Project Title	**:**	Investigating the Chinese and English language proficiency of tertiary students in Hong Kong: Perspectives from the Hong Kong subset of the multilingual student translation corpus
Grantee	**:**	Hong Kong Baptist University
Principal Investigator	**:**	PAN Jun Department of Translation, Interpreting and Intercultural Studies Hong Kong Baptist University
Co-investigators	**:**	CHAN Kar-yan School of Translation and Foreign Languages The Hang Seng University of Hong Kong WANG Honghua School of Translation and Foreign Languages The Hang Seng University of Hong Kong WONG Tak-ming Research Office The Open University of Hong Kong

Final Report

by

Principal Investigator

(a) Title

Investigating the Chinese and English Language Proficiency of Tertiary Students in Hong Kong: Perspectives from the Hong Kong Subset of the Multilingual Student Translation Corpus

(b) Abstract

This project aims to investigate the Chinese and English language proficiency of tertiary students in Hong Kong through the unique lenses of translation. An error-annotated translation learner corpus — the Hong Kong subset of the Multilingual Student Translation (MUST) corpus, was developed following the standard of an international multilingual corpus initiative for the study of translated language of language learners and translation students (Granger & Lefer, 2017). Consisting of Chinese–English and English–Chinese translations of students from over eleven tertiary institutions in Hong Kong, the corpus included over 300,000 word-tokens. It was annotated according to a standardized three-layer error annotation scheme of the MUST initiative, i.e. the Translation-oriented Annotation System (ibid). Apart from student translation data, the study gathered rich contextual information of the source texts, student translators, translation tasks, etc., via a standardized metadata questionnaire of MUST (ibid). Findings of the study suggest that distortion in content transfer was the highest frequency error type among students in both Chinese–English and English-Chinese translations. As far as written language errors were concerned, the top three problems in students' English-Chinese translation was heavy structure (style and situational context), multiword non-term collocation (lexis and terminology), and pronoun reference (cohesion), and the top three in Chinese-English were tense/aspect (grammar), spelling (mechanics), and punctuation (mechanics). These were the most urgent problems that need to be addressed in both secondary and tertiary

level of language teaching. Tailor-made exercises should therefore be developed to help enhance these identified deficiencies in students written Chinese and English language respectively.

(c) Keywords

Student translation
Learner corpus
Chinese language
English language
Language proficiency
Error annotation

(d) Introduction

The cultivation of bilingual (i.e., Chinese and English) personnel has long been a primary goal of education in Hong Kong. Despite the significance of bilingual proficiency enhancement, much remains unknown as to what aspects of tertiary students' bilingual proficiency should be enhanced.

Translation, as one of the prerequisites of bilingual competence, is often used to test and demonstrate the level of one's language proficiency. It has also been used in foreign language teaching as one of the earliest pedagogical tools. Therefore, students' translations consist of valuable data for the study of bilingual proficiency.

This project aims to investigate the Chinese and English language proficiency of tertiary students in Hong Kong through the unique lenses of translation. An error-annotated translation learner corpus was developed — the Hong Kong subset of the Multilingual Student Translation (MUST) corpus, following the standard of an international multilingual corpus initiative for the study of translated language of language learners and translation students worldwide (Granger & Lefer, 2017).

Tapping into the standardized error annotation scheme and rich contextual information of the source texts, student translators, translation tasks, etc. of the MUST initiative (ibid), the project build an error-annotated learner corpus of over 300,000 word tokens that helps to unveil the problematic aspects in the Chinese and English language proficiency of tertiary students in Hong Kong and pinpoint the most urgent problems for improvement. The project can also shed light on the design of language proficiency enhancement strategies catering for the needs of students at tertiary institutions.

