners) often follow certain rules. They may not lead to considerable differences in the use of the vocabulary in both texts. However, if we focus on the non-grammatical factors, it can be noticed that many lexical words have synonyms, hypernyms, and hyponyms. They allow for more variations in word choice. If a text is characterized by more synonyms and hyponyms, its lexical diversity will naturally increase. In example (2), 'creature' (L2W) and 'animal' (L2T and L2W) are synonyms, and 'kangaroo' (L2T and L2W) and 'bird' (L2T and L2W) are hyponyms of 'animal'. These words carry related meanings and allow for some variations in word choice. In (2), sentences in both L2T and L2W seldom repeat the same word, leading to a high diversity. Example (3) provides another instance in which synonyms are used instead of the repetition of the same word. In (3), various adjectives which describe negative emotions, i.e., 'worrying' (L2T), 'disheartening' (L2T), 'unhappy' (L2W), and 'depressing' (L2W), are synonyms. In example (3), both L2T and L2W are characterized by the use of synonyms instead of a single word to express ideas of similar meanings. Besides, (3) mainly reports the situations of older adults. Both L2T and L2W use different expressions to indicate this meaning in the example: 'elderly' (L2T and L2W), 'aged' (L2W), 'older citizens' (L2T), and 'senior citizens' (L2W). Both (2) and (3) show that L2T and L2W have a similar variety of lexical words, resulting in equally high lexical diversity.

4. Discussion

This study compared lexical simplification patterns between L2 learner translation (L2T corpus) and L2 writing (L2W corpus) using a corpus-based approach. The four lexical simplification parameters examined can be roughly classified into two broad categories: informativeness —i.e., lexical density— and lexical diversity —i.e., standardized type-token ratio, core vocabulary coverage, and list head coverage— (Ferraresi *et al.* 2018: 727; Xu and Li 2022: 10–11). The results demonstrate that learners' Chinese-to-English L2 translation is not lexically simpler than L2 writing. While there was no significant difference between the two corpora in all three lexical diversity parameters, L2 learner translation was found to be even lexically denser, i.e., more informative, than L2 writing.

Our study shows that simplification is not confirmed in learner translation when compared to L2 writing. The examples further show that lexical density may be related to the syntactic structure of the texts, while lexical diversity is likely associated with a variety of lexical/content words. In what follows, we will address the possible motivations for these findings from the perspectives of constrained communication, the language background of writers and translators, source language influence, and comparable corpus construction.

The degree of constraints may influence the lexical simplification patterns in communication. As mentioned in Section 1.2, the simplification hypothesis is regarded as one of the potential translation universals in comparison with non-translated native texts (Baker 1996). A possible explanation for this hypothesis is that translation is a form of constrained communication while non-translated native text production is not. Therefore, translated texts are simpler due to a higher cognitive load (Carl and Dragsted 2012) and risk minimization (Pym 2015) on behalf of the translator. Our study mainly focuses on the comparison of L2 writing and comparable translation done by student translators. Such a corpus design maximizes the degree of constraints in the way that language production in L2 becomes a major restriction shared in the two corpora. Our results show that L2 translation and L2 writing share more similarities than differences, highlighting the similar constraints faced by translators and writers who come from a similar background.

From a language background perspective, the texts of L2W and L2T are collected from Hong Kong writers and student translators. Hong Kong has a unique language environment due to its colonial history. Adding to Cantonese/Chinese (L1), English is also an official language in Hong Kong. English is a compulsory course for primary and secondary school students and remains the predominant language in professional settings, such as tertiary education, business, and law (Liu *et al.* 2022: 80). Therefore, Hong Kong English is often regarded as ESL rather than as EFL (Nasseri and Thompson 2021). Examples (1)–(3) also show how L2 writers and L2 learner translators in Hong Kong are able to use various synonyms to express ideas with similar meanings, which results in a variety of lexical words in the texts. Lexical diversity, an indicator of lexical simplification, positively correlates with language proficiency (Jarvis 2002; Crossley and McNamara 2012). We postulate that similar language proficiency and vocabulary

knowledge of the L2 writers and L2 learner translators narrows the differences in terms of lexical patterns.

From a source language perspective, the features of translational language can be subject to source language variation. Ferraresi *et al.* (2018) revealed that texts translated from Italian are lexically denser than original written texts. In contrast, similar results are not observed in the texts translated from French. Ferraresi *et al.*'s (2018) findings suggest that the source language can influence translation activity and alter the lexical density of a text. In our study, lexical density is the only parameter that distinguishes L2 learner translation from L2 writing, and the major difference between the two is the variable of source texts. Therefore, lexical density is likely to be subject to source language influence. Learner translators may be influenced by the source language (Chinese) to produce lexically denser texts than L2 writers. In addition, according to Laufer and Nation (1995: 309), lexical density "depends on the syntactic and cohesive properties of the composition. Fewer function words in a composition may reflect more subordinate clauses, participial phrases and ellipsis." Example (1), above, shows that participial phrases in L2T make a sentence more lexically packed. In short, the source language seems to play a critical role in the syntactic patterns of translations. Since the comparison at the syntactic level is not the focus of this study, further investigation is needed to uncover the relationship between lexical density and syntactic properties of L2 translation and L2 writing.

From the perspective of corpus construction, the degree of comparability may affect the comparison between the two corpora. As House and Kádár (2021: 4–5) argue, "[w]henever we use corpora compiled by others, we need to consider whether the generic, temporal and other features of the corpora are actually comparable." Also, "[i]n any rigorous […] research the size and other features of the corpora investigated need to be as comparable as possible." In traditional comparable corpus-based translation research, researchers mainly compile the corpora by considering the comparability of the genre and size. This study further enhances the comparability of the two corpora by ensuring their semantic sameness, as the texts of L2T are back-translated from that of L2W, that is, both originate from the same source. This may explain why examples (1)–(3) sometimes show similar text structure or vocabulary use. The corpus design may contribute to enhanced similarities between L2 learner translation and L2 writing in the findings. On the other

hand, since the corpora examined in this study are highly comparable, we can be more confident that their differences are likely due to translation factors.

5. Conclusion

This study provides a preliminary picture of the relationship between L2 learner translation and L2 writing through a lexical simplification prism. Our analysis can be summarized in three main points: 1) L2 learner translation is not lexically simpler when compared with L2 writing in the Hong Kong context, 2) lexical density of L2 learner translation is higher than that of L2 writing, and 3) L2 learner translation and L2 writing have similar lexical diversity. We have also discussed that factors such as the degree of constraints in communication, language background of writers and translators, source language, and comparable corpus design may play a part in the results. Through the comparison of L2 learner translation and L2 writing, the findings hint at how L2 learner translation might be influenced by the translation factor (reflected in the differences between the two corpora) and the L2 factor (reflected in the similarities between the two corpora). This can be important for enhancing translation learners' L2 translation competence.

Despite the findings, there are some limitations to the study. First, since the sample size is relatively small, the limited number of texts does not allow to analyze how register might be a possible factor in affecting the various simplification parameters. For future research on the topic, we plan to collect more texts in different registers and consider how register as a variable may affect lexical and syntactic simplification. Second, this study compared L2 learner translation with L2 writing only. Professional L2 translation needs also to be taken into account in order to address the simplification hypothesis properly. The comparison of learner and professional translations will show the extent to which the two differ in the lexical simplification parameters. Third, as lexical density is not only associated with the proportion of lexical words, but also with the syntactic structure of the texts, it is also worthwhile to examine syntactic structures to gain a better understanding of the simplification phenomenon underlying learner translation. All this represents an avenue for future research.

REFERENCES

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Francis Gill and Elena Tognini-Bonelli eds. *Text and Technology: In Honour of John Sinclair*. Philadelphia: John Benjamins, 233–250.

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7/2: 223–243.

Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers ed. *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Philadelphia: John Benjamins, 175–186.

Blum-Kulka, Shoshana and Eddie A. Levenston. 1983. Universals of lexical simplification. In Claus Færch and Gabriele Kasper eds. *Strategies in Interlanguage Communication*. London: Longman, 119–139.

Bolt, Philip and Kingsley Bolton. 1996. The International Corpus of English in Hong Kong. In Sidney Greenbaum ed. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press, 197–214.

Bowker, Lynne and Peter Bennison. 2003. Student translation archive: Design, development and application. In Federico Zanettin, Silvia Bernardini and Dominic Stewart eds. *Corpora in Translator Education*. Manchester: St. Jerome Publishing, 103–117.

Bulté, Bram and Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken and Ineke Vedder eds. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins, 21–46.

Carl, Michael and Barbara Dragsted. 2012. Inside the monitor model: Processes of default and challenged translation production. *Translation: Corpora, Computation, Cognition* 2/1: 127–145.

Chesterman, Andrew. 2004. Hypotheses about translation universals. In Gyde Hansen, Kirsten Malmkjær and Daniel Gile eds. *Claims, Changes and Challenges in Translation Studies*. Amsterdam: John Benjamins, 1–13.

Crossley, Scott A. and Danielle S. McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35/2: 115–135.

Duff, Alan. 1981. *The Third Language: Recurrent Problems of Translation into English*. Oxford: Pergamon Press.

Ferraresi, Adriano, Silvia Bernardini, Maja Petrović and Marie-Aude Lefer. 2018. Simplified or not simplified? The different guises of mediated English at the European parliament. *Meta* 63/3: 717–738.

Frawley, William. 1984. *Translation: Literary, Linguistic, and Philosophical Perspectives*. Newark: University of Delaware Press.

Gonzalez, Melanie. 2013. *The Intricate Relationship between Measures of Vocabulary Size and Lexical Diversity as Evidenced in Non-native and Native Speaker Academic Compositions*. Florida: University of Central Florida dissertation.

Granger, Sylviane and Marie-Aude Lefer. 2020. The Multilingual Student Translation Corpus: A resource for translation, teaching and research. *Language Resources and Evaluation* 54/4: 1183–1199.

Greenbaum, Sidney. 1988. A proposal for an international computerized corpus of English. *World Englishes* 7/3: 315. https://doi.org/10.1111/j.1467-971X.1988.tb00241.x.

Grosjean, François. 2013. Bilingualism: A short introduction. In François Grosjean and Ping Li eds. *The Psycholinguistics of Bilingualism*. Oxford: Wiley-Blackwell, 5–25.

House, Juliane. 2015. *Translation as Communication across Languages and Cultures*. London: Routledge.

House, Juliane and Dániel Z. Kádár. 2021. Introduction. In Dániel Z. Kádár and Juliane House eds. *Cross-Cultural Pragmatics*. Cambridge: Cambridge University Press, 1–12.

Hu, Kaibao. 2016. *Introducing Corpus-based Translation Studies*. Heidelberg: Springer.

Hu, Shirong. 2007. *A Corpus-based Study of the Translation Strategies Used in the Chinese Translations of Hamlet and Othello*. Shanghai: Shanghai Jiao Tong University dissertation.

Jantunen, Jarmo Harri. 2004. Untypical patterns in translations: Issues on corpus methodology and synonymity. In Anna Mauranen and Pekka Kujamäki eds. *Translation Universals: Do They Exist*. Amsterdam: John Benjamins, 101–126.

Jarvis, Scott. 2002. Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing* 19/1: 57–84.

Kortmann, Bernd and Benedikt Szmrecsanyi. 2009. World Englishes between simplification and complexification. In Thomas Hoffmann and Lucia Siebers eds. *World Englishes – Problems, Properties and Prospects*. Amsterdam: John Benjamins, 263–286.

Kruger, Haidee and Bertus van Rooy. 2016. Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide* 37/1: 26–57.

Lanstyák, István and Pál Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13/1: 99–121.

Laufer, Batia and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16/3: 307–322.

Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43/4: 557–570.

Leech, Geoffrey, Paul Rayson and Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Harlow: Longman.

Li, Defeng. 2002. Translator training: What translation students have to say. *Meta* 47/4: 513–531.

Liu, Kanglong and Muhammad Afzaal. 2021. Syntactic complexity in translated and non-translated texts: A corpus-based study of simplification. *PLOS ONE* 16/6: e0253454. https://doi.org/10.1371/journal.pone.0253454.

Liu, Kanglong, Joyce Oiwun Cheung and Nan Zhao. 2022. Learner corpus research in Hong Kong: Past, present and future. *Corpora* 17/Supplement: 79–97.

Lu, Xiaofei. 2012. The relationship of lexical richness to the quality of ESL learners oral narratives. *The Modern Language Journal* 96/2: 190–208.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing toolkit. In Kalina Bontcheva and Jingbo Zhud eds. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore: Association for Computational Linguistics: 55–60.

Mauranen, Anna. 2000. Strange strings in translated language: A study on corpora. In Maeve Olohan ed. *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects*. Manchester: St. Jerome Publishing, 119–141.

McWhorter, John H. 2011. *Linguistic Simplicity and Complexity: Why Do Languages Undress?* Berlin: Mouton De Gruyter.
Nasseri, Maryam and Paul Thompson. 2021. Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing* 47: 100511. https://doi.org/10.1016/j.asw.2020.100511.
Nelson, Gerald. 2006. *The ICE Hong Kong Corpus: User Manual*. London: University College London.
Olohan, Maeve and Mona Baker. 2000. Reporting *that* in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1/2: 141–158.
Pym, Anthony. 2008. On Toury's laws of how translators translate. In Anthony Pym, Miriam Shlesinger and Daniel Simeoni eds. *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury*. Amsterdam: John Benjamins, 311–328.
Pym, Anthony. 2015. Translating as risk management. *Journal of Pragmatics* 85: 67–80.
Saldanha, Gabriela. 2011. Emphatic italics in English translations: Stylistic failure or motivated stylistic resources? *Meta* 56/2: 424–442.
Scott, Mike. 2021. *WordSmith Tools Version 8.0*. Stroud: Lexical Analysis Software.
Toury, Gideon. 2012. *Descriptive Translation Studies – and Beyond*. Amsterdam: John Benjamins.
Tymoczko, Maria. 1998. Computerized corpora and the future of translation studies. *Meta* 43/4: 652–660.
Wen, Tinghui. 2009. *Simplification as a Recurrent Translation Feature: A Corpus-based Study of Modern Chinese Translated Mystery Fiction in Taiwan*. Manchester: University of Manchester dissertation.
Xia, Yun. 2014. *Normalization in Translation: Corpus-based Diachronic Research into Twentieth-century English-Chinese Fictional Translation*. Newcastle upon Tyne: Cambridge Scholars Publishing.
Xu, Cui and Dechao Li. 2022. Exploring genre variation and simplification in interpreted language from comparable and intermodal perspectives. *Babel* 68/5: 742–770.

Corresponding author
Kanglong Liu
The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
11 Yuk Choi Road
Hung Hom, Kowloon
Hong Kong SAR
China
Email: klliu@polyu.edu.hk

received: November 2022
accepted: February 2023

APPENDICES

Appendix 1: The 200 most frequent words extracted from the reference corpus. The written component of the BNC.

The	Would	Any	Go	Came
Of	Her	People	Man	Although
And	There	Should	Well	Few
A	n't	Than	World	Local
In	All	See	Same	Small
To	Can	Very	Most	Before
Is	If	Made	Life	Got
Was	Who	Like	Against	Social
It	Said	Just	Day	'll
For	Do	After	Might	Place
That	What	Between	Under	Case
With	One	Many	Here	Great
He	Its	Years	Does	Off
Be	Into	Way	Another	Always
On	Him	How	Come	've
I	Some	Our	Us	'm
By	Up	Being	Think	're
's	Could	Those	Old	Why
At	When	Such	While	Something
You	Them	Down	Never	Group
Are	So	Make	Where	Went
Had	Time	Through	Each	Want
His	Out	Over	Again	Thought
Not	My	Even	Found	Company
This	Two	Back	Mr.	End
Have	About	Must	Part	Party
But	Then	Know	Say	Per cent
From	No	Year	House	Women
Which	More	Own	Much	Next
She	Other	Still	Used	Both
They	Also	Because	Out of	Men
Or	Only	Too	Number	Find
An	These	Get	Without	Information
Were	Me	Good	Going	Important
As	First	Three	Different	Five
We	Your	Last	Children	Took
Their	May	Take	System	National
Been	Now	However	Put	Often
Has	Did	Government	During	Every
Will	New	Work	Within	State

Appendix 2: The most frequent words extracted from L2T

The	But	Mr.
Of	My	After
To	An	Them
And	Their	Because
A	We	Some
In	She	Years
Is	n’t	So
For	Also	Its
That	More	What
I	If	New
It	When	Than
Be	One	Into
’s	People	First
On	Do	Could
Are	There	Year
Hong	His	Out
Kong	Were	Like
Was	Me	Business
With	Government	China
As	Her	Services
By	Two	Up
This	Only	System
From	All	Most
He	Which	May
Not	Other	Many
At	Would	Any
Have	Who	Public
Will	Been	Did
Said	Had	Still
They	These	Such
You	About	However
Or	Time	Between
Can	Should	
Has	No	

Appendix 3: The most frequent words extracted from L2W

The	But	Should
Of	Their	No
To	Can	Other
And	One	Into
A	My	What
In	Were	Than
Is	We	After
For	n’t	Out
I	She	Years
It	More	Them
That	His	New
Be	Had	Could
On	Which	Because
As	Also	Some
With	Mr.	These
’s	When	First
Was	If	Such
By	Her	Business
Are	Me	May
Hong	There	Many
Kong	Do	Services
From	All	Did
Have	Would	Year
At	People	System
Not	Been	Most
He	Who	Says
You	Two	Public
Or	Government	Any
This	About	Now
An	So	Last
They	Only	Chinese
Will	Its	China
Has	Up	
Said	Time	

RiCL Research in Corpus Linguistics

A corpus-based study of embellishment in translations of the Newbery Medal Awards

Yu Zhai – Bin Xu
Peking University / China
Shandong Normal University / China

Abstract – Embellishment is a stylistic feature of translated children's literature. In recent years, children's reading choices and experiences have been truly thought highly of and, today, the idea that lexical enrichment is good for children —either for their writing or reading experience— is prevailing among children's literature translators and book editors. With this in mind, a small corpus composed of translations of the Newbery Medal Awards was built to figure out whether the phenomenon of embellishment exists in English-Chinese translations of children's literature and, if so, what are the motivations for it. The corpus includes six books selected on four criteria. The study suggests that embellishment is a typical feature of selected translations of the Newbery Awards and that it can be related to both book editing and the translator's own choices.

Keywords – corpus; embellishment; translation; Newbery Medal Awards; children's literature

1. Introduction

In *The Oxford Handbook of Translation Studies*, Lathey (2011: 198) noted that the translation of children's literature was claimed to be a sub-genre of the study of translation. In China, the study of translated children's literature had been largely ignored before the 1980s, when a large amount of children's literature was introduced via translation to fill that gap. Particularly, the analysis of children's literature translations of the Newbery Medal Award has been neglected. The Newbery Medal Awards are awarded annually to the author of the best American literature for children.[1] Books that have won the award are characterised by their diversified topics, including family, love, growth, or ecology, to name just a few. The books chosen to build a corpus of English-Chinese translated children's literature in the present study —namely *The Slave Dancer* (Fox 1974), *Waterless Mountain* (Armer 1932), *Young Fu of the Upper Yangtze* (Fore 1933),

[1] https://www.ala.org/alsc/awardsgrants/bookmedia/newbery

and *Where the Mountain Meets the Moon* (Grace 2010)— also address different topics and are set within different cultural backgrounds.

The term 'embellishment' is mentioned in an unpublished letter by Sharon Creech, the author of *The Wanderer* (2000), while her book was being translated independently by two translators whose versions were in the process of being selected by the chief editor. The letter was addressed to one of the translators, Professor Xu Bin, after he had told the translator that the editor would probably not choose his translation because of the tendency to prefer an enriched version of the original. In her letter, Creech critically used the terms 'enrichment' and 'embellishment' to refer to one of the translations of her original book and, as a result, the editor finally selected Xu's translation.

The phenomenon of embellishment may be considered a type of over-explicitation and may result from translational and editorial choices. The belief that high-quality Chinese children's literature is characterised by complex or flamboyant language may lead the translator and the editor to embellish and enrich the target text. This paper aims to answer two questions concerning the embellishment of translations. The first one is whether embellishment is a typical feature of selected translations of the Newbery Medal Awards. The second one concerns the reasons for the occurrence of embellishment, enrichment, or over-explicitation in the translations.

The paper is organised as follows. Section 2 provides information about the criteria chosen for building a parallel English-Chinese corpus and a description of the texts included in it. Section 3 offers a quantitative and qualitative analysis of the occurrence of embellishment in the corpus. Finally, Section 4 offers the conclusion.

2. BUILDING THE CORPUS

2.1. The theory of Corpus-based Translation Studies (CTS)

Baker (1993: 243) predicted that:

> the availability of large corpora of both original and translated text, together with the development of a corpus-driven methodology will enable translation scholars to uncover the nature of the translated text as a mediated communicative event.

For this reason, Baker suggested designing, building, and analysing different kinds of corpora: parallel, bilingual, multilingual, and monolingual comparable corpora. Her

proposals are often seen as the beginning of Corpus-based Translation Studies (Henceforth CTS). In 1996, the first CTS analysis was carried out at the University of Manchester Institute of Science and Technology (UMIST; Laviosa 2004).

2.2. Criteria in the selection of the texts

Works of literary fiction that have won the Newbery Medal Awards are various and include novels, poems, and short stories. For the research objectives, four criteria have been adopted for building the corpus:

1. Genre limitation: the selection of novels has been restricted to those that had won the Newbery Medal Award.
2. Availability: the novels have been published in China and both the source and the target text are available.
3. Diverse cultural backgrounds: books with different cultural backgrounds have been chosen to figure out the influence of the original texts.
4. Diverse translators: different translations of the same source text have been selected to figure out whether enrichment is influenced by the choices made by different translators.

The four novels which meet the corpus design criteria mentioned above are: 1) *Waterless Mountain* (Du 1932; Gao 1932), 2) *Young Fu of the Upper Yangtze* (Zhong 1933), 3) *The Slave Dancer* (Fu 1974; Li and Ying 1974), and *Where the Mountain Meets the Moon* (Zhang 2010). Further information about the authors, the citation, and the year of publication is provided in Table 1.