(e) Review of literature of the project

1) Language education policy and bilingual proficiency of students in Hong Kong

Hong Kong nurtures a unique language environment, where both Chinese and English have been stipulated as its official languages since the United Kingdom (UK) transferred its sovereignty back to the People's Republic of China (PRC) in 1997[1]. The 1997 turnover had led to a lot of discussion and debate on Hong Kong's language education policy (Evans, 2013; Lin & Man, 2009), which were thereafter ratified as "biliterate and trilingual"[2]. The Education Bureau has spelled out the constitution and current interpretation of this language policy:

> The language education policy of the Government of the HKSAR aims to enable our students to become biliterate and trilingual. We expect that our secondary school graduates will be proficient in writing Chinese and English and able to communicate confidently in Cantonese, English and Putonghua.[3]

To achieve the goal of cultivating biliterate/bilingual[4] talents, many studies were carried out on the implementation and effectiveness of schools' policies of medium of instruction (MOI) (see Evans, 2013; Lin & Man, 2009). Most studies indicated the difficulties in employing English medium of instruction (EMI) in local secondary schools, largely attributable to the inadequacy of students' English language proficiency and lack of school/teacher support (Lin & Man, 2009). However, the Chinese medium of instruction (CMI) policy, prevalently employed after the 1997 turnover, was not well received, mostly due to its mismatch with the EMI policy adopted widely in local higher education institutions, and the lower prestige of the mother tongue in Hong Kong (Evans, 2013; Lin & Man, 2009). Lin (2015) therefore proposed the systematic use of L1 (Chinese) in bilingual classes focusing on content and language integrated learning (CLIL).

In addition to the general discussions and theoretical considerations, a number of empirical studies were conducted to investigate the relationship between MOI and bilingual or monolingual development of students at primary or mostly secondary schools. Tsang (2008), for instance, found that junior-form students' from CMI schools, although obtained higher scores in integrated content subject learning, had lower achievements in English language learning than junior-form EMI students. In addition, senior-form students, while stopping to benefit from the positive effect of CMI from their junior-form education on content subject

learning, continued to get lower scores in English language, and had smaller chances to enter tertiary level education (which often employs EMI) than their EMI peers (also see Evans, 2013; Lin & Man, 2009). Another study, Lo and Lo (2014), through a meta-study of 24 empirical studies, indicated that the application of EMI in secondary schools, while successfully contributing to higher levels of students' English language proficiency, also led to lower levels of Chinese language proficiency and insufficient command of the content knowledge.

Nevertheless, few studies provided insight on the specific aspects for the improvement of students' written Chinese and English proficiency (i.e., the "biliterate" goal). In this regard, Lin and Morrison (2010) tested the impact of MOI in secondary schools on tertiary students' English academic vocabulary, which is a key player of students' academic achievement at the tertiary level. Through comparing their results with Fan's (2001) study carried out at the beginning of the MOI policy change, Lin and Morrison identified a significant decrease in the size of students' English academic vocabulary, partially assignable to the increase of CMI secondary schools. These findings are useful not only for the review of current CMI policy for secondary schools, but also help to pinpoint the special aspects that are worthy of attention.

With the Government's call for "fine-tuning" its language policy (Education Bureau, 2010[5]), there is a need for a systematic investigation of the problematic aspects in current students' bilingual proficiency, upon and after the end of their secondary study. However, such an endeavour is yet to be addressed in the literature.

2) Translation and language education

The relationship between translation and language education is long-standing and far-reaching. Translation was, in the first place, employed as one of the earliest methods of foreign/second[6] language (FL/L2) teaching, i.e., the Grammar-Translation Method, at the outset in the teaching of classic languages of Greek and Latin back in the 16th century (Richards & Rodgers, 2001). With the development of translation studies as an academic discipline, scholars have been calling for differentiating between *translation in language teaching* and *language teaching for translators* — the former treats translation as a "significant component" in language teaching, and the latter focuses on "how translation might most effectively be provided with the kinds of linguistic skills which will help foreign language learners produce socio-functionally adequate texts in the most economic quality-oriented manner possible" (Malmkjaer, 1998, p. 1-2)[7]. In the meanwhile, the use of translation encountered "rejection" in L2 teaching since the Direct Method for language teaching, i.e., the use of only the target language in foreign language class rooms, was introduced towards the turn of the 20th century (Cook, 2010).

Although the application of translation in language teaching has gone through ebb and flow in history, there has been a revived interest in the indispensable relationship between the two, especially in higher education (Laviosa, 2014):

> Since the turn of the century, the debate about the merits of translation as a method of language learning, teaching and testing has been enriched by critical reflections on the value of educational translation as an aid to second language acquisition, as a means of developing metalinguistic competence, as a motivational factor, as an essential skill in today's multilingual societies and globalized world and as an ecological practice that not only recognizes the value and relevance of students' first language but also

> facilitates the creation of multilingual identities and protects linguistic as well as cultural diversity. (p. 28)

This revival of recognizing the role that translation plays in language teaching corresponds well with the recent revitalization of bilingual or mixed MOI (see Lin, 2015, as discussed in the previous section.