Year	Citation	Book	Author
2010	Honour	*Where the Mountain Meets the Moon*	Grace Lin
1933	Winner	*Young Fu of the Upper Yangtze*	Elizabeth Fore
1974	Winner	*The Slave Dancer*	Paula Fox
1932	Winner	*Waterless Mountain*	Laura Adams Armer

Table 1. The selected original books of the Newbery Medal Awards

Where the Mountain Meets the Moon and *Young Fu of the Upper Yangtze* are based on Chinese stories. Grace Lin, the writer of *Where the Mountain Meets the Moon*, is Chinese-American. The author of *Young Fu of the Upper Yangtze*, Elizabeth Fore, spent a long

time in China. By contrast, *The Slave Dancer* and *Waterless Mountain* are set within the background of Indian culture.

The corpus includes two versions of *The Slave Dancer* and two versions of *Waterless Mountain*. Table 2 shows the bibliographical details of the books selected. It is necessary to note that in Du's (1932) version of *Waterless Mountain* the translator is also the editor of the book. Moreover, one of the two translations of *The Slave Dancer* was carried out by two translators, Li Xinxin and Yu Ying. Both *Young Fu of the Upper Yangtze* and *Where the Mountain Meets the Moon* are set within a Chinese cultural background.

Year	English	Chinese	Translator	Publishing house
2010	*Where the mountain meets the moon*	月夜仙踪	Zhang Zizhang	Hebei education publication
1993	*Young Fu of the Upper Yangtze*	扬子江上游的小傅	Zhong Xiaoyu	Jiangsu children
1974	*The Slave Dancer*	"月光号"的沉没	Fu Dingbang	Chinese juvenile and children
1974	*The Slave Dancer*	月光之号	Li Xinxin and Yu ying	Hunan juvenile and children
1932	*Waterless Mountain*	荒泉山	Du Qingong	Tianjin people's fine arts
1932	*Waterless Mountain*	荒泉山	Gao Jie	Harbin

Table 2: Translations of the books selected in the corpus

2.3. Steps and tools

Building and designing a corpus usually involves three steps. First, scanning the original books into .txt files. Second, sampling the texts and cleaning the unrecognised codes. Third, segmenting the texts into token level or character level and POS-tagging them. In this study, the tools used for building the corpus include the website *Tmxall*[2] and software such as *TreeTagger*,[3] *NLPIR-parser*,[4] and *BFSU PowerConc*.[5] An English-Chinese parallel corpus and a reference corpus were used to analyse the phenomenon of embellishment. The latter is the *Original Chinese Children's Literature Corpus* (OCCLC; Zang 2010: 13). The former is the *English-Chinese Translational Children's Literature*

[2] https://www.tmxmall.com
[3] https://cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[4] http://www.nlpir.org/wordpress
[5] http://corpus.bfsu.edu.cn/TOOLS.htm

of the Newbery Medal Awards Corpus (TCLNC). What follows is the general description of these corpora.

OCCLC includes 14 works by six writers representing four periods. Unabridged texts are included with a total number of 265,266 Chinese characters, as shown in Table 3.

Text file	**OCCLC.txt**	**OC1.txt**	**OC2.txt**	**OC3.txt**	**OC4.txt**	**OC5.txt**	**OC6.txt**
File size	2,510,614	163,744	273,748	407,410	731,518	850,486	83,708
Tokens	265,226	19,096	27,986	42,986	71,511	93,586	10,101
Types	15,499	3,040	2,578	3,244	6,076	10,288	2,147
Type/Token Ratio	5.84	16	9	8	9	11	21
Standardised TTR	42.05	40.46	33.3	35.73	40.69	48.84	41.79
STTR basis	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Table 3: General figures of OCCLC

TCLNC was sampled from translations of four Newbery Medal Awards books by four authors and translated by six translators. Whole samples reach around 46,1745 Chinese characters, as illustrated in Table 4.

Text file	**TCLNC**	***Waterless Mountain* (Du Qingong)**	***Waterless Mountain* (Gao Jie)**	***Young Fu of the Upper Yangtze***	***The Slave Dancer* (Li and Yu)**	***The Slave Dancer* (Fu Dingbang)**	***Where the Mountain Meets the Moon***
File size	1,857,406	313,326	310,072	466,740	260,068	224,492	466,740
Chinese characters	461,745	79,307	78,830	121,042	63,549	55,949	121,042
Number of Tokens	319,082	55,595	54,503	80,393	44,754	38,989	80,393
Number of Types	16,326	5,471	5,383	9,740	5,373	5,108	9,740
Type/Token Ratio	5.12	9.84	9.88	12.12	12.00	13.10	12.12
Standardised TTR	46.86	44.59	44.37	50.30	47.42	47.28	50.30
STTR basis	1,000	1,000	1,000	1,000	1,000	1,000	1,000

Table 4: General figures in TCLNC

In addition, there is the *English Children's Literature of the Newbery Medal Awards Corpus* (ECLNC), which is aligned with the TCLNC at sentence level (see Table 5 below).

Text file	**ECLNC**	***The Slave Dancer***	***Waterless Mountains***	***Where the Mountain Meets the Moon***	***Young Fu of the Upper Yangtze***
File size	1,054,728	221,711	269,144	236,782	327,091
Tokens	195,682	41,322	50,652	43,757	59,951
Types	10,990	4,881	4,109	4,154	5,976
Type/Token Ratio	5.62	11.82	8.11	9.50	9.97
Standardised TTR	42.59	44.94	39.61	39.75	45.57
STTR basis	1,000	1,000	1,000	1,000	1,000

Table 5: General figure of ECLNC

3. Analysis of embellishment in TCLNC

Embellishment can be seen as a kind of over-explicitation at the lexical level. Translators and editors may tend to add modifiers in an originally simple sentence for various reasons, and such an addition may be a double-edged sword that can either cause misunderstanding or a better reading experience. It can be hypothesised that the number of embellishments may be influenced by the topic of the source text, the editor's choices, and the translator's choices. In this section, we aim at establishing whether embellishment occurs and, if it does, at calculating the ratio of embellishments in the first 100 sentences in each sample.

3.1. Standardised type/token ratio

Tokens are the running words in a text. The higher the number of tokens, the larger the size of a text. Type refers to any word form in a text. In other words, if the same token appears in a text repeatedly, it can only be counted as one type. The type/token ratio (TTR) can, to some degree, reveal the diversity and richness of the language used by the author or the translator (Baker 2007: 50). However, the different lengths of texts may influence the type/token ratio in different ways, therefore a standardised type/token ratio (STTR) is required.

The calculation of STTR in the two corpora shows that the total STTR of TCLNC is 46.86, which is 4.81 higher than that of OCCLC, but is lower than 42.59 in the source texts. Therefore, words are more diverse in TCLNC than in OCCLC. This is in line with Wang and Qin (2009: 105), who state that "Chinese translational texts have a richer and more diverse manner of using lexical items than Chinese original texts and their English source texts."

3.2. The phenomenon of over-translation type/token ratio

Not only does the type/token ratio reveal features of the translator's language style, but also the ratio of words or characters of target texts to source texts does so. According to Wang (2003: 415), in English-Chinese literary works, the ratio of words or characters of translated texts to source texts is 1:1.55~1:2.1, the common ratio of them is 1:1.65~1:2.1, and the intermediate value is 1.79. If the ratio is larger than the intermediate value, it can be said that the translated text has a tendency towards over-translation. Therefore, the higher the ratio is, the closer it will be to over-translation, and vice versa.

Although the size expansion of a translated text is unavoidable, the criteria provided by Wang (2003) give researchers the momentum to believe that there is a close relation between embellishment and over-translation. It can be assumed that with the rise of the text size ratio, the frequencies of embellishment might also grow. As the data in Table 6 show, one translated text exceeds the common ratio of 1:1.65~1:1.9, let alone the intermediate value of 1:1.79 postulated by Wang (2003) in the analysis of his English-Chinese literature translations corpus. We will assume that embellishment arises from the over-translation phenomenon. Thus, we could relate the embellishing phenomenon to the over-translation feature.

Book	***Waterless Mountain* (Du Qingong)**	***Waterless Mountain* (Gao Jie)**	***Young Fu of the Upper Yangtze***	***The Slave Dancer* (Li and Yu)**	***The Slave Dancer* (Fu Dingbang)**	***Where the Mountain Meets the Moon***
Ratio	1.57	1.56	2.04	1.54	1.36	1.47

Table 6: Ratio of the size of the text

3.3. Frequencies of embellishment and examples of over-translation type/token ratio

Øverås (1998) conducted a study dealing with the hypothesis about explicitation. She built bilingual corpora of English and Norwegian and analysed the first 50 sentences in 40 fictional texts (20 English translations and 20 Norwegian translations). The results showed that explicitation was more frequent in the texts translated into Norwegian than in the texts translated into English. In this section, we follow Øverås' (1998) procedure and show the ratio of embellishment of the six translated texts when compared to their source texts. Since the feature of embellishment is mainly based on lexical addition, the first 100 sentences of each text are analysed at the lexical level.

For each text, we show two typical examples and each example consists of: 1) English sentences from the source texts ([S]), 2) Chinese sentences from the target texts ([T]), and 3) English sentences of back translation ([BT]). The analysis is mainly based on TCLNC, and the frequencies of embellishment in the six texts are shown in Table 7 below.

Book	***Waterless Mountain* (Du Qingong)**	***Waterless Mountain* (Gao Jie)**	***Young Fu of the Upper Yangtze***	***The Slave Dancer* (Li and Yu)**	***The Slave Dancer* (Fu Dingbang)**	***Where the Mountain Meets the Moon***
Ratio	11	7	9	6	4	0

Table 7: Frequencies of embellishment in the texts

The comparison of the data in Tables 6 and 7 seems to suggest that a high tendency towards over-translation is not the only sign that may announce the rise of the phenomenon of embellishment and that both over-translation and embellishment do not have necessary connections. This is shown, for example, in the comparison of *Where the Mountain Meets the Moon* (Zhang) and *The Slave Dancer* (FU). Firstly, both size ratios are within the common scope of literature size ratio, with the latter being much lower than the former. Nevertheless, when it comes to the frequency of embellishment, the latter has a much higher number, while the frequency of embellishment of the former is 0.

If we compare the two translations of *Waterless Mountain* in Tables 6 and 7, it can also be concluded that second editing might cause the increase in the frequency of embellishment. Based on the data, Du Qingong, who translated and edited the translation of Gao Jie, made the text more embellished.

In examples (1)–(4) below, taken from *Waterless Mountain* (Gao) and *Waterless Mountain* (Du), the bold characters are the embellishment of the translated texts. They share the same feature, which is having no equivalent words in the source texts. In (1), the author might not mean to describe how handsome the elder brother is, but just tries to introduce the differences between a young man and a grown man. However, for unknown reasons, the translator chooses to polish the description to show how handsome he might be. It will not cause great damage to add how handsome the character might be, but when the embellishment distorts the original meaning, readers might have negative judgments of some should-be-good characters. For example, this may be the case with the uncle in (2), who always told stories in the winter, and with the father in (3), who had changed his job due to acceptable reasons. In turn, example (4) is the most confusing one. A mom

could stand at the corral with any kind of facial expression waiting for her child, but one may wonder about the addition of the phrase ‘with a smile’. A possible motivation for this is that the translator or editor wanted to describe a warmer family, but such an addition might not provide evidence for what they wanted. At the same time, the addition might deprive readers of their wish to appreciate the author’s original written style.

1. [S-WM] Elder Brother wore his long hair in a knot because he was a grown man.

   [T-Gao] 哥哥已经是成年人了，飞扬的长辫为他**平添了一份英气**。

   [BT-Gao] Elder Brother is a grown man now, the long hair flying on his back **has added some handsomeness to him**.
   [T-Du]哥哥已经是成年人了，飞扬的长辫看起来十分英俊。

   [BT-Du] Elder Brother is a grown man now, the long hair flying on his back **looks very handsome**.

2. [S-WM] He told stories in the wintertime while everyone sat around the fire in the middle of the hogan.

   [T-Gao] 一到冬天，大伙儿都**陆陆续续地**围坐在泥房子中间的营火旁边。每到这会儿，舅舅就**打开了话匣子，将肚子里的新鲜事儿**一件件地说给大家听。

   [BT-Gao] When winter has come, people would cuddle around by fire in the middle of the hogan **one by one**. Uncle would **open his chat box then, telling all the fresh stories** to them one after another.

   [T-Du] 冬天一到，大伙儿都**陆陆续续地**围坐在泥房子中间的营火旁边。每到这时候，舅舅就**打开了话匣子**，将肚子里的新鲜事儿**滔滔不绝地**说给大家听。

   [BT-Du] When winter has come, people would cuddle around by fire in the middle of the hogan **one by one.** Uncle would **open his chat box then, telling all the fresh stories** to them with **an unceasing flow of words**.

3. [S-WM] When he tired of making bracelets and rings, he rode about the desert to look after his cattle.

   [T-Du] 他**心血来潮**不再做打制银手镯、银戒指的工作，而**任性地**骑着马去荒漠担任了一个自在马倌的职位。

   [BT-Du] He suddenly **felt some blood wave flowing in his heart** and would not like to be making **silver** bracelets and rings, but rode about the desert to look after his cattle **petulantly**.

   [T-Gao]他心血来潮放弃了打制银手镯、银戒指的工作，而索性骑着马去荒漠担任了一个自在马倌的职位。

[BT-Gao] He suddenly **felt some blood wave flowing in his heart** and would give up making **silver** bracelets and rings, but **simply** rode about the desert to look after his cattle.

4. [S-WM] Mother met him at the corral and helped him put up the bars.

   [T-Du]妈妈**笑着**站在羊圈旁等他，接过他手上的皮鞭，替他关好羊圈

   [BT-Du] Mother waited him at the corral **with smile,** taking the rope from his hands and put up the sheep bars.

   [T-Gao]妈妈**笑呵呵地**立在羊圈旁，接过他手上的皮鞭，帮他关好羊圈。

   [BT-Gao] Mother waited him at the corral **with laughter,** taking the rope from his hands and put up the sheep bars.

The comparison of the translations of *The Slave Dancer* in Tables 6 and 7 shows that the frequency of embellishment may vary depending on the translators. The author of *The Slave Dancer* has a comparatively high ability to choose complex words and describe sceneries. It is hard not to follow the author's style and enrich the text. However, Fu Dingbang seems to have better self-control regarding embellishment than Li Xingxing and Yu Ying. Yet both show some tendency towards embellishing the text. This is illustrated in (5) and (6). In example (5), 'a garden of flowers' in the source text meant to describe the thread, while Li embellished this nominal description with verbs 'rush to bloom'.

5. [S-SD] By candlelight, the warmth of the colors made me think the thread would throw off a perfume like a garden of flowers.

   [T-Li]柔和的烛光下，暖烘烘的缤纷色彩让人觉得这些线团会散发出扑鼻的香气，仿佛满院子的花儿**争相绽放**。

   [BT-Li] By soft candlelight, the warmth of the colors made me think the thread would throw off a perfume like a garden of flowers **rush to bloom**.

   [T-Fu] 在柔和的烛光下，各种漂亮的颜色也变得暖烘烘的，仿佛正在向外散发香气，就像花园里的鲜花一样。

   [BT-Fu]By soft candlelight, **beautiful colors** turn to be warm, like a garden of flowers throw off a perfume.

6. [S-SD] I had seen damask and gauze and velvet and silk...

   [T-Li] 我见过**溜滑的**缎子、**蝉翼般的**薄纱、**柔软的**天鹅绒、**轻柔光滑的**绢丝……

   [BT-Li] I had seen **smooth** damask, gauze **like the wings of cicadas**, **soft** velvet, **light and smooth** silk...

   [T-Fu] 上好的料子我看过很多，锦缎、薄纱、天鹅绒、丝绸都有……

   [BT-Fu] I had seen many fine materials such as damask and gauze and velvet and silk ...

If, in examples (1)–(6) and example (9), the addition of modifiers is considered a form of embellishment, what is added in examples (7) and (8) are mental activities. Such a kind of embellishment is no longer confined to the lexical level, but also involves the sentence level. In example (6), the original text only lists the name of the materials. However, the target text does not only list the names, but also describes the tactile of the material. The embellishment may be able to convey the translator's knowledge and feeling about those clothes, but the description is comparatively redundant and difficult to control. In example (7), an additional sentence is added in the translation, which further explains why the mother tells them they should feel fortunate. However, this kind of embellishment seems to be redundant in both syntax and meaning.

7. [S-SD] Then my mother would mention how fortunate we were to live in New Orleans where we did not suffer the cruel extremes of temperature that prevailed in the north.

   [T-Fu]**然而每每**这**个**时**候**，妈妈**却告**诫**我**们**要**对现**在在新奥**尔**良的生活心**怀**感激**，这**里不像北方有寒冷侵**袭，**更何况还经常是晴天**。

   [BT-Fu] Then my mother would admonish us to be grateful for our life in New Orleans where we did not suffer the cruel extremes of temperature that prevailed in the north, **not to mention the fact that it's often sunny.**

   [T-Li]**每当这时，妈妈就会跟我们絮叨，说幸亏我们住在新奥尔良，不然就得忍受北方的严寒酷暑。**

   [BT-Li] Then my mother would say that we are lucky to live in New Orleans where we did not suffer the extremely cold winder and hot summer that prevailed in the north.

8. [S-SD] I had never heard anyone called such a name before.

   [T-Fu]**我**对这样**一个名字念念不忘，我从未想**过**有人会叫星星。**

   [BT-Fu] I **couldn't get such a name out of my head**, and I never thought anyone called Xingxing before.

   [T-Li]以前从没有听过哪个人叫这个名字。

   [BT-Li] I had never heard anyone called such a name before.

9. [S-SD] I imagined the splendid house I would live in, my gardens, my carriage and horses.

   [T-Fu] 我尽情地幻想，幻想自己金碧辉煌的房子，幻想拥有自己的花园、马车还有马儿。

   [BT-Fu] I imagined the splendid house I would live in, my gardens, my carriage and horses.

   [T-Li]我还有一处华丽无比的住所，有**漂亮的**花园，有专属马车和马。

   [BT-Li] I imagined the splendid house I would live in, my **beautiful** gardens, my exclusive carriage and horses.

The comparison of *Young Fu of the Upper Yangtze* and *When the Mountain Meets the Moon* in Tables 6 and 7 also shows that fictional background cannot determine the frequency of the embellishment in translations. Though they share the same story background of China, the frequency of embellishment differs in both texts. This difference may be noticed in the redundancy test, which refers to the size ratio in Table 6, and it is the frequency of embellishment that makes it more obvious, as illustrated in (10)–13), below. In (10), 'told of its wonders' is translated as "*每每吹嘘起城里的奇观种种，就吐沫横飞、滔滔不绝* ('talk like waves that won't stop with their spittle flying everywhere'). Despite the translation of 'wonder', the translator extends the meaning of 'talk' into 'brag about' and adds two groups of four-character idioms *吐沫横飞* ('spittle flying everywhere') and "滔滔不绝 ('talk like waves').

10. [S-YF] In his village men who counted it a privilege to visit this city once in a lifetime had told of its wonders.

    [T-Yu]**在他**们**村里，那些**亲**身到**过重庆**的人无不把**这**种**经历视为**极大的殊荣，每每吹嘘起城里的奇**观**种种，就吐沫横飞、滔滔不绝。**

    [BT-Yu] In their village, those who have been to Chongqing in person regard this experience as a great honor. Whenever they brag about the wonders of the city, they would **talk like waves that won't stop with their spittle flying everywhere**.

In example (11) below embellishment is illustrated in that the preposition ‘in’ is translated as 一窝蜂地涌到 (‘in some place like a swarm of bees’), but the source text never mentions information about how they got in the theatre and tea houses. Unlike example (11), in (12) three phrases are added to the adverb ‘tortuously’, namely, ‘winds its way’ (*蜿蜒蜒蜒*), ‘for hundreds and thousands of times’（*百转千回*) and ‘goes on and on’（*源源不断*）. In Chinese, it is common for writers to use four-character idioms to polish the text. Example (13) exhibits the same type of embellishment as examples (7)–(8) and adds a redundant explanation to the source text. In (13), ‘beneath’ means that Fu’s mother thought wood or bamboo is doubtlessly below the plasters, therefore there were cracks and holes which made the place not suitable to live. The translator adds a redundant explanation *外面用灰泥一涂了事*（‘Plasters cover the outside and everything will be finished’）. This kind of embellishment enlarges the size of the target text, but does not make the text lose its original meaning. Even though the phenomenon of embellishment is not rare, the risk of twisting the original writing style still exists.

11. [S-YF] when there is time for play, enjoy themselves in handsome tea houses and theaters.

    [T-Yu]等闲下来了，就**一窝蜂地**涌到漂亮的茶馆和戏院里找乐子。

    [BT-Yu] When there is time for play, they **flock to** the beautiful teahouses and theaters like a swarm of bees for fun.

12. [S-YF]to the east, its main artery of life, the Yangtze-kiang, flowed tortuously for fifteen hundred miles before it reached Shanghai and the coast and emptied its muddy stream into the blue Pacific.

    [T-Yu]东面是重庆的生命主干道扬子江，**江水蜿蜿蜒蜒、百转千回，连绵几千里**直奔上海，把浑浊的河水**源源不断地**送入蔚蓝色的太平洋。

    [BT-Yu] To the east is the Yangtze-kiang River, the main lifeline of Chongqing. The river flows tortuously, **winds its way for hundreds and thousands of times, goes on and on for thousands of miles** before it reached Shanghai, **continuously** sending a steady stream of muddy water into the blue Pacific.

13. [S-YF] Wood or bamboo is doubtless beneath, but that will make it no better a place in which to live.

    [T-Yu]保不定这墙就是用木头或竹子做的基，**外面用灰泥一涂了事**，那可就彻底没法住人了！

> [BT-Yu] Wood or bamboo is doubtless beneath. **Plasters covers the outside and everything will be finished**. But that will make it no better a place in which to live!

4. CONCLUSION

The analysis presented here shows that the phenomenon of embellishment exists in a sample of translations selected from the Newbery Medal Awards. At least in TCLNC, five out of the six books analysed contain occurrences of embellishment and enrichment. Despite the frequency of zero books, *The Slave Dancer* —translated by Du Qingong— has the highest frequency of occurrences, while *Where the Mountain Meets the Moon* — translated by Zhang Zizhang— has the lowest frequency. There is also one sample with a frequency of zero, and this is due to the samples that were selected.

This study has also contributed to a new perspective as regards the over-exploitation study of translation theory by borrowing the term 'embellishment' from the source text author. Furthermore, the data have shown that embellishment seems to have little to do with the texts' redundancy and the fiction's story background, but is rather related to the editing and the translator's own choices. The data in the study (see Tables 6 and 7) show that the text with the highest size ratio (*Young Fu of the Upper Yangtze*) does not exhibit the highest frequencies of embellishment. In the sample texts analysed here, there is no evidence that the translators' cultural background could give them the confidence to embellish the texts. However, if the language in the source texts is rich and diversified, the translated texts may unavoidably be characterised by the phenomenon of embellishment and enrichment. It seems sensible to state that translators and editors should consider both target readers and the author when it comes to book translation. Translators may be suggested to firstly make sure that they convey the basic meaning of the source text in their translations. It may be sensible to state that the number of embellishments may increase if editors or translators pay less attention to the accurate delivery of the message between the author and target readers. Embellishment may indeed improve the target reader's reading experience, but it may also have the potential of distorting the original style or intention by the author. This research, however, has only focused on applying a corpus-based methodology to analyse the phenomenon of embellishment in existing translations. Further research is required to analyse the true feeling of target readers and the motivations for the embellishment of texts by editors or even if the source text shows a simple and concise style.

REFERENCES

Armer, Laura Adams. 1932. *Waterless Mountain*. New York: Dover Publications.

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Gill Francis and Elena Tognini-Bonelli eds. *Text and Technology: In Honor of John Sinclair*. Amsterdam: John Benjamins, 233–250.

Baker, Mona. 2007. Corpus-based translation studies in the academy. *Journal of Foreign Languages* 5: 50–55.

Creech, Sharon. 2000. *The Wanderer*. New York: Harper Collins.

Du, Qingong. 1932. *荒泉山* (*Waterless Mountain*). Hong Kong: Tianjin People's Fine Arts Publishing House.

Fore, Elizabeth. 1933. *Young Fu of the Upper Yangtze*. London: Square Fish Press.

Fox, Paula. 1974. *The Slave Dancer*. New York: Simon & Schuster Press.

Fu, Dingbang. 1974. *月光号"的沉没* (*The Slave Dancer*). Shanghai: Chinese Juvenile and Children's Publishing House.

Gao, Jie. 1932. *荒泉山* (*Waterless Mountain*). Harbin: Harbin Publishing House.

Grace, Lin. 2010. *Where the Mountain Meets the Moon*. Boston: Little, Brown and Company Press.

Lathey, Gillian. 2011. The translation of literature for children. In Kirsten Malkmjaer and Kevin Windle eds. *The Oxford Handbook of Translation Studies*. Oxford: Oxford University Press, 198–214.

Laviosa, Sara. 2004. Corpus-based translation studies: Where does it come from? Where is it going? *Language Matters* 35/1: 6–27.

Li, Xinxin and Yu Ying. 1974. *月光之号* (*The Slave Dancer*). Hunan: Hunan Juvenile and Children's Publishing House.

Øverås, Linn. 1998. In search of the third code: An investigation of norms in literary translation. *Meta* 43/2: 571–588.

Wang, Kefei. 2003. Sentence parallelism in English-Chinese/Chinese-English: A corpus-based investigation. *Foreign Language Teaching and Research* 35/6: 410–416.

Wang, Kefei and Qin Hongwu. 2009. A parallel corpus-based study of general features of translated Chinese. *Foreign Language Research* 1: 102–105.

Zang, Guangya. 2010. *A Corpus-based Study of Language Use in Translated Chinese Children's Literature*. Qufu: Qufu Normal University.

Zhang, Zizhang. 2010. *月夜仙踪* (*Where the Mountain Meets the Moon*). Hebei: Hebei Education Publication

Zhong, Xiaoyu. 1933. *扬子江上游的小傅* (*Young Fu of the Upper Yangtze*). Nanjing: Jiangsu Children's Publishing House.

Corresponding author
Yu Zhai
Peking University
School of Software and Microelectronics
No.5 Yi He-yuan Road
100091, Beijing
China
Email: zhaiyu@stu.pku.edu.cn

received: November 2022
accepted: August 2023

RiCL Research in Corpus Linguistics

Multilingual parallel corpus: An institutional resource for terminology development at the University of South Africa (Unisa)

Koliswa Moropa – Bulelwa Nokele
University of South Africa / South Africa

Abstract –The indigenous African languages of South Africa are not fully developed to provide for specialised terminology and were considered unsuitable for use as languages of tuition and research. This was used as a scapegoat for not utilising these languages in the South African education system. Since 1994, however, terminology development has been one of the key priorities of democratic South Africa. The institutions of Higher Learning have been mandated to develop and intellectualise the indigenous languages for teaching, learning and research. In line with this, this article aims to address the problem of unavailability of scientific or technical terms by illustrating how a multilingual corpus —from which multilingual glossaries as resources for tuition and research— can be compiled. Adopting a qualitative descriptive approach, suitable source texts in English and their translations in various African indigenous languages, namely, IsiZulu, IsiXhosa, IsiNdebele, SiSwati, Tshivenda, and Xitsonga were selected from the University study material for inclusion in the multilingual parallel corpus. *ParaConc*, a software that is suitable to query parallel texts, was used to align and extract terms from the corpus. The study demonstrates how parallel texts can be useful in developing scientific and technical terms. The University of South Africa can become the centre of corpus compilation for the intellectualisation of the official indigenous South African languages, since it is the only university in the country that caters for all these languages.

Keywords –corpus; corpus compilation; terminology development; multilingual parallel corpus; indigenous languages

1. INTRODUCTION

This article provides information on the main topics to be considered when designing a corpus and tries to answer the question of why it is imperative for the University of South Africa (Unisa) to move towards corpus compilation to promote terminology development. Commenting on the important issue of language development and the Unisa Language Policy (2016), Alexander (2003: 18) observes that:

> [m]ere language planning cannot bring about the fundamental shifts in consciousness and in behaviour which are necessary to lift the indigenous languages of Africa to a different historical trajectory.

Alexander finds it illogical to believe that it is possible to think of an African renaissance without the development and intellectualisation of the South African indigenous languages. He calls on Higher Education Institutions (HEIs) to participate in facilitating and promoting the development of these languages in such a manner that they can be used in all official functions as formal academic languages in HEIs.

Since 1994, language development has been one of the key priorities of the democratic South Africa. The Constitution of the Republic of South Africa, Act 108 of 1996, brought about changes regarding the status of indigenous African languages, declaring nine of them to be official. Section 6 (1) of the Constitution stipulates that "the official languages of the Republic of South Africa are Sepedi, Sesotho, Setswana, Tshivenda, Xitsonga, Afrikaans, English, SiSwati, IsiNdebele, IsiXhosa, and IsiZulu." Before that, the education policies of the apartheid regime neglected the indigenous languages, recognising only Afrikaans and English as official languages. Likewise, the colonial bilingual education system marginalised all other indigenous languages and, therefore, the technical/scientific registers of these languages remained underdeveloped (Moropa and Shoba 2017).

1.1. Language Policy for Higher Education

A milestone in the commitment to language development and promotion of multilingualism in institutions of Higher Learning was the *Language Policy for Higher Education* (LPHE; Department of Education 2002), which was revised in Department of Education (2020). The revised version of the policy framework declares that since the propagation of the LPHE in 2002, little progress has been made in exploring and exploiting the potential role of indigenous African languages in facilitating access and success —as well as in the intellectualisation of these languages— in Higher Education. LPHE requires HEIs to develop language policies together with implementation plans.

In 2016, Unisa amended its language policy, and in 2017 it adopted the implementation plan entailing that, as a national university a) it acknowledges that there are eleven official languages in South Africa and ensures that, together with South African Sign Language, they enjoy parity of esteem and equitable treatment (Section 4.1.1), and b) it endeavours to support all the official languages of South Africa (Section 4.1.6).

As an "African university, in the service of humanity, shaping futures," Unisa commits to building capacity for all official South African languages (Unisa Language Policy 2016, Section 4.1.6). The initiative of corpus compilation as a resource for terminology development is thus a step towards promoting the indigenous African languages as languages of teaching, learning and research at Unisa, and this constitutes one step towards fulfilling this obligation.

1.2. The research problem and aim

Different scholars have argued that one of the reasons hindering the development and intellectualisation of African languages is the lack of terminology (cf. Madiba 2001; Gauton and De Schryver 2004). Different intervention strategies have been suggested to sort this out. Madiba (2001) suggests a pragmatic approach which entails borrowing and indigenisation of terms from source languages. Moropa (2005) recommends the use of the concordancer *ParaConc*[1] for term identification and term creation strategies when dealing with technical texts. Addressing the lack of specialised quadrilateral dictionaries, Mlambo *et al.* (2021) also use *ParaConc* to identify lexical items in English and their equivalents in Xitsonga, SiSwati and IsiNdebele, to compile a quadrilateral dictionary. Their broader aim is to facilitate communication across languages and development of minority languages, which will ultimately promote multilingualism in South Africa. These are some studies that have not only voiced concern about the lack of terminology in the indigenous languages of South Africa, but have also proposed solutions to overcome this problem.

As the largest —and the only— university that serves all language communities in South Africa, Unisa is well positioned to contribute towards the noble mandate of developing all previously disregarded languages to end the negative narrative of lack of terminology. Therefore, the University has embarked on a project involving translating study material from English into all the South African official languages. The translated material is part of the corpus that could be used to extract and/or develop specialised terminology. It is also imperative to note that the translation of these study materials is not only for mere availability in the various indigenous African languages, but for access of information in a language that is understandable to the reader. That is why paraphrasing

[1] https://paraconc.com/

is one of the popular translation strategies that has been used. Set within this background, the present article seeks to illustrate how a corpus can assist in the development and intellectualisation of the indigenous languages.

The paper is structured as follows. Section 2 discusses the notion of 'corpus' and how corpora are being compiled in South Africa, whereas Section 3 offers a qualitative discussion of the results. Section 4 concludes the article.

2. CORPUS COMPILATION

2.1. Definition of 'corpus' and types of corpora

Originally, the term 'corpus' (plural 'corpora') meant any collection of writings in a processed or unprocessed form by a specific author. According to Baker (1995: 225), with the advancement of corpus linguistics, this definition changed in three important ways:

1. Nowadays, a corpus is primarily a collection of texts held in machine-readable form, which can be analysed (semi)-automatically in a variety of ways.
2. A corpus is no longer restricted to written texts but can also include spoken texts.
3. A corpus may include a number of texts from a variety of sources by many writers and speakers on a multitude of topics.

What is important is that a corpus is assembled for a *particular purpose* (emphasis added) and according to explicit design criteria to ensure that it is *representative* (emphasis added) of the given area or sample of language for which it aims to account for.

In his classification, Sinclair (1995) distinguishes various types of corpora such as reference, monitor, parallel and comparable corpora. A parallel corpus is a "collection of texts, each of which is translated into one or more other languages than the original" (Sinclair 1995: 32). A parallel corpus can be bilingual when it comprises original texts and their translated versions of the same source language. For example, in the South African context it can be English Source Text (original) [ST] – IsiZulu Target Text (translated) [TT]; English [ST] – Tshivenda [TT], or *vice versa*. The parallel corpus is multilingual, as it contains translations into several target languages. For example, tutorial letters in various disciplines are translated from English into other official South African languages to make the texts accessible to all (e.g., English [ST – Setswana –IsiNdebele – Xitsonga [TTs], etc.).

2.2. Corpus-based research for terminology development in South African languages

Corpus-based research for terminology development in various indigenous African languages has been widely conducted in South Africa. Madiba (2004), Gauton and De Schryver (2004), Moropa (2005, 2007), Ndhlovu (2016), and Shoba (2018) have shown how specialised corpora can be used to develop indigenous African languages and retrieve terminological information. Madiba (2004), who analyses the *Special Language Corpora for African Languages* (SPeLCAL), illustrates how parallel corpora can be used as tools for developing the indigenous languages of South Africa. The SPeLCAL project was born out of the need for language resources to support the implementation of South Africa's multilingual language policy adopted after the democratic changes of 1994. Madiba (2004) makes use of *Multiconcord*[2] to analyse translation equivalents of terms such as 'act', 'legislation', 'rule', 'order', and 'law' in a parallel corpus of English-Tshivenda texts of The Constitution of the Republic of South Africa (1996). His findings revealed that translation equivalents, as well as inconsistencies in the translation, can be identified and that the SpeLCAL corpora could be useful for terminographers and lexicographers.

Gauton and De Schryver (2004) demonstrate how special-purpose multilingual and parallel corpora can be used as a translator's tool in finding suitable equivalent terms when translating technical texts from English into Zulu. In their study they make use of a) the *University of Pretoria Zulu Corpus* (PZC), an electronic corpus compiled at the University of Pretoria which comprises literary texts, religious texts, internet files, and pamphlets in Zulu and a total amount of five million words (first case study), and b) the *University of Pretoria Internet English Corpus* (PIEC), an electronic corpus of 12.4 million English words retrieved from online sources (second case study.)[3] In the first case study, multilingual corpora are used to investigate the terminology used in the translation of HIV/AIDS texts. HIV/AIDS terminology was identified in both corpora by resorting to the *KeyWords* function in *WordSmith Tools*.[4] In the second case study, parallel corpora dealing with labour issues are scrutinised to investigate labour and determine the usefulness of such corpora as a resource for the translation of technical texts into Zulu.

[2] https://artsweb.cal.bham.ac.uk/pking/multiconc/l_text.htm
[3] https://sadilar.org/index.php/en/
[4] https://lexically.net/downloads/version5/HTML/index.html?keywords_start.htm

Moropa (2004, 2005, 2007) also investigates how corpus-based research may contribute to the development of strategies for translating financial and technical texts into IsiXhosa. Moropa's research illustrates the benefits of parallel texts and computer tools, such as *ParaConc*, in the development of terminology and also their usefulness for translators as terminology developers. Proper alignment and good quality translations are also emphasised for translations to be used as resources for terminology development.

Along the same lines, Ndhlovu (2016) resorts to the *English-Ndebele Parallel Corpus* (ENPC; Ndhlovu 2012) to extract bilingual terminology for the creation of an English-Ndebele medical dictionary. The corpus comprises English STs and equivalent Ndebele TTs with multiple translations collected from Zimbabwean non-governmental and governmental organisations.

Likewise, Shoba (2018) explores how parallel corpora can be analysed with the use of the concordancer *ParaConc* to extract bilingual terminology that can be used to create specialised bilingual dictionaries. Shoba (2018) follows a corpus-based approach because it quickly, efficiently, and accurately allows the extraction of bilingual terms in their immediate contexts.

2.3. South African universities and corpus compilation

To illustrate the state of affairs of corpus compilation as an institutional resource in South Africa, we will show two examples of HEIs in the country that have made progress in corpus building: a) the University of Pretoria and b) the University of KwaZulu-Natal. The Department of African Languages of the University of Pretoria has been involved in corpora compilation since the early 1990s. It compiled the *Pretoria Sesotho sa Leboa Corpus* (PSC; Prinsloo 1991), which started with 156,000 running words and comprises millions of words nowadays. Similarly, there is a joint project between the Departments of African Languages of the Universities of Pretoria and Ghent that has generated large corpora for all official South African languages, with sizes averaging several million tokens per language (De Schryver and Daniëlle 2005).

The University of KwaZulu-Natal, in line with its language policy and plan, has established a centre specialising in designing an *IsiZulu National Corpus*[5] (INC) and an

[5] https://inzc.ukzn.ac.za

IsiZulu term bank as key enablers in the development of Human Language Technologies (HLT). The University of KwaZulu-Natal, through its University Language Development and Planning Office (ULDPO), has undertaken the development of computer programmes and technologies for the study and use of IsiZulu. These include a) INC, which comprises 20 million tokens, b) an IsiZulu term bank for a variety of disciplines, c) an IsiZulu spellchecker software for writing and editing in IsiZulu with interfaces in both IsiZulu and English, and d) an electronic Zulu lexicon. In essence, they focus on developing IsiZulu as a language of teaching and learning at the university (Khumalo *et al.* 2019).

2.4. Corpus compilation at Unisa

The *Unisa Corpus* discussed here mainly represents Language for Special Purposes (LSP) since the University has adopted a multilingual approach to the creation of learning, such as the use of glossaries that can assist non-proficient students of English to access specialised subject fields in their preferred language. LSP refers to language that tends to be formal and contains a highly specialised vocabulary. LSP texts are restricted and precise and typically feature, amongst others, an abundance of specialised terms. LSP phrases and terminology are chiefly aimed at serving the communication needs of specialists. LSP language strongly contrasts with Language for General Purposes (LGP), which is of common usage.

In implementing its language policy in 2016, the University of South Africa, through its strategic project *Transformation: Building capacity for South African languages*, started collating Unisa texts that have been translated from English several South African official languages. This collection of texts includes an ST and its TTs and it is the first step in corpus compilation, known as ‘corpus design’. The texts comprise tutorial material. The tutorial texts are grouped according to domains that represent various disciplines: education, economic and management sciences, accounting sciences, human sciences, law, agriculture and environmental studies, science, and engineering and technology, as well as the sub-disciplines within these. As the research progresses, other domains will be determined by the type of texts which the institution generates.

Selecting and collecting suitable texts for inclusion in the corpus, as well as obtaining corpus analysis tools, are crucial steps in corpus-based compilation and

research. A corpus needs to be carefully compiled and the quality of the investigation is directly related to the quality of the data (Granger 1998). A corpus is not simply a collection of texts, but rather the appropriate design of a corpus depends on what it is meant to represent. What is deposited into the corpus determines the outcome. As Leech (1998) aptly notices, paying proper attention to quality and design criteria always takes twice as long as one would think and it may take ten times as much effort. Creating a corpus requires a number of researchers, for example, those responsible for selecting the criteria in identifying the texts to be included, for collecting the texts, for grouping the texts according to the individual domains, medium and time, size, etc. De Schryver and Daniëlle (2005) identify three steps in the process of corpus compilation: a) corpus design, b) text collection, and c) text encoding. The size of a corpus is described in terms of the size of individual texts. For a parallel corpus, it is the size of the ST and TTs and the total number of words in the corpus.

In the present study, the corpus selected contains texts from the disciplines of human sciences and accounting sciences, respectively. Table 1 below shows the names of the individual STs and languages of translation. The author of the STs is identified as the Faculty, and the translator is the Language Unit that manages the translation workflow.

	STs and author/s **Faculty of Human Sciences** **Faculty of Accounting Sciences**	**TTs** **Translator: Language Unit**
1	Honours in Development Studies Code: DVAALD_2019_TL_301_B)	*Izifundo Zentuthuko* (IsiZulu) *Dithuto tša Tlhabollo* (Sepedi)
2	Introductory Financial Accounting Code: (FAC_2022_TL_101_B)	*Selelekela ho Akhaonting ya tsa Ditjhelete* (Sesotho) Introductory Financial Accounting (Setswana) Introductory Financial Accounting (Sepedi) *Intshayelelo ngoCwangciso lwezeMali* (IsiXhosa) *Isethulo Sokubalisisa Izimali* (IsiZulu) *Singeniso Kutekuphatfwa Kwetimali* (SiSwati) Introductory Financial Accounting (IsiNdebele) Introductory Financial Accounting (Tshivenda) Introductory Financial Accounting (Xitsonga)

Table 1: Translated tutorial letters for DVAALD and FAC

2.5. Corpus tools

The value of corpora as sources of data lies in that the data may be accessed with the use of software tools. Text retrieval programmes, commonly referred to as 'concordancers', are the most widely linguistic software tool used. Concordancers allow counting words and sequences of words and sort them in a variety of ways. They also provide information on how words combine with each other in the text. In addition, they can carry out comparisons of entities in two corpora and bring out statistically significant differences. Software tools such as *WordSmith Tools*, *Multiconcord*, *ParaConc*, *Sketch Engine* (Kilgarriff *et al.* 2014), and many others enable researchers to carry out searches, which they could never hope to do manually.

In the present analysis, *ParaConc* has been selected as a software programme designed primarily as a search tool that works with parallel texts. *ParaConc* combined with a suitable set of texts can be used as a full context bilingual dictionary or as linked bilingual discourse of translation equivalences and present the user with a) multiple instances of the search term, and b) a large context for each instance of the search term, thereby allowing a thorough analysis of usage in terms of the equivalences between two languages (Barlow 2008: 12).

To illustrate the application of software tools, we provide some screenshots, one from a Development Studies Tutorial Letter (DVAALD_2019_TL_301_B) translated into IsiZulu (see Figure 1), and another one from Introduction to Financial Accounting (FAC_2022_TL_101_B) translated into SiSwati (see Figure 2). It is worth noticing that the African languages we used for illustration and analysis in this study are IsiNdebele, IsiXhosa, IsiZulu and SiSwati. Since FAC 1501 has more TTs and *ParaConc* takes a maximum of four files at a time —English (ST) and IsiXhosa, IsiNdebele and IsiZulu— SiSwati was loaded separately.