Whilst many studies naturally press for the use of translation as a means for L2 teaching or enhancement (see Laviosa, 2014), the application and benefits of translation in first language (L1) education have also been, although far less frequently, touched upon. Horner and Lu (2012) suggested that translation, as a translingual approach, can be employed in tertiary-level English writing classes in the United States (US), whereby both native and non-native English speakers in the classes can collaboratively improve their understanding of writing in a wider sense. They extended the notion of "teaching writing in English" to "rewriting English". Their translingual approach features the construction of multilingual identity and preservation of cultural diversity mentioned in Laviosa (2014).

In addition, Ngan (2009) addressed the relevance between bilingualism and translation, and proposed the incorporation of bilingual representation method in biliteracy training. The author defines bilingual representation as "a complicated process which involves selecting from the TL[8] corresponding counterparts of the SL with reference to the use of the SL in the context of the source text (ST)" (p. 41), a method often employed in practical translation. Moreover, Sidiropoulou (2015), focusing on modal markers, illustrated the usefulness of translation-related parallel data in foreign language teaching.

Apart from the pedagogical application of translation in language teaching, the notion of linguistic competence and translation competence are mutually inclusive. Linguistic competence is naturally included in the components of translation competence (PACTE 2003, p. 58) as "the underlying system of knowledge needed to translate".

In a different manner, language learning/teaching, in particular L2 teaching, implies a translational component. The notion of "competence" in language learning/teaching has been extended from Chomsky's (1965, p. 4) "linguistic competence" to "communicative competence" (Hymes, 1972). The former referred to "the speaker-hearer's knowledge of his language" (cf. the notion of "performance", i.e., "the actual use of language in concrete situations", Chomsky, 1965, p. 4). The latter was further divided into linguistic competence, sociolinguistic competence, discourse competence, and strategic competence (Canale & Swain, 1980). Communicative language teaching (CLT) was therefore built upon these competence components, within which, translation (also interpreting) was taken as "the fifth skill", in addition to the four basic skills of reading, writing, listening, and speaking in L2 teaching (Naimushin, 2002). Moreover, in Selinker's (1992) model of interlanguage (IL, i.e., the language produced by L2 learners) competence, translation skill is taken as an important indicator of L2 competence.

Moreover, translation has been used widely in language assessment (Tsagari & Floros, 2013). Ricardo-Osorio (2008), through a survey of FL learning outcomes assessment methods of undergraduate programmes in the US, showed that translation was the fourth most widely used assessment method, following faculty designed tests, student papers and projects, and student presentations. Likewise, Sun and Cheng (2013), through an empirical study, found that translation was a valid measure for students' FL competence.

* Notes:

1. See GovHK website: https://www.gov.hk/en/about/abouthk/facts.htm.

2. See GovHK website: http://www.policyaddress.gov.hk/pa99/english/espeech.pdf.

3. See Education Bureau's website: http://www.edb.gov.hk/en/edu-system/primary-secondary/applicable-to-primary-secondary/sbss/language-learning-support/featurearticle.html.

4. *Bilingual* in this study refers to *biliterate*, since this study concerns only the written Chinese and English. Similar uses can also been found in Lin & Man (2009).

5. See Education Bureau's website: http://www.edb.gov.hk/attachment/en/edu-system/primary-secondary/applicable-to-secondary/moi/2nd_moi_booklet.pdf.

6. The term *foreign language (FL)* and *second language (L2)* teaching are used interchangeably in the study.

7. This study focuses on *language learners* instead of *translation learners*. Therefore, it relates more to *translation for language learning* rather than *language teaching for translators*. However, these two are not entirely indispensable as suggested in the revived yet mutually-enriching relationship of the two disciplines of translation and language teaching (see Cook, 2010; Laviosa, 2014).

8. SL refers to the source language in translation, and TL the target language. Likewise, ST stands for source text and TT for target text.