Figure 1: *ParaConc* file menu: DVAALD_2019_TL_301_B

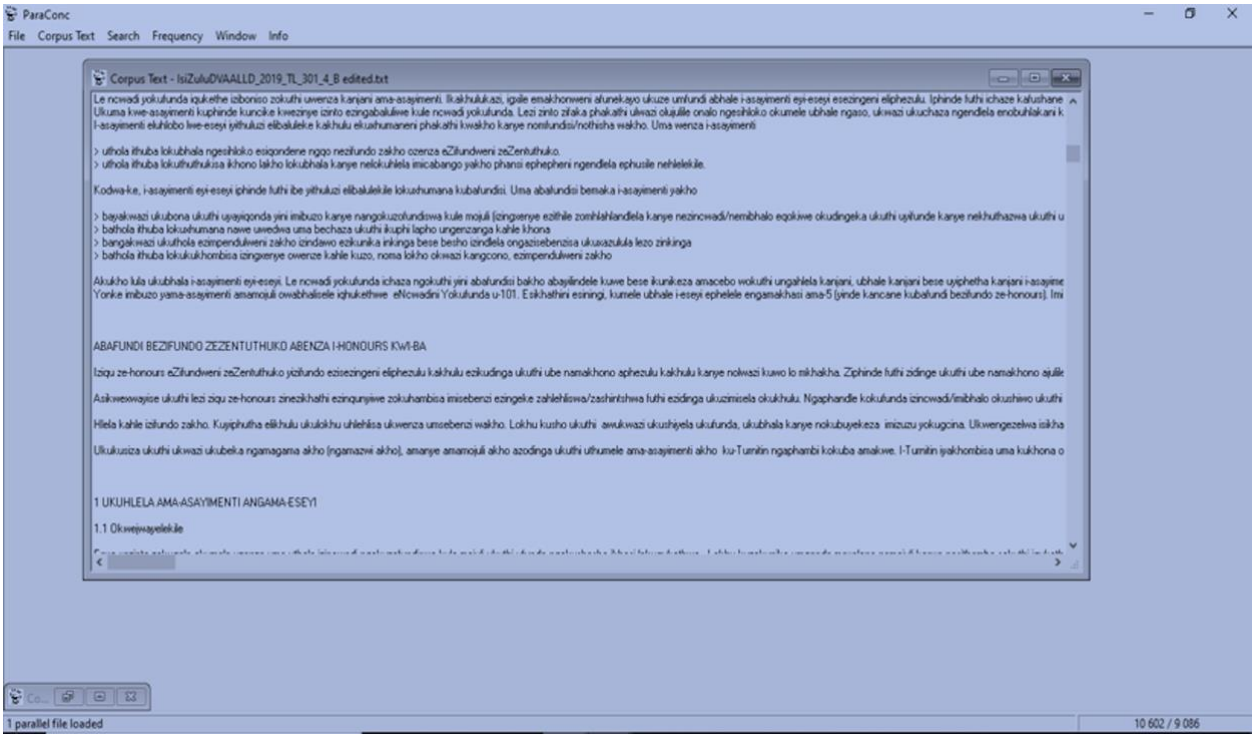


Figure 2: IsiZulu plain text – DVAALD_2019_TL_301_B

The processing of the files may take some time and it is advisable to use the workspace option to minimise the number of times that texts must be uploaded. There is no real limit to the size of the corpus loaded. Even though the corpus files appear to be loaded in the programme ready for searching, in practice, *ParaConc* does not load the whole text, but rather switches chunks of text in and out of memory as one does the searches. This means that the programme should theoretically be able to handle any size of text (see Figure 3

of parallel texts loaded in workspace). Since *ParaConc* does not have a list of the South African indigenous languages, it should be noted that French Canadian is used to label IsiZulu (cf. Figure 3). Finally, Afrikaans, 'Additional 1' and 'Additional 2' are used as labels in Figure 4. In the FAC1501 corpus files, 'Afrikaans' is a label for IsiXhosa, 'Additional 1' is a label for IsiNdebele and 'Additional 2' is for IsiZulu.

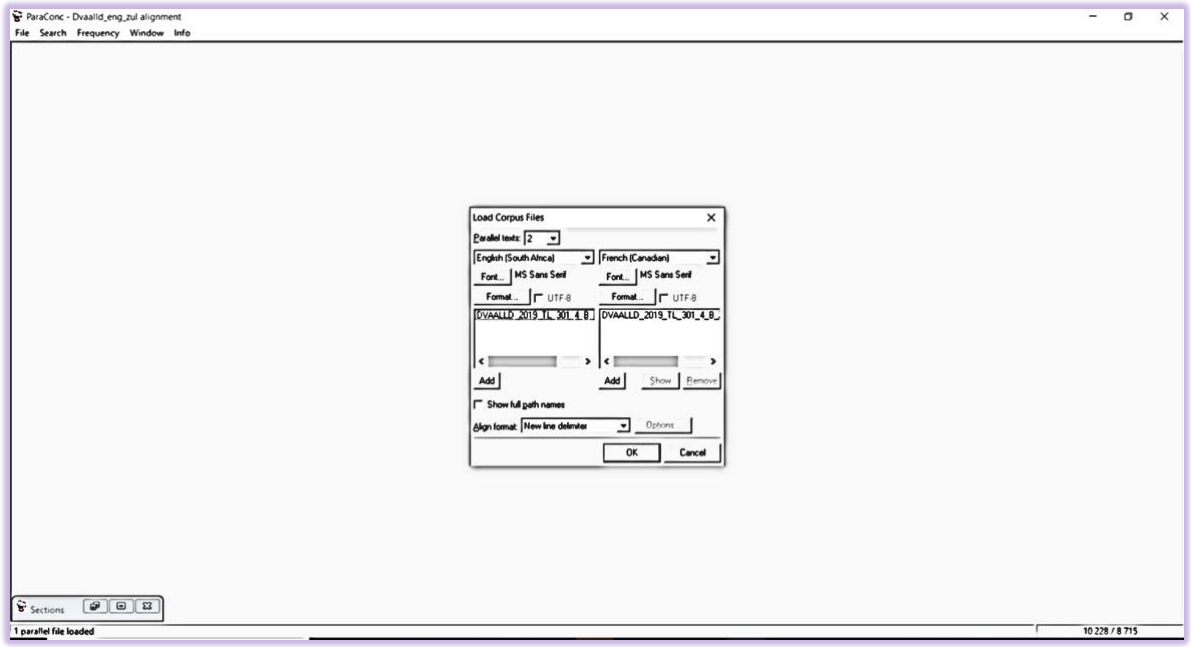


Figure 3: Parallel texts loaded in workspace

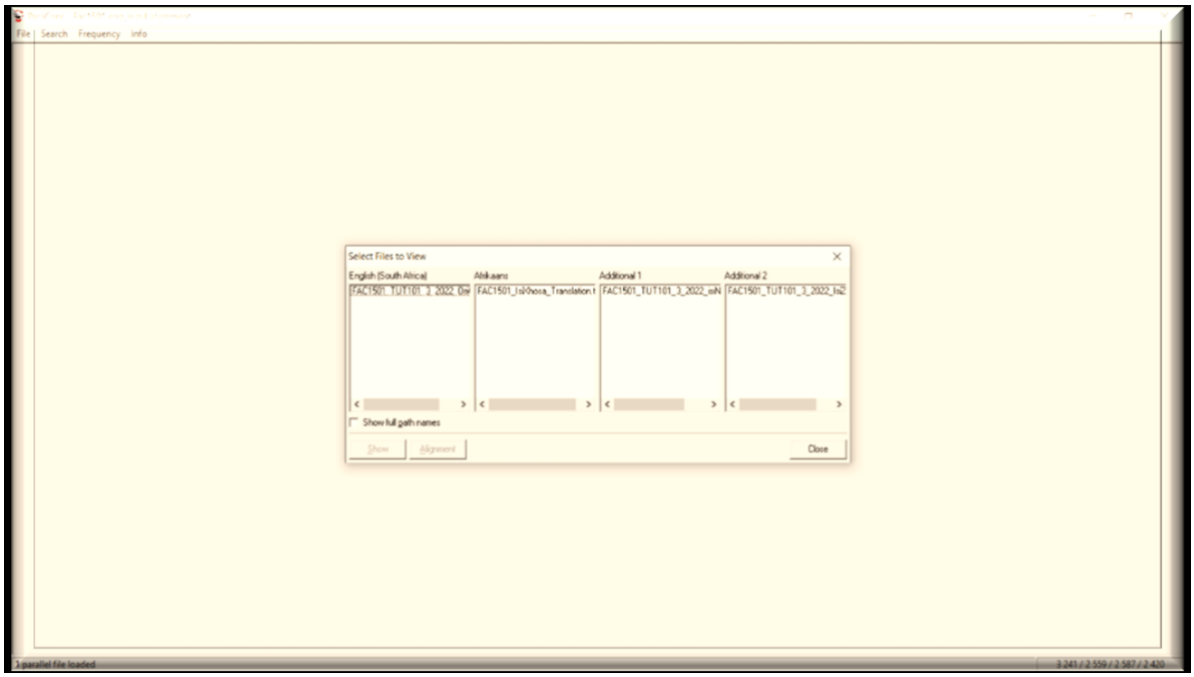


Figure 4: FAC1501 parallel texts loaded in workspace

2.6. Creating a word list

Once the texts have been loaded on *ParaConc*, the corpus frequency commands create a word list for the two parallel texts and the results are displayed in two parallel windows. The word list gives an overall idea of what information can be searched for, and it is very useful for the identification of terminology. The word list may be arranged according to the order of frequency or alphabetically (Figures 5 and 6). The word list in Figure 6 is arranged according to the order of frequency in English, IsiXhosa, IsiNdebele and IsiZulu.

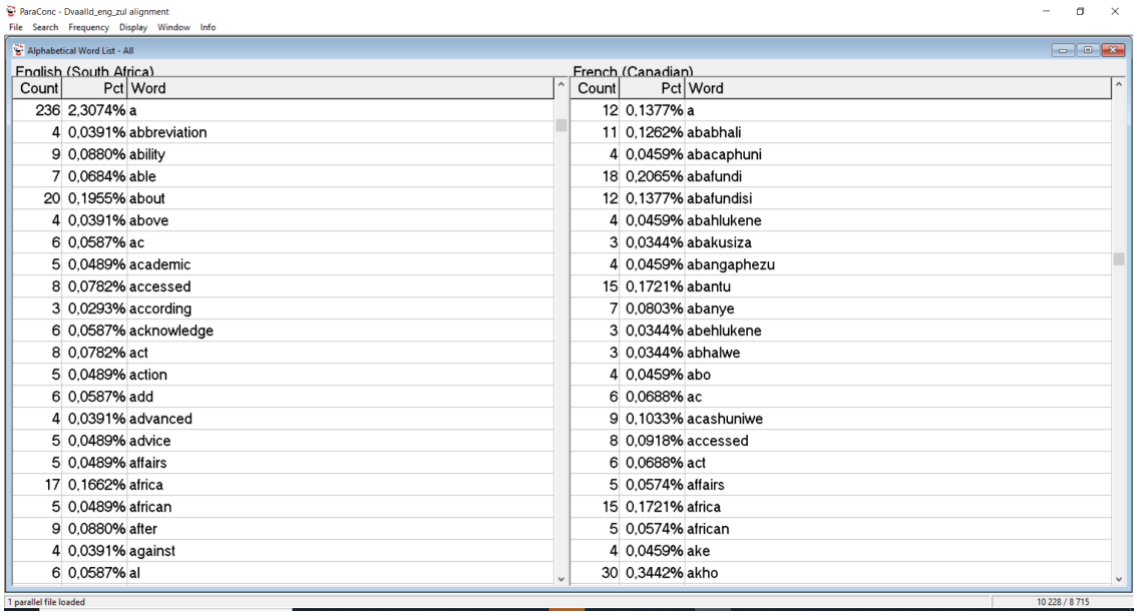


Figure 5: Alphabetical word list – DVAALLD English – IsiZulu

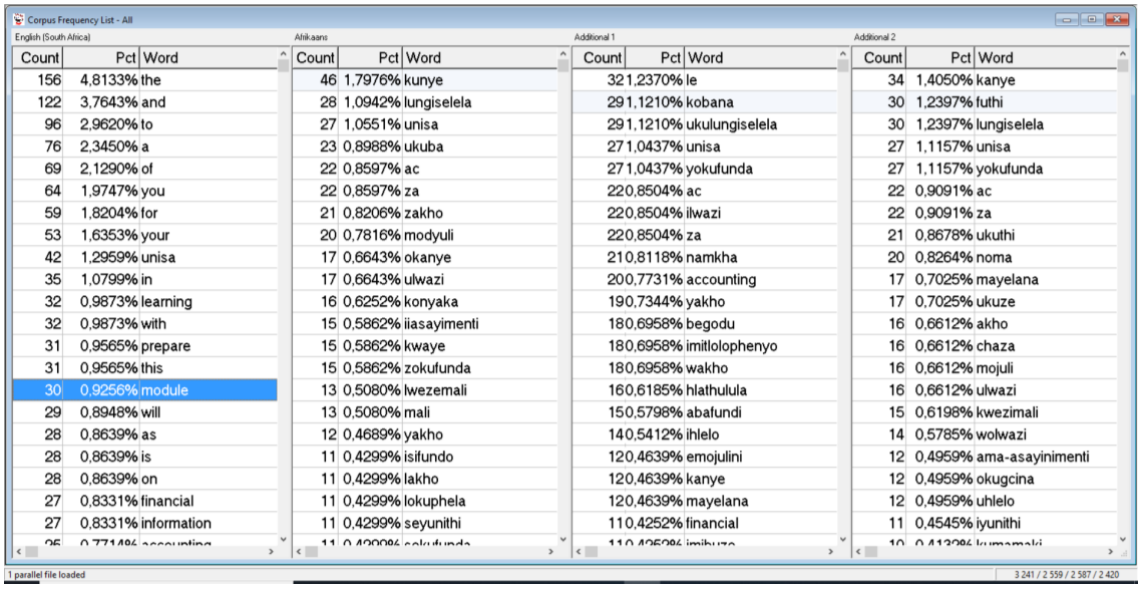
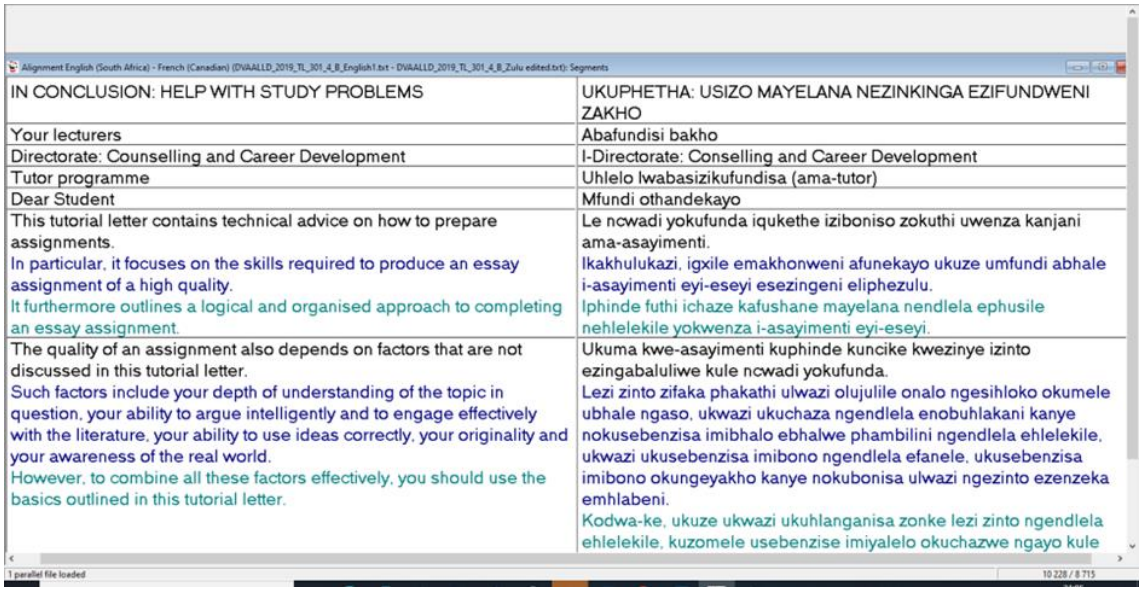


Figure 6: Frequency order – FAC1501 English – IsiXhosa – IsiNdebele – IsiZulu

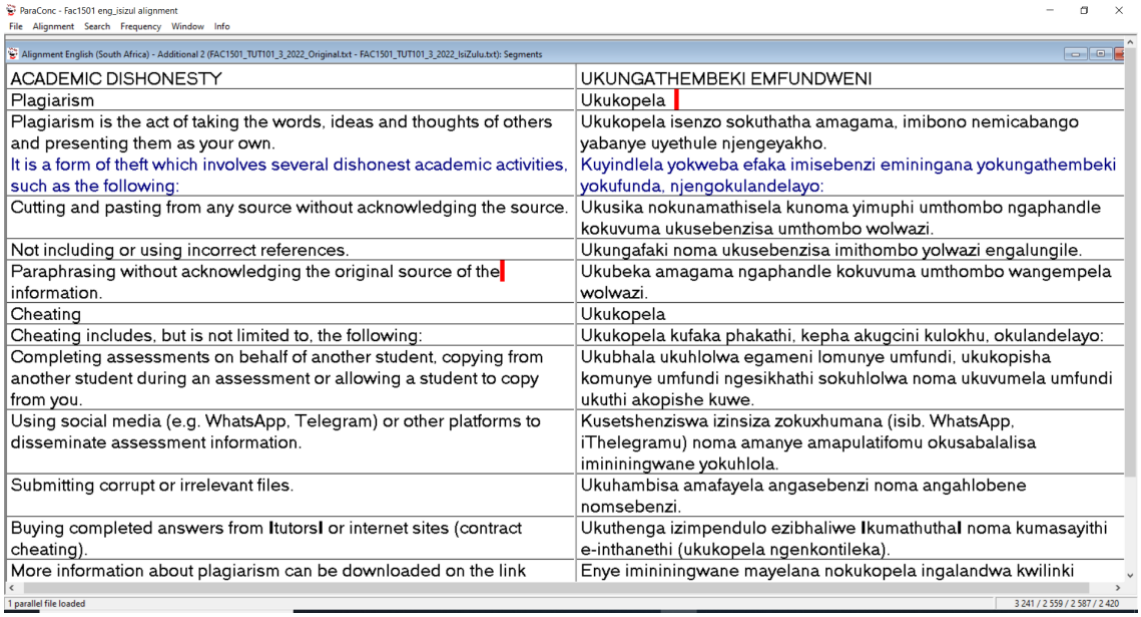
2.7. Alignment

The successful analysis of parallel texts depends on alignment. Alignment creates links between the ST and the TT. In the alignment process, the texts are matched at sentence level so that a sentence in the ST finds a corresponding sentence in the TT (cf. Figures 7 and 8).



IN CONCLUSION: HELP WITH STUDY PROBLEMS	UKUPHETHA: USIZO MAYELANA NEZINKINGA EZIFUNDWENI ZAKHO
Your lecturers	Abafundisi bakho
Directorate: Counselling and Career Development	I-Directorate: Conselling and Career Development
Tutor programme	Uhlelo lwabasizikufundisa (ama-tutor)
Dear Student	Mfundi othandekayo
This tutorial letter contains technical advice on how to prepare assignments. In particular, it focuses on the skills required to produce an essay assignment of a high quality. It furthermore outlines a logical and organised approach to completing an essay assignment.	Le ncwadi yokufunda iqukethe iziboniso zokuthi uwenza kanjani ama-asayimenti. Ikakhulukazi, igxile emakhonweni afunekayo ukuze umfundi abhale i-asayimenti eyi-eseyi esezingeni eliphezulu. Iphinde futhi ichaze kafushane mayelana nendlela ephusile nehlelekile yokwenza i-asayimenti eyi-eseyi.
The quality of an assignment also depends on factors that are not discussed in this tutorial letter. Such factors include your depth of understanding of the topic in question, your ability to argue intelligently and to engage effectively with the literature, your ability to use ideas correctly, your originality and your awareness of the real world. However, to combine all these factors effectively, you should use the basics outlined in this tutorial letter.	Ukuma kwe-asayimenti kuphinde kuncike kwezinye izinto ezingabaluliwe kule ncwadi yokufunda. Lezi zinto zifaka phakathi ulwazi olujulile onalo ngesihloko okumele ubhale ngaso, ukwazi ukuchaza ngendlela enobuhlakani kanye nokusebenzisa imibhalo ebhalwe phambilini ngendlela ehlelekile, ukwazi ukusebenzisa imibono ngendlela efanele, ukusebenzisa imibono okungeyakho kanye nokubonisa ulwazi ngezinto ezenzeka emhlabeni. Kodwa-ke, ukuze ukwazi ukuhlanganisa zonke lezi zinto ngendlela ehlelekile, kuzomele usebenzise imiyalelo okuchazwe ngayo kule

Figure 7: DVAALLD English-IsiZulu-aligned parallel texts



ACADEMIC DISHONESTY	UKUNGATHEMBEKI EMFUNDWENI
Plagiarism	Ukukopela
Plagiarism is the act of taking the words, ideas and thoughts of others and presenting them as your own. It is a form of theft which involves several dishonest academic activities, such as the following:	Ukukopela isenzo sokuthatha amagama, imibono nemicabango yabanye uyethule njengeyakho. Kuyindlela yokweba efaka imisebenzi eminingana yokungathembeki yokufunda, njengokulandelayo:
Cutting and pasting from any source without acknowledging the source.	Ukusika nokunamathisela kunoma yimuphi umthombo ngaphandle kokuvuma ukusebenzisa umthombo wolwazi.
Not including or using incorrect references.	Ukungafaki noma ukusebenzisa imithombo yolwazi engalungile.
Paraphrasing without acknowledging the original source of the information.	Ukubeka amagama ngaphandle kokuvuma umthombo wangempela wolwazi.
Cheating	Ukukopela
Cheating includes, but is not limited to, the following:	Ukukopela kufaka phakathi, kepha akugcini kulokhu, okulandelayo:
Completing assessments on behalf of another student, copying from another student during an assessment or allowing a student to copy from you.	Ukubhala ukuhlolwa egameni lomunye umfundi, ukukopisha komunye umfundi ngesikhathi sokuhlolwa noma ukuvumela umfundi ukuthi akopishe kuwe.
Using social media (e.g. WhatsApp, Telegram) or other platforms to disseminate assessment information.	Kusetshenziswa izinsiza zokuxhumana (isib. WhatsApp, iThelegramu) noma amanye amapulatifomu okusabalalisa imininingwane yokuhlola.
Submitting corrupt or irrelevant files.	Ukuhambisa amafayela angasebenzi noma angahlobene nomsebenzi.
Buying completed answers from **I**tutors**I** or internet sites (contract cheating).	Ukuthenga izimpendulo ezibhaliwe **I**kumathuth**a**l noma kumasayithi e-inthanethi (ukukopela ngenkontileka).
More information about plagiarism can be downloaded on the link	Enye imininingwane mayelana nokukopela ingalandwa kwilinki

Figure 8: FAC1501 English-IsiZulu-aligned parallel texts

The alignment process is very important for the successful operation of the software. If the sentences are not aligned properly (as often happens), a menu of options makes it possible to split and merge sentences or segments (see Figure 9: note drop-down menu).

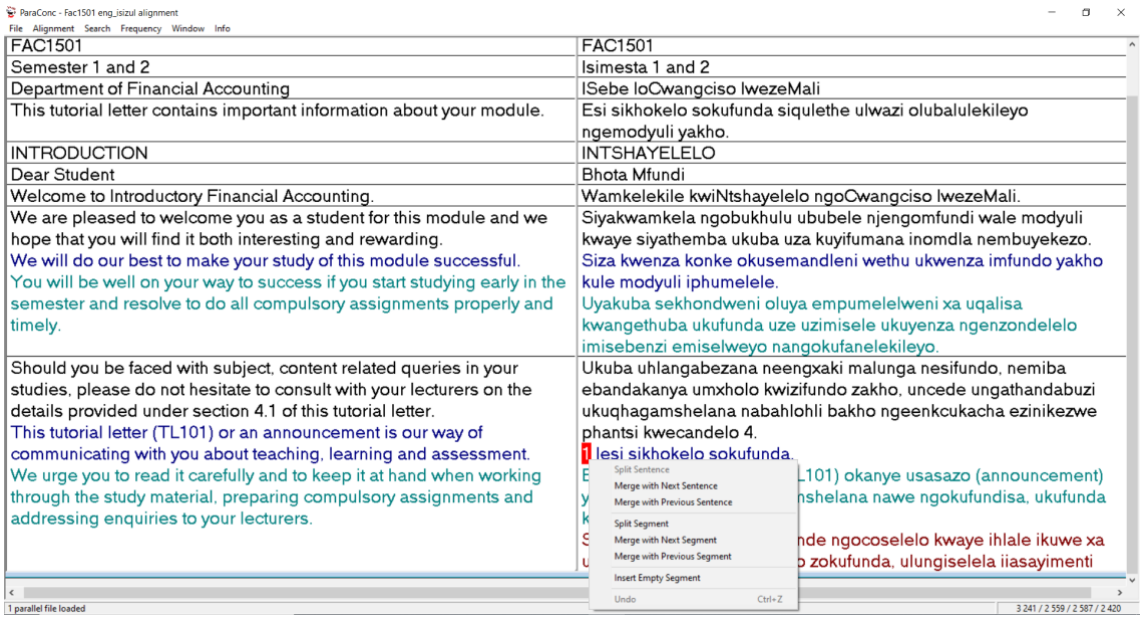


Figure 9: Alignment – split and merge sentences

3. DISCUSSION

One of the advantages of using a corpus is that researchers may access words in their real context (Moropa 2004). Context provides knowledge patterns that offer more information about the use of a term. This can only be possible by compiling a corpus of a specific domain in all languages. Such an exercise helps in defining equivalences. Moreover, researchers become empowered to coin words in their languages or use other word formation strategies. Table 2 shows the texts that were queried from human and accounting sciences.

STs	**TTs and language**	**Size/words**
Tutorial letter 301 for Honours in Development Studies (DVAALLD)		10,228
	Izifundo Zentuthuko (IsiZulu)	8,715
Tutorial letter 101 Introductory Financial Accounting (FAC1501)		3,241
	Isikhokelo sokufunda 101 – *Intshayelelo ngoCwangciso lwezeMali* (IsiXhosa)	2,559
	Incwadi yokufundisa 101 – Introductory Financial Accounting (IsiNdebele)	2,587
	Incwadi yokufundisa 101 – *Isethulo Sokubalisisa Izimali* (IsiZulu)	2,420
	Incwadzi Yekufundzisa 101 – *Singeniso Kutekuphatfwa Kwetimali* (SiSwati)	2,767

Table 2: The texts analysed in the study

When the texts are properly aligned, the researcher can search for a specific term and its translation equivalent using *ParaConc*. In this study, a keyword search for tutorial letter in the DVAALLD file was performed. The window showed two texts, the English text above and the IsiZulu translation below. The search for the English word 'tutorial' (Figure 10) had nine hits: the keyword is highlighted in blue and its collocates in red.

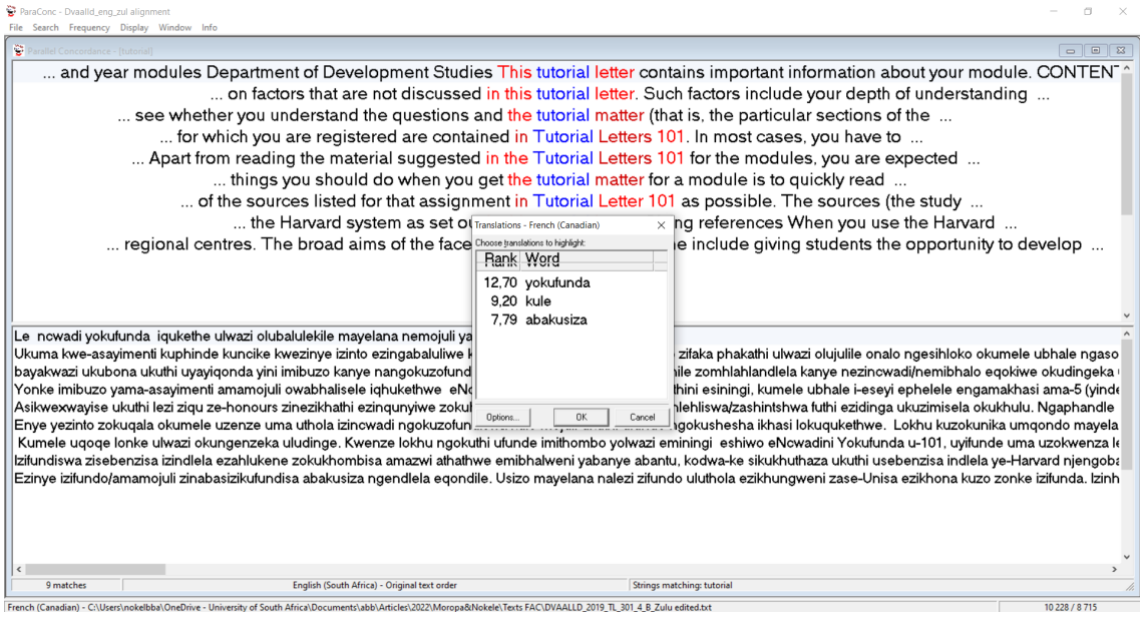


Figure 10: Identifying translation equivalents for tutorial letter in IsiZulu (DVAALLD)

To find the translation equivalent of 'tutorial', we right-clicked on the IsiZulu window, and a drop-down menu with options appeared. We then scrutinised the results and found that the equivalent for IsiZulu in the DVAALLD file is *incwadi yokufunda*. Similar results were found in the FAC1501 file: IsiNdebele *incwadi yokufundisa*, IsiZulu *incwadi yokufunda*, and SiSwati *tincwadi lekufundzisa*. IsiXhosa showed a different result which was *isikhokelo*. Although the translators had different solutions for 'tutorial letter': *incwadi yokufundisa/lekufundzisa* ('a book for teaching'), *incwadi yokufunda* ('a book for learning') and *isikhokelo* ('a guide'), they all essentially imply the same knowledge information that involves learning.

A search for 'trial balance' (FAC1501) yielded four results and the translation equivalents in SiSwati were paraphrases such as *simo semabhuku etimali* and *simo setimali*. IsiNdebele used indigenisation *ithrayalibhalans* ('trial balance'), while IsiXhosa settled for *ibhalansi yolingo*, which is the equivalent for 'general ledger' (cf. Figure 11), another common term in accounting. IsiNdebele and IsiXhosa translators opted for *ileja*, an indigenisation strategy throughout the texts, while IsiZulu and SiSwati translators used

paraphrasing: in the case of IsiZulu *ibhuku elijwayekile lemali* (‘a book that is common for finance / cash’), *ibhukwana* (‘a small book’), *ibhukwana elijwayelekile* (‘a small book that is common’), *ibhukwana elivamile* (‘a book that is known’) and, in the case of SiSwati, *umculu wekubikwa kwemali* (‘a book that records money/finances’), *umculu wemarikhodi etetimali* (‘book for financial records’), and *kumarekhodi embiko wetetimali* (‘in the record of financial report’).

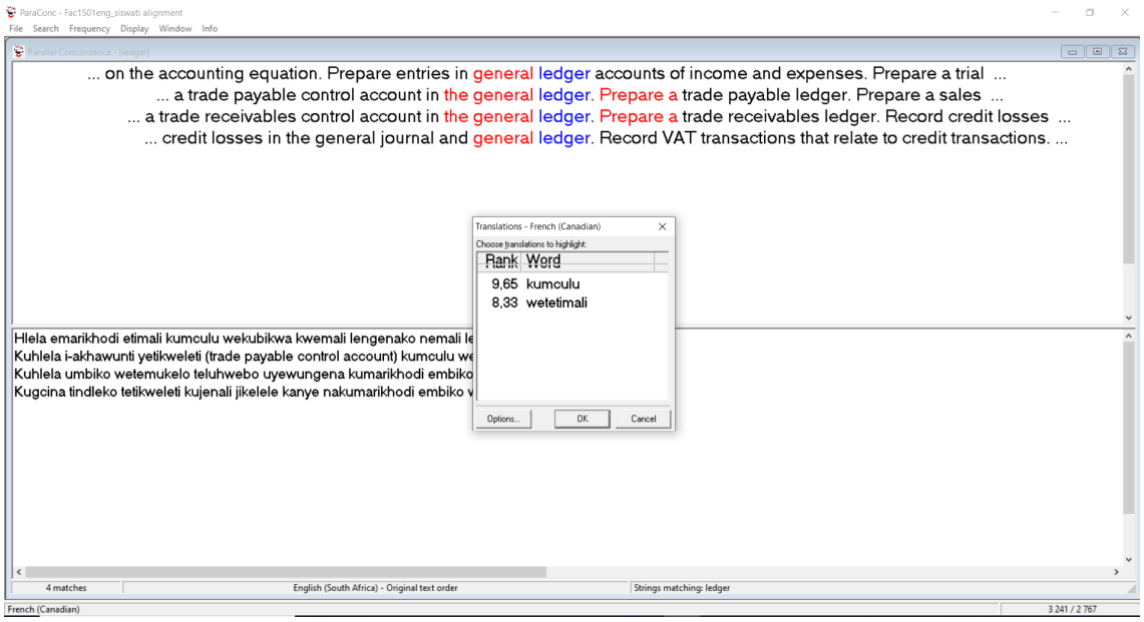


Figure 11: Identifying translation equivalent for general ledger in SiSwati (FAC1501)

Another finding was the translation of the word ‘assignment’ in IsiNdebele as *umtlolophenyo* and *sivivinyo* in SiSwati, which are coined or derived words, while the IsiXhosa and IsiZulu used indigenisation *iasyimenti/iasayinimenti*. Earlier on, we pointed out that a corpus can be used as resource from which terms can be drawn. For example, if IsiXhosa does not have a term for ‘assignment’, as illustrated earlier, and does not make use of a borrowed word, the researcher can consult the other sister languages in the Nguni language family and borrow it from them. In this instance, IsiXhosa translators can learn from IsiNdebele, which opted for the indigenous term *umtlolophenyo*, or they can coin their own IsiXhosa word related or similar to the IsiNdebele word. *Umtlolophenyo* is a compound word which is derived from *umtlolo* (‘writing’) and *phenyo* (‘research’). If the IsiXhosa translators decide to learn from IsiNdebele, they may use compounding and form *ubhalophando* as an equivalent for ‘assignment’. This way a new term may be formed as a synonym. For ‘journal’ and ‘receipts’, borrowing was used and indigenisation of the loan word was adopted, that is, the term was spelt according to the orthographical

rules of the borrowing language: IsiXhosa *ijenali*, *(yee)risithi*, SiSwati *ijenali*, and IsiNdebele *ijenali* and *(yama)risidi*.

Another useful tool in *ParaConc* is the hot word tool, which helps in identifying both possible translations and the knowledge patterns such as collocations or synonyms. Hot words are selected by looking at the frequency of words. The top-ranking words may include translations, translations of collocations, and collocations of the search word. This is illustrated in Figure 12, which shows the occurrences of 'accounting' in the English-SiSwati alignment.

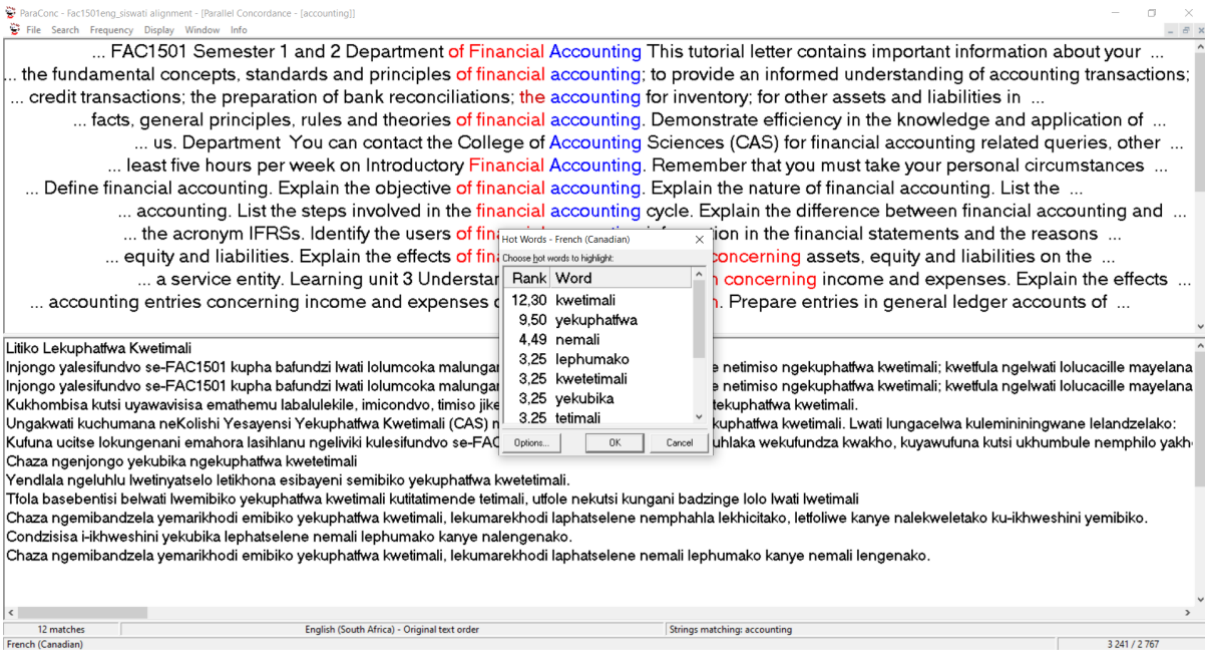


Figure 12: Hot words for 'accounting' (FAC1501-English - SiSwati)

The search revealed several hot words suggesting SiSwati equivalents, of which *kwetimali* (12.30) and *yekuphatfwa* (9.50) ranked the highest. Results for IsiNdebele revealed 'accounting' as a loan word, which suggests that most of the time accounting was reproduced and not translated. 'Financial' with a co-occurring word, as in 'financial accounting', was also not translated. The high frequency words for IsiXhosa were *lwezemali*, *lwemali*, and *mali*, which suggests that accounting in IsiXhosa has to do with money. The information revealed for IsiZulu was *kwezimali* and e*zibalweni*. When comparing the four languages it seems clear that IsiXhosa, IsiZulu and SiSwati conceptualise 'accounting' in the same way, because the highest-ranking words are *lwezemali*, *kwezimali* and *kwetimali*, respectively. In this sense, IsiNdebele can use the same translation strategy instead of using the loan word(s).

Other terms that were queried were generic terms attested in all level-one tutorial letters. These terms were ‘module’, ‘electronic reserves’, ‘internet’, ‘online’, and ‘plagiarism’. Similarity in the translation of certain words was observed during our analysis. ‘Module’ was translated as *imojuli* in IsiZulu and IsiNdebele, and *imodyuli* in IsiXhosa. The only difference between these words is in spelling and pronunciation. The translation strategy is the same in all three languages, that is, indigenisation. The equivalent for SiSwati is *tifundvo* (‘lesson’), which is different from other sister languages. SiSwati opted for an indigenous term that carries a similar meaning. *Imojuli/imodyuli* and *tifundvo* can then be taken as referring to the same term.

‘Electronic reserves’, ‘internet’, and ‘online’ are among the terms that generally pose a challenge to translators of African languages. The analysis of the five texts revealed that translators sometimes use ‘internet’ and ‘online’ interchangeably. IsiXhosa and IsiZulu, for instance, opt for *ku-inthanethi* or *ngeintanethi*, while IsiNdebele and SiSwati use *nge-online/ya-online/ta-online*. IsiXhosa sometimes uses *ngomoya*, which refers to waves rather than wind. This shows that translators understand how the ‘internet’ or ‘online’ terms work. The hot word tool revealed that the use of ‘online’ in the SiSwati texts has a ranking of 7.21, while *nge-online* in IsiNdebele has a ranking of 6.26, and *ku-inthanethi* has a ranking of 6.71 in the context where ‘online’ is used. These rankings suggest that the words ‘internet’ and ‘online’ are treated as synonyms in these languages and can therefore be incorporated in their lexicon.

‘Plagiarism’ is another term that is difficult to translate into African languages. Most of the time it is paraphrased in long sentences, as in *ukuthatha umsebenzi womnye umntu uwenze owakho* (‘to take one’s work and present it as your own’), in IsiXhosa, or *ubunikazi bomsebenzi ekungasiwo wakho i-plagiarism* (‘ownership of work that is not your own’) in IsiNdebele. IsiZulu and SiSwati simplify this explanation by opting for *ukukopela* (‘copying’) or *kukopa* (‘cheating’). It is interesting to note that there is a kind of uniformity among the languages regarding the translation of academic dishonesty: *ukungathembeki emfundweni* (‘to be dishonest in education’) in IsiZulu and IsiNdebele, *kungetsembeki emsebentini* (‘to be dishonest in your work’) in SiSwati, *ungathembeki kwimfundo* (‘to be dishonest in education’) in IsiXhosa. There is also curriculum transformation: *ukuguqulwa kwekharikhulamu* (‘changing in the currriculum’) in IsiZulu, *ukutjhugululwa kwekharikhyulamu* (‘changing in the currriculum’) in IsiNdebele, *tingucuko kukharikhulamu* (‘changing in the currriculum’) in *SiSwati*, and *Utshintsho*

kwikharityhulamu (‘changing in the currriculum’) in IsiXhosa. Such similarities confirm the acceptance of the terms or concepts by the community.

The analysis has also revealed that a corpus can be beneficial in terminology development when looking at the knowledge information and knowledge patterns displayed in the texts. Researchers and translators can retrieve information from such knowledge patterns when developing terms in their own languages. Another advantage of using a corpus is that researchers can see the translation strategies adopted by other language practitioners when dealing with equivalence or lack of terminology in their own languages. This then puts them in a better position to make informed decisions.

The process explained above illustrates how a parallel corpus can be used to extract terms and their translations, thereby enriching the indigenous languages. Corpora as authentic resources provide terms in use. The frequency of use also indicates whether the term has been accepted by the speech community. The similarity that is observable in how IsiNdebele, IsiXhosa, IsiZulu, and SiSwati translators deal with specialised terminology indicates that these languages are being developed and also shows that related languages can borrow from each other to solve the problem of lack of terms.

The step towards corpus building and use of software to query data will improve the process of the identification of terms, both in monolingual and parallel platforms. In dictionaries, information about frequency or generalisation of use is not provided in a consistent manner, whereas such facts can be obtained somewhat more readily from parallel texts. As Bowker (2000: 21) states, parallel corpora contain a range of terms that is wider than dictionaries, present terms in context, and are more current than dictionaries. They allow translators to acquire both specialised conceptual and linguistic knowledge about terms and, as they are available in electronic form, they can increase the scale and speed of a translator’s research. Teubert (2005: 98) concludes that:

> many words in our ordinary language have unspecific meanings which cannot be described without referring to the context in which they occur, but this is what dictionaries, due to their constraints in space, cannot do effectively.

The meaning of words is created in texts and this renders specialised status to terms, which is why it is fundamental to identify the context where texts are produced and to describe the texts that constitute the corpus (Moreira 2014).

4. CONCLUSION

Parallel texts present terms in authentic contexts allowing terminologists and translators to acquire specialised conceptual and linguistic knowledge. The extraction of terminology from parallel corpora (considering the different language pairs) is not only feasible but also extremely useful in developing scientific and technical vocabulary for large-scale use in Unisa. In addition, it is worth noting that the researcher, that is, the linguist, verifies the list of terms. History shows that the indigenous African languages are capable of drawing on their own resources and can create the necessary terms from their own vocabulary.

To reinforce language transformation within Unisa, corpus compilation and the use of software tools should be part of the agenda of language development for teaching, learning, and research. This is in line with the language policy framework that encourages institutions of Higher Learning in South Africa to develop language plans and strategies that will enhance the development and promotion of indigenous African languages as centres of research and scholarship. By constructing corpora as resources for linguistic research, Unisa will also be acknowledging the authors of the texts, as well as the language practitioners who translate, edit, and proofread the texts that are public. Unisa is the only university in South Africa that, in its transformation agenda, has committed to build capacity for all official South African languages. Furthermore, it is the only university in South Africa that teaches all official South African languages. Electronic written corpora should serve as a tool for the development of South African languages, given that a corpus provides the researcher with a wealth of linguistic data instantly. Corpus-based linguistic research within the University of South Africa will transcend terminology development because the corpus-based approach can be applied to empirical investigations in almost any discipline with its attendant linguistic challenges.

REFERENCES

Alexander, Neville. 2003. *African Renaissance and the Use of African Languages in Tertiary Education*. Cape Town: The Estate of Neville Edward Alexander.

Baker, Mona Baker. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7/2: 223–243.

Barlow, Michael. 2008. *ParaConc and Parallel Corpora in Contrastive and Translation Studies*. Houston: Athelstan.

Bowker, Lynne. 2000. Towards a methodology for exploiting specialized target language corpora as translation resources. *International Journal of Corpus Linguistics* 5/1: 17–52.

Department of Education. 2002. *Language Policy for Higher Education*. Pretoria: Government Printers.

Department of Education. 2020. *Language Policy for Higher Education*. Pretoria: Government Printers.

De Schryver, Gilles-Maurice and Jacobus Daniëlle. 2005. Managing eleven parallel corpora and the extraction of data in all official South African languages. In Walter Daelemans ed. *Multilingualism and Electronic Language Management*. Pretoria: Van Schaik, 100–122.

Gauton, Rachélle and Gilles-Maurice De Schryver. 2004. Translating technical texts into Zulu with the aid of multilingual and/or parallel corpora. *Studies in the Languages of Africa* 35/1: 148–161.

Granger, Sylviane. 1998. *Learner English on Computer*. London: Longman.

Khumalo, Langa, Valentine Azom and Peter Olukanmi. 2019. The design and implementation of a corpus management system for IsiZulu National Corpus. In Martin Doerr, Oyvind Eide, Oddrun Gronvik and Bjorghild Kjelsvik eds. *Humanists and the Digital Toolbox*. Oslo: Novus Forlag, 179–196.

Kilgarriff, Adam, Vít Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. The Sketch Engine: Ten years on. *Lexicography* 1/1: 7–36.

Leech, Geoffrey. 1998. Learner corpora: What they are and what can be done with them. In Sylviane Granger ed. *Learner English on Computer*. London: Longman: xiv–xx.

Madiba, Mbulungeni. 2001. Towards a model for terminology modernisation in the African languages of South Africa. *Language Matters* 32/1: 53–77.

Madiba, Mbulungeni. 2004. Parallel corpora as tools for developing indigenous languages of South Africa. *Language Matters* 35/1: 133–147.

Mlambo, Respect, Nomsa Skosana and Muzi Matfunjwa. 2021. The extraction of terminology list using ParaConc for creating a quadrilingual dictionary. *Southern African Linguistics and Applied Language Studies* 39/1: 82–91.

Moreira, Adonay. 2014. A methodology for building a translator and translation oriented terminological resource. *InTralinea Online Translation Journal*. https://www.intralinea.org/specials/article/2032

Moropa, Koliswa. 2004. A parallel corpus as a terminology resource for Xhosa: A study of strategies used to translate financial statements. *Language Matters* 35/1: 162–178.

Moropa, Koliswa. 2005. *An Investigation of Translation Universals in a Parallel Corpus of English-Xhosa Texts*. Pretoria: University of South Africa dissertation.

Moropa, Koliswa. 2007. Analysing the English-Xhosa parallel corpus of technical texts with ParaConc: A case study of term formation processes. *South African Linguistics and Applied Language Studies* 25/2:183–205.

Moropa, Koliswa and Feziwe Martha Shoba. 2017. Language and terminology development in IsiXhosa: A history. In Russell H. Kaschula, Pamela Maseko and H. Ekkehard Wolff eds. *Multilingualism and Intercultural Communication: A South African Perspective*. Johannesburg: Wits University Press, 76–91.

Ndhlovu, Ketiwe. 2012. *An Investigation of Strategies Used by Ndebele Translators in Zimbabwe in Translating HIV/AIDS Texts: A Corpus-based Approach*. Alice: University of Fort Hare dissertation.

Ndhlovu, Ketiwe. 2016. Using ParaConc to extract bilingual terminology from parallel corpora: A case of English and Ndebele. *Journal of Literary Criticism, Comparative Linguistics and Literary Studies* 37/2: 1–12.

Prinsloo, Daniël Jacobus. 1991. Towards computer assisted word frequency studies in Northern Sotho. *South African Journal of African Languages* 11/2: 54–60.

Republic of South Africa. 1996. *The Constitution of the Republic of South Africa* (Act 108 of 1996). Pretoria: The Government Printer. https://www.gov.za/documents/constitution-republic-south-africa-1996

Sinclair, John. 1995. Corpus typology: A framework for classification. In Gunnel Mechers and Beatrice Warren eds. *Studies in Anglistics*. *Acta Universitatistockholmienses*. Stockholm: Almqvist & Wicksell, 17–33.

Shoba, Feziwe Martha. 2018. *Exploring the Use of Parallel Corpora in the Compilation of Specialized Bilingual Dictionaries of Technical Terms: A Case Study of English and IsiXhosa*. Pretoria: University of South Africa dissertation.

Teubert, Wolfgang. 2005. Language as an economic factor: The importance of terminology. In Geoffrey Barnbrook, Pernilla Danielsson and Michaela Mahlberg eds. *Meaningful Texts: The Extraction of Semantic Information from Monolingual and Multilingual Corpora*. London: Continuum, 96–106.

University of South Africa. 2016. *Unisa Language Policy*. https:www.unisa.ac.za/policies

Corresponding author
Koliswa Moropa
University of South Africa
E-mail: koliswa04@gmail.com
Department of Tuition Support and Facilitation of Learning
Muckleneuk Campus
1 Preller Street
Pretoria 0002
South Africa

received: March 2023
accepted: May 2023

APPENDIX

Appendix 1: Terms extracted from the corpus

ST term	TT term	Language	Translation strategy
Tutorial letter	*Incwadi yokufunda*	IsiZulu	Paraphrase
	Incwadi yokufundisa	IsiNdebele	Paraphrase
	Tincwadi lekufundzisa	SiSwati	Paraphrase
	Isikhokelo	IsiXhosa	Substitution
Trial balance	*Ibhalansi yomzamo*	IsiZulu	Paraphrase
	Simo semabhuku etimali simo setimali	SiSwati	Paraphrase
	Ithrayalibhalansi	IsiNdebele	Indigenisation
	Ibhalansi yolingo	IsiXhosa	Paraphrase
General ledger	*Ibhuku elijwayekile lemali/ Ibhukwana elijwayelekile/ ibhukwana elivamile*	IsiZulu	Paraphrase
	Ileja	IsiNdebele	Indigenisation
	Umculu wekubikwa kwemali/ umculu wemarikhodi etetimali / kumarekhodi embiko wetetimali	SiSwati	Paraphrase
	Ileja	IsiXhosa	Indigenisation
Assignment	*Iasayinimenti*	IsiZulu	Indigenisation
	Umtlolophenyo	IsiNdebele	Substitution
	Sivivinyo	SiSwati	Substitution
	Iasayimenti	IsiXhosa	Indigenisation
Journal (of receipts)	*Ijenali (yamarisidi)*	IsiZulu	Indigenisation
	Ijenali yamarisidi	IsiNdebele	Indigenisation
	Ijenali	SiSwati	Indigenisation
	ijenali (yeerisithi)	IsiXhosa	Indigenisation
Module	*Imojuli*	IsiZulu	Indigenisation
	Imojuli	IsiNdebele	Indigenisation
	Tifundvo	SiSwati	Substitution
	Imodyuli	IsiXhosa	Indigenisation
Accounting	*Ukubalwa kwezimali*	IsiZulu	Paraphrase
	Ukubalwa kwemali	IsiNdebele	Paraphrase
	Ubalo lwezimali	SiSwati	Paraphrase
	ucwangcisomali	IsiXhosa	Substitution
Financial accounting	*Ukubalwa kwemali yokubalwa kwemali wokubalwa kwemali*	IsiZulu	Paraphrase
	Financial accounting	IsiNdebele	Loan word
	Ukuphatfwa kwetimali	SiSwati	Paraphrase
	Ucwangcisomali	IsiXhosa	Substitution
Inventory control	*Ulawulo lokusungula*	IsiZulu	Substitution
	Ulawulo lwenani yepahla	IsiNdebele	Paraphrase
	Luhlelo loluchaphaluhlu Lwemphahla	SiSwati	Paraphrase
	Ulawulo lwempahla	IsiXhosa	Substitution

ST text term	TT text term	Language	Translation strategy
Online	*Ku-inthanethi*	IsiZulu	Indigenisation
	Nge-online	IsiNdebele	Loan word
	Online	SiSwati	Loan word
	Ya-online		
	Ta-online		
	Ngomoya	IsiXhosa	Substitution,
	Ngeintanethi		indigenisation
Purpose	*Inhloso*	IsiZulu	Substitution
	Umnqopho	IsiNdebele	Substitution
	Injongo	SiSwati	Substitution
	Injongo	IsiXhosa	Substitution
Outcomes	*Imiphumela*	IsiZulu	Substitution
	Imiphumela	IsiNdebele	Substitution
	Imiphumela	SiSwati	Substitution
	Iziphumo	IsiXhosa	Substitution
Academic dishonesty	*Ukungathembeki emfundweni*	IsiZulu	Paraphrase
	Ukungathembeki emfundweni	IsiNdebele	Paraphrase
	Kungetsembeki emsebentini	SiSwati	Paraphrase
	Ukungathembeki kwimfundo	IsiXhosa	Paraphrase
Electronic reserves	*Imithombo eku-inthanethi*	IsiZulu	Paraphrase
	Imithombo eyi-Electronic reserves (e-reserves)	IsiNdebele	Paraphrase plus loan word
	lwati lolugcinwe emishinini	SiSwati	Paraphrase
	izixhobo ezigcinwe ngoomatshini	IsiXhosa	Paraphrase
Assessment criteria	*Indlela yokuhlola*		Substitution
	Amaqhinga wokuhlola		Paraphrase
	Indlela yekuhlolwa kwebafundzi		Paraphrase
	Iindlela zovavanyo		Substitution
Curriculum transformation	*Ukuguqqulwa kwekharikhulamu*	IsiZulu	Substitution
	Ukutjhugululwa kwekharikhyulamu	IsiNdebele	Substitution
	tingucuko kukharikhulamu	SiSwati	Substitution
	Utshintsho kwikharityhulamu	IsiXhosa	Substitution
Resources	*izinsizakusebenza*	IsiZulu	Substitution
	Imithombo	IsiNdebele	
	Tinsita tekufundza resources	SiSwati	Paraphrase
	Izibonelelo zokusebenza	IsiXhosa	Paraphrase
Log on	*Ngena*	IsiZulu	Substitution
	Loga	IsiNdebele	Indigenisation
	Condza	SiSwati	Substitution
	Ngena	IsiXhosa	Substitution

ST text term	TT term	Language	Translation strategy
Define	*Chaza*	IsiZulu	Substitution
	Hlathulula	IsiNdebele	Substitution
	Chaza	SiSwati	Substitution
	Chaza	IsiXhosa	Substitution
Identify	*Thola* *Khomba*	IsiZulu	Substitution
	Tjengisa	IsiNdebele	Substitution
	Tfola	SiSwati	Substitution
	Chonga *Bonisa*	IsiXhosa	Substitution
Explain	*Chaza*	IsiZulu	Substitution
	Hlathulula	IsiNdebele	Substitution
	Chaza	SiSwati	Substitution
	Cacisa	IsiXhosa	Substitution
Plagiarism	*Ukukopela*	IsiZulu	Indigenisation
	Ubunikazi bomsebenzi ekungasiwo wakho i-plagiarism	IsiNdebele	Paraphrase plus loan word
	Kukopa	SiSwati	Indigenisation
	Ukuthatha umsebenzi womnye umntu uwenze owakho	IsiXhosa	paraphrase
Development Studies	*Izifundo zezentuthuko*	IsiZulu	Substitution
Critical reading	*Ukufunda ngokucubungula*	IsiZulu	Paraphrase
References	*Ukukhonjiswa kolwazi oluthathwe emibhalweni yabanye abantu*	IsiZulu	Paraphrase
Bibliographic details	*Imininingwane yokushicilelwa*	IsiZulu	Substitution
Online assignments	*Ama-asayimenti atyelwa ngamakhompyutha (athunyelwa online)*	IsiZulu	Paraphrase
Critically discuss	*Xoxa ngokucubungula*	IsiZulu	Substitution
Distinguish	*Yehlukanisa*	IsiZulu	Substitution
Framework	*Uhlaka*	IsiZulu	Substitution

RiCL Research in Corpus Linguistics

Combining corpora with other language resources and tools in pedagogic audiovisual translation

Ruska Ivanovska-Naskova
Ss. Cyril and Methodius University in Skopje / North Macedonia

Abstract – This study focuses on the potential of combining various types of language resources and tools in pedagogic audiovisual translation in university level courses. It argues that the direct use of *ad-hoc* corpora compiled by students can be combined with other tools such as bilingual dictionaries, online resources and subtitling software in performing concrete translation tasks. The study reports on the positive results of the translation activity conducted with students of the degree course in the Italian Language and Literature program at the Ss. Cyril and Methodius University of Skopje in 2018. The first part of the study reflects on certain tendencies in the field of intersection between language pedagogy and audiovisual translation and presents concrete examples of this type of pedagogic tasks applied in teaching Italian as a foreign language. The central part of the study presents various aspects and stages of the activity: its aim, context, choice of video material, the complexity of the language of the videos, the tools used, the translation strategies, the creation of glossary, the revision of the subtitles and the discussion of the feedback. The study concludes with the results of the questionnaire and potential prospects for enhancing the task and reuse of the translated material for compilation of parallel corpus.

Keywords – corpus for pedagogic purposes; audiovisual translation; Italian as a foreign language; Macedonian

1. INTRODUCTION

Audiovisual translation (AVT) is defined as a "transfer of multimodal and multimedial content across languages and/or cultures" (Pérez-González 2020: 30). The multimodal aspect is related to the variety of signs that construct the message in the case of audiovisual material, such as language signs, images and music, while the multimedial nature refers to the different modes through which the message is mediated to the viewer. The recent technological development, the digitalization shift on a world scale and globalization tendencies have contributed to the introduction of new types of AVT, expansion of this field and its opening in the research topics towards adjacent domains (Pérez-González 2019, 2020). One of these areas of contact is language pedagogy,

undergoing itself a transformation driven by similar forces as in the case of AVT (Laviosa and González-Davies 2020). For example, the need to re-examine the role and potential of the L1 in language pedagogy in the process of learning a new one has renewed the interest in translation as pedagogic practice and has spurred the introduction of various types of translation-driven pedagogic practices, such as, for example, the use of AVT (Laviosa 2020).

The present paper reports on one teaching experience of the author of this article that uses the potential of AVT as a pedagogic practice in university context. It explores the possibility of using subtitling activities in the classroom both for language learning and for acquiring skills and knowledge related to various types of language resources and tools, such as dictionaries, corpora and subtitling software. Section 2 reflects upon some concepts that inspire the adopted pedagogical model and presents several examples of audiovisual translation activities in the teaching and learning of Italian as a foreign language. After these preliminaries, the core of the study is developed in Section 3, which describes various stages of the experimented activity: choice of input and analysis of the language, terminology search, subtitling, glossary compilation and review of translation. It also reflects on students' opinion about the activity and the prospects for its more frequent use as a teaching practice in degree courses. Finally, Section 4 provides some concluding remarks.

2. AUDIO-VISUAL TRANSLATION FOR PEDAGOGIC PURPOSES

The growing interest of using AVT in the language pedagogy is closely related to the recent development of the wider category of multilingual pedagogies that involve the use of two or more languages in the teaching and learning process (Laviosa 2020: 272). Multilingual pedagogies are becoming more important in today's plurilingual reality, conditioned by migration, movement and the need to create links between languages in real life and in the learning process. As Laviosa (2015: 85) points out:

> [...] multilingual pedagogy [...] has considerable potential for developing the ability to operate between languages, allowing learners to enter the traffic of meaning and preserving global semiodiversity and glossodiversity. In order to unlock the untapped potential of multilingual teaching methods, it is crucially important to carry out interdisciplinary research that brings together scholars and educators working in literary, film and media studies as well as many convergent fields of applied linguistics.

Pedagogic translation is the domain where the interaction between translation studies and educational linguistics occurs. The exchange between these two fields stimulates the use of innovative and motivating practices in language teaching and learning (Laviosa 2019: 196). As pointed out in Laviosa (2020), under this perspective language learning is seen as an action:

> in the classroom context, this language as an action perspective means that learners engage in meaningful activities as varied as projects, presentations and investigations. These activities are intended to engage students' interest and encourage language growth through perception, interaction, planning, research, discussion and the co-construction of academic output of various kinds. During such action-based work, language development occurs when it is carefully scaffolded by the teacher as well as by the students working together (Laviosa 2020: 273).

One of the translation pedagogies that has the potential to involve students in such activities is AVT. The interest in using AVT in the language classroom coincides with the growing in corpus of subtitled and revoiced text on a global scale in the last decades and, consequently, with the growing importance of AVT within translation studies (Bolaños-García-Escribano and Díaz-Cintas 2020). The particularity of AVT —both as a translation activity on its own and as pedagogic practice— relies on the multimodal and multimedial nature of audiovisual texts. They create challenges both in terms of their comprehension and translation, and in terms of the presentation of the translated texts in an appropriate mode.

Translators today make use of various types of resources and tools (O'Brien and Rodríguez Vázquez 2020). Reflecting on the education of translators in the field of AVT, Bolaños-García-Escribano and Díaz-Cintas (2020: 211) draw attention to the need for the development of different abilities related to digital technologies:

> in the particular case of AVT training, the main difference with other translation specialisms, be they literary or non-literary, lies in its multimodal and multimedia nature, which calls for transversal abilities closely related to digital technology and audiovisual literacy [..] [T]he instrumental competence seems to be particularly relevant in the case of AVT courses as it entails the mastery of AVT-specific software and the ability to work with a plethora of multimedia files and technologies.

The *Process in the Acquisition of Translation Competence and Evaluation* research group (PACTE 2005) defines the need to develop the instrumental competence in

translators as "knowledge related to the use of documentation sources and information technologies applied to translation" (PACTE 2005: 611), which recalls the idea of developing skills in students to use language resources and tools in and outside the language classroom. This is related to the growing interest in using computers and technology in language learning and teaching, which, in the last decades, has given rise to the well-established domain of Computer-aided Language Learning (CALL).

Reflecting on the link between technology, translation and language teaching and learning on a broader scale, Enríquez Raído *et al*. (2020) suggest the new term 'Computer-assisted L2 Learning and Translation' (CAL2T). This further articulation of the field of CALL in the areas of intersection with translation studies is again related to idea to use translation as pedagogic practice, as noted in Enríquez Raído *et al*. (2020: 278)

> we propose to introduce the term computer-assisted L2 learning and translation (CAL2T) with the aim of (1) re-conceptualizing L2 translation as a core skill in contemporary translator training, and (2) re-evaluating the pedagogical potential of L2 translation to further foster linguistic and intercultural mediation skills in other learning contexts involving the use of a second, or additional, language.

As far as AVT is concerned, the idea of involving students in concrete subtitling tasks has been explored in various studies (Incalcaterra McLoughlin and Lertola 2014, 2015; Incalcaterra McLoughlin 2019; Ivanovska-Naskova and Talevska 2021). They focus on the possibility of using subtitling activities for acquiring particular language knowledge and developing skills in students, such as writing skills, pronunciation, independent learning or translation skills (Incalcaterra McLoughlin 2019: 486). As Incalcaterra McLoughlin (2019: 488) argues, the most recent studies encourage the introducing of these practices in the curriculum in a systematic manner:

> current research in this area is prioritizing the systematic integration of AVT tasks in the language curriculum. By providing detailed guidance on how to achieve an optimal level of curricular integration, specialists aim to ensure that AVT tasks are no longer dealt with as isolated add-ons, but combined with an array of pre- and post-task activities to help learners elicit and recall information and assist trainers with the feedback delivery process.

As far as future directions of research in this field, Incalcaterra McLoughlin (2019: 492) concludes that the studies need to expand in terms of language pairs, learning environments and degree of students' familiarity with AVT. As in the case of translators'

education, students need to develop skills to use new tools, to combine them and to keep up with the fast-changing technological environment.

As far as audiovisual translation and the teaching of Italian as a foreign language is concerned, the number of studies dealing with AVT in educational contexts is constantly growing. In what follows, we briefly discuss some examples that are related to teaching Italian as a foreign language, which is also the context of the model presented in this paper.

Incalcaterra McLoughlin and Lertola (2014: 75) postulate a pedagogical model based on subtitling activities to be fully integrated in the foreign language curriculum. The model, grounded in the *Common European Framework of Reference for Languages*,[1] involves subtitling activities conducted in five stages: 1) motivation (presentation of the subtitling activity); 2) global perception (showing of the L2 audiovisual input); 3) analysis (deconstruction and comprehension of the L2 input); 4) synthesis (translation and subtitling of the video); and 5) reflection (discussion on the subtitling activity). This model has been implemented at the National University of Ireland (Galway) in four consecutive academic years, as part of a regular language course of Italian as a foreign language. Its evaluation —based on the students' opinion about their subtitling experience— is positive since they find subtitling a very motivating activity. Incalcaterra McLoughlin and Lertola (2014) encourage the use of this flexible model to other language pairs, learning environments, students' levels and learning objectives.

In another study, the same authors focus on the use of the platform *ClipFlair*[2] for teaching Italian in an online environment (Incalcaterra McLoughlin and Lertola 2015; Romero 2015). This platform was specifically designed in an EU-funded project to be used in foreign language teaching and learning through subtitling and revoicing of video material and can be used online for classroom or distance learning environments or autonomous learning. The model combines various types of intralingual subtitling activities, such as reordering of subtitles, inserting missing words or keywords, transcribing subtitles, and revoicing activities.

In a previous study, Lertola (2012) explores the possibility of using subtitle activities for vocabulary building. The subtitling task is integrated in an Italian studies

[1] https://www.coe.int/en/web/common-european-framework-reference-languages
[2] http://clipflair.net/

curriculum in a university context and is piloted with a small group of learning students. Lertola (2012: 69) concludes that the results of this small-scale experimental subtitling activity are in line with:

> the positive results obtained in recent studies on the use of the subtitling practice as an effective pedagogical tool in the FL class, and it [the subtitling activity] greatly encourages further research on the topic.

Laviosa (2015) explores the possibility of developing the translingual and transcultural competence through subtitling. The paper reports on a subtitling activity as a pedagogic methodology in a professional development course for secondary school EFL teachers and demonstrates that the proposed translating practices can be successfully integrated in language learning and teaching.

The study by Ivanovska-Naskova and Talevska (2021) is another example of subtitling activity whose aim is to increase students' awareness about the culture of the language they study and the differences with their own. This pedagogic activity, which mirrors Laviosa's (2015) idea that translation can be used for the acquisition of linguistic and transcultural competences, involves the subtitling of a documentary film about the feminist movement in Italy in the 1960s and 1970s. The video, which contains numerous cultural elements, proves very stimulative for the students who appreciate the possibility to learn about the culture they study through subtitling, and to present their translations to a wider audience at the screening of the documentary at a local festival.

Finally, Romero Ramos (2012) examines the idea of using subtitling activities for learning one variety of Italian, namely the Neapolitan dialect. This pedagogic activity is conducted with the *Learning via Subtitling* software (LVS; Romero Ramos 2012), specifically designed to be used in foreign language teaching and learning. Beside the area for viewing the video material and editing the subtitles, this software integrates two more windows: one for viewing the instructions for the activity and a communication window for the teacher and the student connected with the subtitles. The activity had a positive impact on the students' knowledge about the specific variety and their motivation to learn the language through subtitling.

3. TRANSLATION TASKS: SUBTITLING SHORT VIDEOS IN THE CLASSROOM AND COMPILATION OF A GLOSSARY

The present paper reports on a classroom experience inspired by recent tendencies in language pedagogy and translation presented in Section 2. It specifically involves two translation tasks that combine both individual and collaborative work: subtitling video-material and the compilation of a glossary. The learning objective is to strengthen the instrumental competence in university students of modern languages, in particular the development of skills related to terminology search, terminology management and subtitling. The key idea underpinning this activity is that various types of language resources and tools such as dictionaries, encyclopedias, search engines, corpora and translation software can be successfully combined in pedagogic translation at a university context.

3.1. Outline of the activity and the context

The translation activity consisted in subtitling short videos from Italian to Macedonian and in the compilation of a glossary. It was introduced in a double-semester module on software and tools for language learning, language teaching, and translation. This elective module is part of a four-year degree course in Italian Language and Literature at the Ss. Cyril and Methodius University of Skopje. The course includes activities with different types of language resources, such as electronic dictionaries, virtual libraries, online exercises software, terminology management tools, corpora, concordancers, tools for computer-assisted translation, and subtitling. In the module, the activities are mostly performed with freeware or a demo-versions of commercial software. The course is intended for students from both translation/interpreter and teacher stream with a B2-language level in their third academic year.

The translation activity was carried out in the academic year 2017–2018. 14 Macedonian-speaking students were involved in eight sessions of 45 minutes each. A small part of the translation assignment was carried out as homework.

The activity had several stages. In the first stage the teacher presented the activity, the resources, and the input. The students became acquainted with the language of the videos and analyzed the terminology used. By drawing upon previous translation experience, they outlined various translation strategies in a group discussion. In the

second stage, each student translated one video. The teacher reviewed the translation, provided the students with feedback, and engaged them in a group discussion on the translation strategies by drawing their attention to specific translation examples. In the third stage, the students finalized the subtitles by confirming or refining their initial choices in light of the discussion and teacher's comments. They also worked collaboratively in a shared document on the creation of the glossary.

3.2. Translation input

The translation input consisted of short educational videos on various topics related to art. The decision to focus the activity on this specific linguistic genre is closely related to its importance when it comes to the study of Italian as a foreign language. The Italian language is mainly studied because it is strongly associated with the Italian culture and its art (Magnatti 2016; Pizzoli 2018). Topics related to art are also present, though to a minor extent, in materials for beginners and they are more frequent in upper levels of language command.[3] Another motivation for the choice of the language of art is its significance when it comes to the spread of Italianisms in other languages. Some of the most frequent Italian loanwords in other languages derive from the semantic field of art (Pizzoli 2018: 154).

The complexity and the heterogeneity of the Italian art language depends on different factors (Biffi 2010). It is related to the osmosis between the language of art and other types of language, such as the standard or the literary language, the great diversity of texts that deal with arts (technical texts, reviews, inventories, archive materials, notes, and literary texts) and the existence of subfields (painting, sculpture, architecture, and restoration). The specific nature of the language of art is mainly related to its lexis rather than to its morphosyntactic structure.

Because of its importance and its complexity, the language of art poses many challenges both to the teachers and the students of Italian as a foreign language. The

[3] See the syllabus for the PLIDA certification for Italian-language proficiency (*Sillabo della Certificazione PLIDA*) at https://plida.it/certificazione-plida/documenti.html.

general manuals for Italian address this issue through texts that deal with art topics. Also, several specialized manuals have been published in the last decade.[4]

The input consisted in 14 short educational videos (1–1.5 min. each) of the portal Treccani Scuola on various topics related to art.[5] At the time when the activity was conducted, this portal contained educational videos on various topics related to natural and social sciences and art. The playlist *Arte*, for example, offered more than 250 short videos about artistic movements, artworks, and artists. The video material was chosen for its quality, duration, and appropriateness for the language level of the students. The verbal and the visual message in the input complement each other. The main challenges in the process of translation related to the input regard the terminology used, the density of terms and the speed of the speech.

The language of the videos presents some particular features both at the structural and the lexical level. As far as the grammar is concerned, some of the most frequent features are the use of the *passato remoto*, impersonal verbs, the passive voice, and relative clauses. Lexis, however, is the most salient feature of the language of the videos. In line with Biffi's (2010) general outline of the composition of the lexis in Italian art discourse, a large part of the lexis consists of art and architecture terminology and collocations (for illustrative examples, see column 1 in Table 1). The second largest lexical group consists of general language words (see examples in column 2 of Table 1). The third and the smallest group involves terms from other domains, such as geometry, religion or history (see examples in column 3 of Table 1). The examples are classified according to the information about the distribution of the lexical items in Italian as presented in the *Dizionario della lingua italiana De Mauro*.[6] The examples in Column 1 are designated with the acronyms TS arch (*Tecnico-specialistico architettura* 'Technical-specialist architecture') and TS arte (*Tecnico-specialistico arte* 'Technical-specialist art'), the terms in Column 2 with FO (*Fondamentale* 'Fundamental'), AU (*Alto u*so 'High use'), AD (Alta disponibilità 'High availability'), and CO (Comune 'Common), while those in Column 3 with acronyms for other domains, such as history, geometry and

[4] For the general manual see Piantoni *et al.* (2017), for the specialized manuals see Bigliazzi *et al.* (2013) and Andriuzzi (2017).
[5] The videos are available at https://www.youtube.com/c/TreccaniScuola (accessed on 19 June 2018).
[6] https://dizionario.internazionale.it/

religion. In general, the density of terms is higher in videos devoted to architectonical works, which makes them more difficult to translate when compared to others.

Art language	**General language**	**Other domains**
Navata 'nave'	*Arco* 'arch'	*Ellittico* 'elliptic'
Transetto 'transept'	*Disegno* 'drawing'	*Rettilineo* 'rectilinear'
Cripta 'crypt'	*Contorno* 'outline'	*Triangolare* 'triangular'
Ábside 'apse'	*Proporzione* 'proportion'	*Oncentrico* 'concentrico'
Campata 'span'	*Decorazione* 'decoration'	*Sfera* 'sphere'
Contrafforte 'buttress'	*Ritratto* 'portrait'	*Consacrare* 'to consecrate'
Complesso 'complex'	*Pilastro* 'pilar'	*Canonizzazione* 'canonization'
Tamburo 'tambour'	*Affresco* 'fresco'	*Minoico* 'Minoan'
Calotta 'calotte'	*Culto* 'cult'	
Lanterna 'lantern'	*Fondo* 'background'	
Coro 'choir'	*Maiolica* 'majolica'	
Costolone 'rib'	*Facciata* 'façade'	
Pronao 'pronaos'	*Nicchia* 'niche'	
Oculus 'oculus'	*Composizione* 'composition	
Iconografía 'iconography'	*Simbólico* 'symbolic'	
Gotico 'Gothic'	*Astratto* 'abstract',	
Chiaroscuro 'chiaroscuro'	*Retrostante* 'at the back (of)'	
Volta 'vault'	*Ieratico* 'solemn'	
Allegoria 'allegory'	*Bizantino* 'Byzantine'	
Arco rampante 'flying buttress	*Rinascimentale* 'Renaissance'	
Chiesa a una navata 'single nave church'	*Classico* 'classical'	
Arco a tutto sesto 'round arch'	*Augusteo* 'of Augustus'	
Cupola a sesto acuto 'pointed dome'	*Romántico* 'Romantic'	
Edificio a pianta centrale 'central-plan building'		
Colonna corinzia 'Corinthian column'		

Table 1: Examples of the terminology in the subtitled videos

3.3. Language resources and translation tools

In the translation process several tools were combined. The students used various paper and electronic dictionaries and terminological databases, such as the *Italian-Macedonian Dictionary* [Italijansko-makedonski rečnik] (2015, paper edition), *Dizionario della lingua italiana De Mauro* (Internet edition),[7] the set of dictionaries available at the site of the newspaper *Corriere della Sera*[8] and the portal *Wordreference*.[9]

With the use of texts from online sources, the students also compiled small-size comparable corpora with *AntConc* (Anthony 2014).[10] They also conducted free research

[7] https://dizionario.internazionale.it/
[8] https://www.corriere.it/
[9] https://www.wordreference.com/
[10] The students got first familiar with the basic functions of *AntConc* while performing another translation task in the same module related to the translation of legal texts from Italian to Macedonian.

on the web. The software *Subtitle Workshop* was used for the subtitling.[11] The translation process did not follow any strict pattern and the students were free to start and carry out their research as they thought was best for a given term.

The students received teacher's feedback about the quality of the translation and assistance with technical issues regarding the subtitles. Upon examining the feedback individually and in group discussion, they revised their translations and handed in the final version of the videos (see one example of the parallel text in Appendix 1). The videos were published in a private *YouTube* profile and shared with the rest of the group (see Appendix 2).[12] In the final stage of the activity, the students created collaboratively a bilingual glossary of more than 150 terms with *Google Drive*. Some of the terms of this shared resource are linked to the translated videos in those contexts in which the given term is used (see Appendix 3).

3.4. Students' translation

The review of students' translations showed that they had understood the texts of the videos and that they managed to conduct even complex terminological research. The majority of the terms and collocations had been translated correctly, such as the following examples: *abside – апсида* 'apse', *peristilio – перистил* 'peristyle', *transetto – попречен кораб* 'transept' */ трансепт*, *contrafforte – потпорен ѕид* 'buttress' and *volta a crociera свод во вид на крст / крстовиден свод* 'rib vault'.

Example (1) illustrates some of the challenges the students faced during translation:

(1) La cupola, divisa a spicchi, viene realizzata sopra un *tamburo* ottagonale intervallato da finestre circolari che illuminano l'interno. Sulla sommità viene posta una *lanterna* utile a conferire una maggiore stabilità (*La cupola di Santa Maria del Fiore di Filippo Brunelleschi a Firenze*, Treccani Scuola).

(1a) Куполата поделена на делови е изградена над осмоаголна *цилиндрична конструкција* разделена со округли прозорци коишто ја осветлуваат внатрешноста. На врвот е поставен *светилник* корисен за да даде поголема стабилност (traduzione di E.S.).[13]

[11] http://subworkshop.sourceforge.net/ (accessed on 7 September 2022).
[12] Copy-right issues and the migration of the video content of *Treccani Scuola* from *YouTube* to its own portal prevented publishing the translated videos separately on *YouTube*.
[13] The Macedonian translations in the section are given in their initial and not in their reviewed version. See Appendix 1 for an example of a full text of one of the videos and its final version of the translation.

The first type of error is represented by the translation of the term *lanterna* 'lantern': the student did not understand that the word has a particular meaning in the architectonic discourse and, instead of using the correct Macedonian term *lanterna*, s/he translated it with its common meaning (*svetilnik* 'lighthouse'). In the same example, the term *tamburo* 'tambour' is translated with the syntagm *cilindrična konstrukcija* ('cylindrical construction'), which is an approximative translation with respect to the more precise term *tambur*. Other similar examples are *redica stolbovi* ('sequence of columns') instead of *kolonada* for the term *colonnato* 'colonnade', *izvor na svetlina* ('source of light') instead of *okulus* for the Italian term *oculus* or *oculo* 'oculus', *svetlo temno* ('bright dark') instead of *kjaroskuro* for the Italian term *chiaroscuro* ('chiaroscuro'). The main reason for these errors may be that, although the terms are used in specialized texts in Macedonian, they are not present in the Macedonian dictionaries, so the students experienced difficulties in finding the exact equivalents.

The case of the names of art works is also interesting. In most cases the translation was correct, and this was mainly due to the fact that the described art works are worldwide known. As shown in (2), the names were usually transcribed: for some of them this was the most suitable solution (*Pjeta* for Michelangelos's *Pietà* 'Michelangelo's Piety'), while for others, the corresponding name should have been used instead (*Bah* or *Bahus* in Macedonian for Michelangelo's Bacco, and not *Bako*). In the case of Bramante's Tempietto, the student decided to include both the translation (*mal hram* 'little temple') and its transcription (*Tempijeto*).

(2) L'oprera che meglio esemplifica il risultato di queste ricerche è il Tempietto di San Pietro in Montorio (*Il linguaggio classico e il modulo vitruviano: L'attenzione per il voltoIl tempietto di S. Pietro in Montorio di Bramante*, Treccani Scuola).

(2a) Делото што најдобро го илустрира резултатот од овие проучувања е малиот храм (Темпијето) на Св. Петар во Монторио (traduzione di T.M.).

At the structural level, the errors concern cases in which the boundaries between the utterances were not identified correctly. The texts of the videos are quite short, but dense with information, with small or no pauses between the utterances, which causes difficulties in identifying the structure of the discourse and reconstructing it in the source-text. This is the case in (3), below. The dependent clauses in the target-text that form one utterance with the main clause are separated in the Macedonian version, which might compromise its understanding in the target-text.

(3a) Quest'ultima è chiusa *da due bracci rettilinei leggermente divergenti per restringere la visuale* della facciata e farla sembrare più stretta e più alta (*Gli spazi sacri: Colonnato e piazza di San Pietro in Vaticano a Roma*, Treccani Scuola)

(3b) Овој вториот е затворен *од две малку дивергентни правоаголни раце. За да го ограничи погледот* на црковната фасада и да направи да изгледа потесен и повисок (traduzione di E.P.).

3.5. Class discussion and questionnaire

The students' answers to a questionnaire, which were collected at the end of the activity (see Appendix 4), reveal that students consider the experience very positive and motivating, mostly because they learned how to translate specialized discourse. They like the fact that the activity involved translation of video material and felt challenged to perform the task. They point out that, in general, they understood the content of the videos and that they feel to have expanded their knowledge about the topic of the video. Their answers further reveal that the most useful resources for the terminology search were the bilingual dictionary, the multilingual databases, and the specialized texts.

The students adopted various strategies in the translation process. Usually, they started their search by consulting the bilingual dictionary or some multilingual online resource, frequently engaging also a third language, and continued with testing their hypothesis for translation equivalents in various types of specialized texts. The students state that the *ad-hoc* corpora they created were useful in few cases in the phase of checking the translation equivalent through identification of examples with that particular word in authentic texts. Still, they consider the corpora as one of the outputs in the translation process that can be reused for similar translation tasks in the future. The main challenges with the process of corpus compilation and, in general, with the authentic specialized texts is the difficulty to find such texts and to decide whether the text has been automatically translated from another language.

All the students appreciate the collaborative work on the glossary: they think that this type of work is timesaving on a long run, it creates larger and thus more useful resources in comparison to individual work, and they find it fun. They also think that the teacher's comments and class discussion were very useful and helped them significantly

to improve the final version of their translation. The main difficulty in the translation process in general regards the terminology research, especially the difficulty to render rare terms in the target language. Some of the students faced difficulties during the final stage of the translation process, such as technical issues regarding the use of the subtitling software and the lack of practical skills for creating and synchronizing the subtitles. Most of them shared the view that they would like to have more activities of this type in their studies.

4. Conclusions

Although it is not possible to draw definite conclusions because of the limited nature of this study, this teaching experience is in line with previous studies in this field, which conclude that the AVT is a stimulating pedagogic practice for the students. The fact that students feel that they have gained new knowledge both in terms of the vocabulary and the topic of the video is another important point that encourages this type of pedagogic practices also for spurring the intellectual curiosity in students and their personal growth. As far as the instrumental competence is concerned, the students' answers and the quality of the translations confirm that the objective of the activity was achieved and that the students appreciate the possibility to perform activities with different tools and especially.

As far as the corpora are concerned, this study shows that, particularly in the case of this rare language pair, they cannot be the only resource in the translation process. Nevertheless, they can be combined with other resources and reused later on in other translations task. Another important conclusion that can be drawn is that the activity raised the students' awareness of the need to develop skills to work with various resources and tools, to create reusable resources by themselves or collaboratively and to share them.

As also suggested in previous studies (Incalcaterra McLoughlin and Lertola 2014), this type of translation task can be included in other degree courses which are focused on the development of the instrumental competence and also in translation. When introducing this type of activity, attention should be paid to the selection of the input and the tools used. The input should be stimulative, comprehensive and the tools adopted, should be up to date. More time and attention should be paid to the creation of the corpus and to the development of critical skills in evaluating the reliability of the source of the texts. In the case of this specific translation task and in this particular context, some

modifications can be introduced. There are loanwords from Italian in the art and architecture terminology in Macedonian and some examples can be used to establish links between the two languages and to improve students' motivation in the first phase (Saržoska and Ivanovska-Naskova 2021: 208). Moreover, the video texts present some recurring patterns which can be used as examples for stimulating metalinguistic reflection. Another similar pedagogic experience (Ivanovska-Naskova and Talevska 2021) showed that translation activities are particularly motivating and rewarding when students know that their translation will be made public. The input could consist of text or material that can be freely published online or presented in other forms with its translation. This would significantly increase the motivation and the responsibility of students. Another possible development of the activity is the creation of parallel corpora with the translated texts. These corpora can grow with translations of various generations of students, thereby becoming a valuable pedagogic and research tool.

REFERENCES

Andriuzzi, Rossana. 2017. *L'italiano dell'arte: Corso di Lingua Italiana*. Milano: Hoepli.

Anthony, Laurence. 2014. *AntConc* (Version 3.2.4). Tokyo: Waseda University. https://www.laurenceanthony.net

Biffi, Marco. 2010. La lingua dell'arte e critica dell'arte. In Raffaelle Simone, Gaetano Berruto and Paolo D'Achille eds. *Enciclopedia dell'Italiano.* Roma: Istituto della Enciclopedia Italiana, 106–108.

Bigliazzi, Maria Silvia, Mariella Colombini and Massimiliana Quartesan. 2013. *La Lingua dell'Arte per gli Studenti Stranieri: Arte, Moda e Design*. Siena: Becarelli.

Bolaños-García-Escribano, Alejandro and Jorge Díaz-Cintas. 2020. Audiovisual translation: Subtitling and revoicing. In Sara Laviosa and Maria González-Davies eds., 207–225.

Enríquez Raído, Vanessa, Frank Austermühl and Marina Sánchez Torrón. 2020. Computer-assisted L2 learning and translation (CAL2T). In Sara Laviosa and Maria González-Davies eds., 278–299.

Incalcaterra McLoughlin, Laura. 2019. Audiovisual translation in language teaching and learning. In Luis Pérez-González ed. *The Routledge Handbook of Audiovisual Translation*. London: Routledge, 483–497.

Incalcaterra McLoughlin, Laura and Jennifer Lertola. 2014. Audiovisual translation in second language acquisition: Integrating subtitling in the foreign-language curriculum. *The Interpreter and Translator Trainer* 8/1: 70–83.

Incalcaterra McLoughlin, Laura and Jennifer Lertola. 2015. Captioning and revoicing of clips in foreign language learning using Clipfair for teaching Italian in online learning environments. In Catherine Ramsey-Portolano ed. *The Future of Italian Teaching: Media, New Technologies and Multi-Disciplinary Perspectives*. Newcastle upon Tyne: Cambridge Scholars Publishing, 55–69.

Ivanovska-Naskova, Ruska and Irina Talevska. 2021. La traduzione audiovisiva e l'insegnamento dell'italiano LS: Un'esperienza didattica. In Margarita Borreguero Zuloaga ed. *Acquisizione e Didattica dell'italiano: Riflessioni Linguistiche, Nuovi Apprendenti e uno Sguardo al Passato* (Vol. 1). Berlin: Peter Lang, 649–658.

Laviosa, Sara. 2015. Developing translingual and transcultural competence through pedagogic subtitling. *Linguaculture* 1: 72–88.

Laviosa, Sara. 2019. Translanguaging and translation pedagogies. In Helle V. Dam, Matilde Nisbeth Brøgger and Karen Korning Zethsen eds. *Moving Boundaries in Translation Studies*. London: Routledge, 181–199.

Laviosa, Sara. 2020. Language teaching. In Mona Baker and Gabriela Saldanha eds. *Routledge Encyclopedia of Translation Studies* (third edition). London: Routledge, 271–275.

Laviosa, Sara and Maria González-Davies. 2020. Introduction: A transdisciplinary perspective on translation and education. In Sara Laviosa and Maria González-Davies eds., 1–8.

Laviosa, Sara and Maria González-Davies eds. 2020. *The Routledge Handbook of Translation and Education.* London: Routledge.

Lertola, Jennifer. 2012. The effect of the subtitling task on vocabulary learning. In Anthony Pym and David Orrego-Carmona eds. *Translation Research Project 4*. Tarragona: Universitat Rovira i Virgili, 61–70.

Magnatti, Michele. 2016. La didattica dell'arte a stranieri: Esperienze e riflessioni. *Bollettino Itals* 63: 61–78.

O'Brien, Sharon and Silvia Rodríguez Vázquez. 2020. Translation and technology. In Sara Laviosa and Maria González-Davies eds., 264–277.

PACTE = *Process in the Acquisition of Translation Competence and Evaluation*. 2005. Investigating translation competence: Conceptual and methodological Issues. *Meta* 50/2: 609–619.

Pérez-González, Luis. 2019. Rewiring the circuitry of audiovisual translation: Introduction. In Luis Pérez-González ed. *The Routledge Handbook of Audiovisual Translation.* London and New York: Routledge, 1–12.

Pérez-González, Luis. 2020. Audiovisual translation. In Mona Baker and Gabriela Saldanha eds. *Routledge Encyclopedia of Translation Studies* (third edition). London: Routledge, 30–34.

Piantoni, Monica, Chiara Ghezzi and Rosalla Bozzone Costa. 2017. *Nuovo Contatto B2: Corso di Lingua e Civiltà Italiana per Stranieri*. Torino: Loescher.

Pizzoli, Lucilla. 2018. Italiano e italianismi nel mondo: Osservazioni sulla ricerca dei neologismi. In Raffaella Bombi ed. *Italiano nel Mondo: Per una Nuova Visione.* Udine: Forum, 151–158.

Romero Ramos, Lupe. 2012. L'uso dei sottotitoli per l'apprendimento delle varietà regionali dell'italiano: Un'esperienza didattica con il programma LvS. *inTRALinea*. *Special issue: The Translation of Dialects in Multimedia II*. http://www.intralinea.org/specials/article/sottotitoli_per_apprendimento_dellitaliano (21 March 2022).

Romero, Lupe. 2015. L'uso del doppiaggio e del sottotitolaggio nell'insegnamento della L2: Il caso della piattaforma ClipFlair. *Lingue Linguaggi* 15: 277–284.

Saržoska, Aleksandra and Ruska Ivanovska-Naskova. 2021. I neoitalianismi in macedone tra dizionari e testi. In Saržoska, Aleksandra ed. *Atti del Convegno Internazionale L'italianistica nel Terzo Millenio: Le Nuove Sfide nelle Ricerche Linguistiche*,

Letterarie e Culturali. 60 Anni di Studi Italiani all'Università Ss. Cirillo e Metodio di Skopje. Skopje: Università Ss. Cirillo e Metodio di Skopje, 207–217.

Corresponding author
Ruska Ivanovska-Naskova
Ss. Cyril and Methodius University in Skopje
Blaže Koneski Faculty of Philology
Department of Italian Language and Literature
Goce Delčev 9a
1000 Skopje
North Macedonia
Email: rivanovska@flf.ukim.edu.mk

received: November 2022
accepted: February 2023

APPENDICES

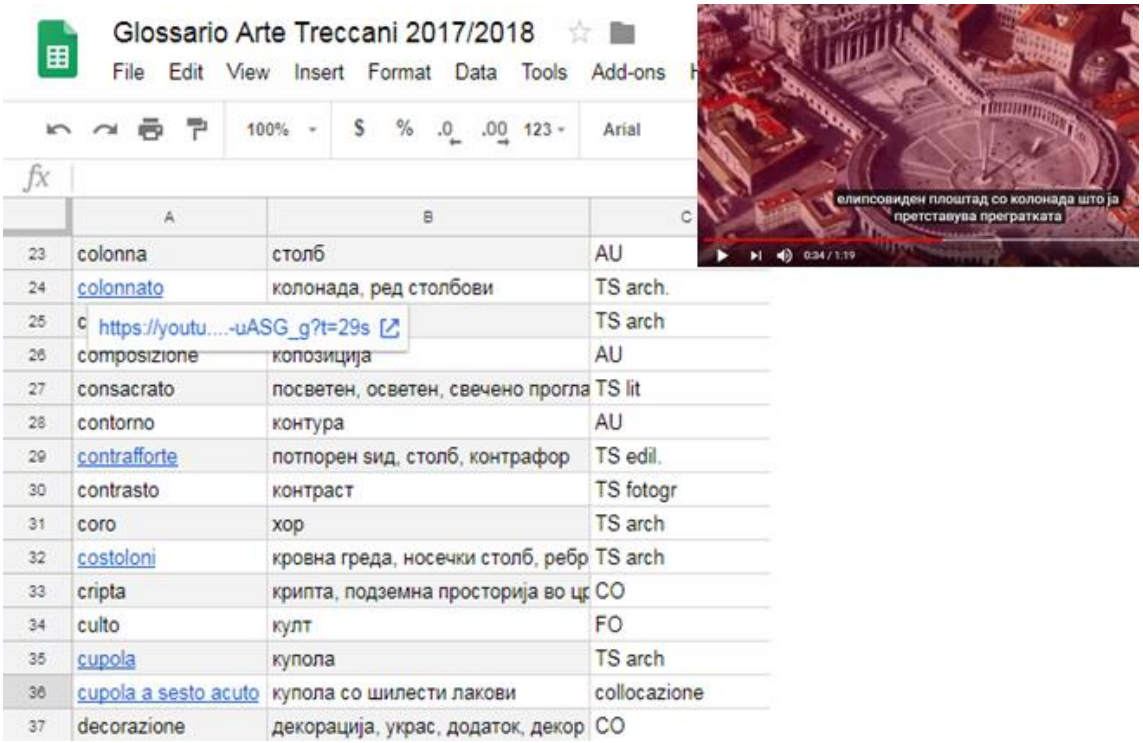
Appendix 1: Example of the text of one video and the final version of the translation

Basilica di San Francesco ad Assisi	Базиликата на Св. Франциск во Асизи
Nel quadro della diffusione del Gotico in Italia particolare rilevanza assume la Basilica di San Francesco ad Assisi centro dell'ordine mendicante dei Francescani. L'edificio fu costruito per celebrare la canonizzazione di San Francesco due anni dopo la sua morte e consacrato nel 1253. Concepita per rispondere alla duplice finalità di luogo di sepoltura del santo e meta di pellegrinaggio, la Basilica è articolata in due livelli sovrapposti: una chiesa inferiore o cripta e una superiore destinata alla predicazione. Entrambe le chiese sono una sola navata con un transetto e un abside sorrette all'esterno da lunghi e contrafforti cilindrici e in basso da archi rampanti. L'aula inferiore di fatto funge da basamento all'ambiente soprastante. Lo si intuisce dalle proporzioni schiacciate delle ampie volte a crociera impostate su archi a tutto sesto e poggianti su pilastri bassi e massicci. Nella chiesa superiore la vasta e slanciata navata è divisa in quattro campate a base quadrata. Le campate sono coperte da volte ogivali rete da alti pilastri a fascio addossati alle pareti completamente coperte di affreschi.	Во рамките на ширењето на готиката во Италија, особена важност има базиликата на Св. Франциск во Асизи, центар на монашкиот ред на Францисканците. Градбата е изградена со цел да се прослави канонизацијата на Св. Франциск, две години по неговата смрт, а осветена е во 1253. Осмислена за да ја задоволи двојната намена за гроб на светецот и место за аџилак, базиликата е поделена на две нивоа поставени едно врз друго долна црква или крипта и горна црква за проповедање. Двете цркви се еднокорабни со трансепт и апсида однадвор потпрени на долги и цилиндрични потпорни столбови, а надолу на лакови. Долната просторија, всушност, е основата на горниот дел. Тоа може да се забележи од широките заоблени сводови во вид на крст поставени врз полукружни лакови и потпрени на ниски и масивни столбови. Во горниот дел широкиот извишен наос е поделен на 4 дорати со квадратна основа Доратите се покриени со шилести сводови, поткрепени од снопчести столбови на страните и целосно покриени со фрески.

(Translated by A. T.)

Appendix 2: Playlist of the translated videos published in a private *YouTube* profile

Appendix 3: Italian-Macedonian glossary of art terms



	A	B	C
23	colonna	столб	AU
24	colonnato	колонада, ред столбови	TS arch.
25	c		TS arch
26	composizione	композиција	AU
27	consacrato	посветен, осветен, свечено прогла	TS lit
28	contorno	контура	AU
29	contrafforte	потпорен ѕид, столб, контрафор	TS edil.
30	contrasto	контраст	TS fotogr
31	coro	хор	TS arch
32	costoloni	кровна греда, носечки столб, ребр	TS arch
33	cripta	крипта, подземна просторија во цр	CO
34	culto	култ	FO
35	cupola	купола	TS arch
36	cupola a sesto acuto	купола со шилести лакови	collocazione
37	decorazione	декорација, украс, додаток, декор	CO

Appendix 4: The questionnaire

1. Was the video translation a useful activity? Why do you think so?

2. Did you appreciate the collaborative creation of a glossary? Please explain why.

3. Was the feedback on the translation useful?

4. Did you have any prior knowledge about the topic of the video?

5. Did you experience any difficulties regarding the comprehension of the video?

6. What resources did you use in the translation?

7. What was the main difficulty you experienced during the activity?

8. Would you like to have more often similar translation tasks in your studies?

9. Please add any comment you might have.

Thank you.

RiCL Research in Corpus Linguistics

Review of Moessner, Lilo. 2020. *The History of the Present English Subjunctive: A Corpus-based Study of Mood and Modality*. Edinburgh: Edinburgh University Press. ISBN: 978-1474-43801-8. https://doi.org/10.1515/9781474438018

Erik Smitterberg
Uppsala University / Sweden

The development of the subjunctive mood in English is of great interest to language historians not only in English linguistics but also from a cross-linguistic perspective, since this mood has evolved along different paths in the Germanic languages. Moessner's corpus-based book-length study, which treats the present subjunctive in Old, Middle, and Early Modern English (henceforth OE, ME, and EModE, respectively), is thus a very welcome contribution to research. The book contains six main chapters and an epilogue; I will first provide brief and selective summaries of these main sections of the monograph before proceeding to the evaluation of Moessner's findings.

The first, introductory chapter creates a research space for the book, discusses the subjunctive as a concept and its treatment in previous work, addresses Moessner's choice of corpus and method, and outlines the structure of the book. Moessner considers the subjunctive to be "identified by its form as a realisation of the category mood" and to express "one of several kinds of root modality" (p. 241). Root modality occurs "when an illocutionary act is intended to get the world to match the words" (p. 12). Moessner uses a subset of the *Helsinki Corpus* as her material is based on, among other things, the high degree of representativity of that corpus.[1] However, as she did not have access to a comparable corpus that bridges the gap between EModE and Present-Day English, she does not include the period after 1710 in her study. As regards retrieval, Moessner opts

[1] https://varieng.helsinki.fi/CoRD/corpora/HelsinkiCorpus/

for manual scrutiny and uses close readings to identify subjunctives and their competing variants in the texts examined to ensure high recall and precision (p. 17). These competing variants are the imperative, the indicative, and modal constructions,[2] though not all of them are relevant to each clause type. Moessner excludes from the scope of her study (i) forms of the verb *be* where the subjunctive and indicative are distinct but where other verbs do not have an equivalent distinction, and (ii) past-tense verb forms, with the rationale that the distribution of the verb *be* and of strong and weak verbs with subjects of different persons and numbers might otherwise skew quantitative comparisons. Each relevant token is coded for a large number of relevant parameters; the exact parameters used depend to some extent on the clause type the verb phrase occurs in.

Chapters 2–5 address main, relative, nominal, and adverbial clauses, in that order; each chapter provides a period-by-period discussion of OE, ME, and EModE. The scope of investigation frequently differs with the period under study, as the subjunctive is formally distinct from other forms in more contexts in OE than in EModE; the normalised frequency of relevant verb phrases thus typically decreases over time.

Moessner begins Chapter 2, on main clauses, by identifying modal constructions and (in the second person) the imperative as the other variants considered, before proceeding to the analysis. The proportion of subjunctives decreases between OE and ME but not between ME and EModE, where the percentage of imperatives drops (a result that is partly related to the relative frequency of second-person and third-person contexts). Modal constructions account for an increasing share of relevant tokens; *shall* predominates in ME, while *can* and *will* become frequent in EModE. Statutory texts — and, in OE, prose texts and texts on religious instruction— favour the subjunctive throughout; in EModE, this is partly a result of the use of formulaic structures.

Chapter 3 is devoted to adnominal relative clauses, where the variants are subjunctives, indicatives, and modal constructions. The subjunctive begins to lose ground in this context even during the OE period, with 25 per cent of relevant tokens being subjunctives overall. In ME, the subjunctive accounts for a mere 4 per cent of the variant field, and it appears to virtually die out in relative clauses during the EModE period. Statutory texts favour the subjunctive in all periods, as do prose texts in OE and

[2] Some semi-modal constructions, as well as structures like OE *uton* + infinitive, are also included in the modal category in many of Moessner's tables.

ME. In OE and ME, restrictive relative clauses appear to favour the expression of root modality (modals or subjunctives), as do southern texts. In all periods, use of the subjunctive and/or modal constructions is promoted by an expression of root modality in the matrix clause, while epistemic modality in the matrix clause correlates with indicative verb phrases in the relative clause. Moessner interprets this pattern as a tendency towards 'modal harmony', a term borrowed from Huddleston and Pullum (2002: 179–180), with root or epistemic modality expressed in both clauses.

Nominal clauses are the topic of Chapter 4. Here the relevant variants are subjunctives, indicatives, modal constructions, and, in ME and EModE, imperatives.[3] Moessner excludes infinitive clauses from her analysis as she considers them to be in competition with finite *that*-clauses in general rather than with, for instance, the subjunctive in such finite clauses. Nevertheless, she acknowledges that the increased frequency of non-finite clauses may have contributed to the decline of the subjunctive (pp. 102–103). The proportion of subjunctives decreases between OE and ME and between ME and EModE; in OE and EModE, there are also decreasing proportions of subjunctives within the periods themselves. Statutory texts favour subjunctives in all three periods, as do *that*-clauses when compared with other clause types and —in OE and ME— texts on secular instruction and prose compared with verse. There are clear tendencies towards modal harmony at least in OE and ME. Moessner also supplies a detailed analysis of which functions of nominal clauses (subject, adjectival complement, etc.) favour the subjunctive in different periods.

Chapter 5 addresses adverbial clauses, which are categorised according to their semantic role. The results reveal a continuous decline of the proportion of subjunctives from OE via ME to EModE, although ME again stands out in evincing no steady development within the period itself. In OE and ME, Southern and Kentish texts, statutory texts, and texts providing secular instruction favour subjunctives. In the former period, prose is also more hospitable to the subjunctive than verse. Secular instruction continues to promote subjunctives in EModE, whereas modal constructions appear to replace it in statutory texts. As regards semantic roles, concessive and conditional clauses —and, in OE, clauses of purpose/result— have the highest percentages of

[3] Moessner (pp. 8, 126, 132–133) seems to analyse direct speech following a reporting clause as a nominal clause: if the direct speech takes the form of an imperative clause, it is thus included as an imperative nominal clause. In contrast, Huddleston and Pullum (2002: 1027) argue that, in such cases, the complement of the reporting verb "is not a subordinate clause of any kind" and that the structure "involves the embedding of a **text**, not of clauses as such" [emphasis original].

subjunctives overall. Individual conjunctions (e.g., *ær* in OE) or subclasses of semantic roles (e.g., posterior time in ME) may also prefer the subjunctive. In several contexts, an expression of volition in the matrix clause increases the likelihood of a subjunctive in the dependent clause.

Chapter 6 provides an overall picture of the subjunctive in each period, followed by a concise account of the subjunctive in OE, ME, and EModE taken together. The comparison of periods enables Moessner to show, for instance, that the subjunctive is the most prevalent in nominal clauses in OE and ME but in adverbial clauses in EModE relative clauses are the least hospitable to the subjunctive in all periods. The final section demonstrates, among other things, that the simplification of the verbal paradigm may not have affected the distribution of the subjunctive as much as might be supposed, as most subjunctives occurred in the third person singular, where the subjunctive and indicative have remained distinct, even in OE. Moessner also suggests that the decrease in modal harmony between matrix clauses that express root modality and nominal clauses in EModE might be seen as a reduction of redundancy: if the matrix clause already expresses root modality, subjunctive marking in the nominal clause is typically redundant. The diachronic tendency for infinitive clauses, which do not express mood, to replace finite *that*-clauses contributes to the same reduction in redundancy. In the brief epilogue that concludes the book, Moessner mentions the same tendency in adverbial clauses, where there is less change overall, but where the indicative may replace the subjunctive and reduce redundancy where the latter "did not add a meaning component" (p. 243). Moessner concludes by mentioning the need for a study of the subjunctive's development from Late Modern English onwards and some desiderata for such a study.

Moessner's book is clearly the result of careful and painstaking work. Detailed surveys of previous research, problems of classification, etc. accompany the presentation of results throughout. As regards her own data, based on Tables 6.1, 6.9, and 6.17 (pp. 203, 214, and 223, respectively), I conclude that the analysis covers a total of 14,254 verb phrases, 4,398 (31%) of which are subjunctives. Selecting that number of tokens through manual scrutiny and analysing them on a large number of contextual parameters must be considered a remarkable achievement. This method also allows Moessner to include contexts that are notoriously difficult to identify through

computerised searches without loss of recall even in tagged and parsed corpora, such as relative clauses with zero relative markers.

The comprehensive scope of Moessner's analysis also enables her to reach conclusions that would not necessarily have been apparent from an analysis of only one clause type. For instance, her enlightening account of modal harmony between matrix and dependent clauses clearly benefits from her being able to show that this phenomenon occurs in different types of dependent clause. Similarly, the special status of statutory texts as promoting the subjunctive stands out because this phenomenon recurs in several different linguistic contexts.

Another clear strength of Moessner's study is the mixture of quantitative analyses and detailed observations on individual tokens, genres, and so on. In several places, this sheds light on distributions of data that may otherwise have resisted explanation. For instance, her identification of a shift from inheritance-centred to testator-centred wills in OE enables her to account for the drop in directive speech acts in this text category during the period (pp. 36–38). Her close reading of the corpus texts doubtless facilitates reaching such insights.

There are also a few areas where the study could be improved. The first of these concerns the use of statistics. As made clear above, Moessner opts for a quantitative analysis using raw frequencies, percentages, and —where relevant— normalised frequencies, and includes an impressive number of independent variables that may affect the distribution of variants. However, it is very difficult to establish which of the potential factors have an independent effect on the distribution of variants when only one variable is considered at a time. Moessner does attempt to consider the simultaneous influence of several factors, for instance by considering tokens that feature several characteristics which favour the subjunctive (p. 115), but a multifactorial analysis of each variant field could have made this far clearer. It is possible that Moessner does not believe in null-hypothesis significance testing —see Koplenig (2019) for a recent critique of this practice— but the advantages and drawbacks of refraining from more detailed statistical analysis should have been discussed explicitly in the book.

Moessner's variant fields are another area where more explicit discussion would have been desirable. Although Moessner does not discuss her methodological choice in those terms, the main perspective of the book is implicitly variationist rather than text-

linguistic —in Biber *et al.*'s (2016) sense— in that the distribution of "the subjunctive and its competitors is measured in terms of the relevant verbal syntagms, not in terms of the size of the texts in which they are attested" (p. 20). Raw frequencies are thus primarily turned into percentages of occurrence rather than normalised frequencies.

One basic tenet of the variationist paradigm is that the variants of the linguistic variable should be different "ways of saying the same thing" (Tagliamonte 2012: 2). Yet it seems clear that not all the tokens included by Moessner are equivalent in this respect. As Moessner does not separate epistemic and root meanings of modal auxiliaries in main clauses, not all modal constructions included in tabulations can be replaced with subjunctives without change of meaning. As Moessner acknowledges, this makes ME and EModE figures for main clauses somewhat difficult to compare, as in ME *shall* dominates the distribution, while in EModE *will* and *can* —which are more likely than *shall* to be epistemic— are frequent (p. 56). Based on Figure 6.1 (p. 229), Moessner also draws conclusions as regards which variants have replaced the subjunctive by conflating results for her four clause types; but as the imperative is not an option in all clauses, cumulative percentages of these four variants do not necessarily provide a true picture of the actual choices language users made. Against this background, I would have appreciated a more explicit discussion of Moessner's perspective on variation.

Differences in meaning also play a role in Moessner's interpretations of some other results. The smaller share of subjunctives in relation to imperatives and modal constructions in ME than in OE main clauses is interpreted in terms of the hypothesis that the ME variants "which were preferably used, namely imperatives and modal constructions, especially those with the modal *shall*, expressed a stronger type of deontic modality than their OE predecessors with their greater share of subjunctives" (p. 46). Moessner also suggests that a preference for "more face-threatening" directive speech acts in ME main clauses "in turn offers a new explanation of the decreasing frequency of subjunctives between OE and ME" (p. 46). In addition to the discussion of whether variants are ways of saying the same thing if some of them express stronger modality than others (see above), this interpretation raises interesting questions of more general relevance. If ME does express stronger root modality than OE does, it would be of great interest to attempt to uncover potential reasons for such a difference. Alternatively, one might assume that other factors affect the distribution. For instance,

strength of the root modality expressed by the main variants considered by Moessner may have shifted between OE and ME; expressions that are not included by Moessner but also express root modality may have affected the variant field (for instance, as Moessner acknowledges (p. 235n), adverbs that express modality —very understandably— fall outside the scope of her study); or there may be problems regarding the comparability of the OE and ME samples. More extensive discussion of this issue would have been welcome.

The book is well written and well edited as a whole. Moessner's account is easy to follow despite a few typos and a number of run-on sentences, and the summaries of results are very reader friendly. In addition, as Moessner notes (p. 19), it is fully possible for readers to focus on only one clause type by accessing the relevant chapter directly.

In sum, Moessner's account is a valuable and very welcome contribution to research on the subjunctive in English (and, by extension, in other Germanic languages). The results presented in her book are also likely to be an important source of inspiration for further work on the topic, not least as regards (i) Late Modern English developments, and (ii) forms of the subjunctive not covered by Moessner, such as the past tense and additional forms of the verb *be* where subjunctive and indicative forms are distinct.

REFERENCES

Biber, Douglas, Jesse Egbert, Bethany Gray, Rahel Oppliger and Benedikt Szmrecsanyi. 2016. Variationist versus text-linguistic approaches to grammatical change in English: Nominal modifiers of head nouns. In Merja Kytö and Päivi Pahta eds. *The Cambridge Handbook of English Historical Linguistics*. Cambridge: Cambridge University Press, 351–375.

Huddleston, Rodney and Geoffrey K. Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Koplenig, Alexander. 2019. Against statistical significance testing in corpus linguistics. *Corpus Linguistics and Linguistic Theory* 15/2: 321–346.

Tagliamonte, Sali A. 2012. *Variationist Sociolinguistics: Change, Observation, Interpretation*. Chichester: John Wiley & Sons.

Reviewed by
Erik Smitterberg
Uppsala University
Department of English
P.O. Box 527
SE-751 20
Uppsala
Sweden
E-mail: erik.smitterberg@engelska.uu.se