9. See the ICLE website: http://www.fltr.ucl.ac.be/fltr/germ/etan/cecl/Cecl-Projects/Icle/icle.htm.

10. See the TeleNex website: http://www.telenex.hku.hk/telec/pmain/opening.htm.

11. See the Longman Learners' Corpus website: http://www.pearsonlongman.com/dictionaries/corpus/learners.html.

12. See the CLC website: http://www.cambridge.org/gb/cambridgeenglish/better-learning/deeper-insights/linguistics-pedagogy/cambridge-english-corpus.

13. See the CHILDES website: http://childes.talkbank.org/.

(f) Theoretical and/or conceptual framework of the project

1) Learner corpus and language learning

The study of learner language has long been an important aspect in language learning research, including both L1 and L2 acquisition as well as bilingual development (see Poulisse, 1999). At the core of learner language study is what Corder (1967/1983) refers to as "the systematic errors of the learner from which we are able to reconstruct his[/her] knowledge of the language to date, i.e., his[/her] *transitional competence*" (p. 168). This notion of language errors, due to its limitation of only addressing the static and negative side of learner language, has later been developed into the IL hypothesis, which addresses learner language from a developmental point of view, defined by Selinker (1972/1983) as "a separate linguistic system based on the observable output which results from a learner's attempted production of a TL [Target Language, in this case the second language the learner is attempting to learn] norm" (p. 176, elaborations added).

As discussed in the previous section, translation is an important indicator of IL competence (Selinker, 1992). Al Khafaji (2007) compared translation to translanguage, i.e., "a transitionally unstable linguistic entity that evolves during acts of translation along intersecting stages in a 'trip' stretching from the ST towards the TT during which hybrid 'language' comes

into being banking on the linguistic and social potentials of the SL and TL" (p. 473). Therefore, translation, in the sense of translanguage, is reflective of the (in)competence in a SL and TL.

The development of corpora has contributed greatly to the research and practice of language learning. Steward, Bernardini, and Aston (2004) defined three primary areas where language corpora and language learners are associated with each other. The primary one is *corpora by learners*, defined as the development of corpora that "can be used to study features of interlanguage" (p. 2). The second area is *corpora for learners*, referring to those "designed to benefit learning by allowing teachers and material designers to provide better descriptions of the language to be acquired" (p. 6). The last area concerns *corpora with learners*, relating to "activities designed to help leaners use corpora and to acquire linguistic knowledge and skills through their use" (p. 8).

The study of corpora by learners has great potential to the investigation of "the systematic errors of the learner" (Corder, 1967/1983, p. 168), or the systematic analyses of the IL, and can therefore shed light on the development of focused "teaching methods and contents … so as to speed acquisition" (Steward, Bernardini & Aston, 2004, p. 3). Granger (2002), in particular, gives a definition of "learner corpora" in FL/L2 learning that can be extended to L1 learner corpora as well:

> Computer learner corpora are electronic collections of language textual data assembled according to explicit design criteria for a particular language teaching purpose. They are encoded in a standardized and homogeneous way and documented as their origin and provenance. (Adapted from Granger, 2002, p. 7)

Nevertheless, the greatest challenge in learner corpus research lies in "identifying and classifying errors, and hypothesising 'correct' version corresponding to the learner's intentions" (Steward, Bernardini & Aston, 2004, p. 3). The lack of a consistent and comprehensive error classification system and the painstaking efforts involved in the annotation process may have led to the limited progress in learner corpus research (also see Granger, 1998).

Despite its difficulty, pioneering efforts have been made for learner corpus research. One of the most significant outcomes is the International Corpus of Learner English (ICLE; see Granger, 1998; Granger et al., 2009), the "best-known" learner corpus (McEnery, Xiao & Tono, 2006, p. 66) with 3.7 million English words in size and composing of essays written by advanced English learners from 16 different L1 backgrounds (Granger et al., 2009)[9]. ICLE features a standardized learner profile questionnaire and an error annotation scheme designed specifically for the language learners (ibid). The corpus has helped to greatly advance the study of learner language and the development of learner-corpora-informed L2 teaching resources (Granger, 2003; Steward, Bernardini & Aston, 2004), for example, the Teachers of English Education Nexus (TeleNex), as "a computer network providing continuous professional support to English language teachers in Hong Kong primary and secondary schools"[10]. The project website includes both data of student problems and teaching implications (see Granger, 2003).

Apart from the ICLE, other major international initiatives of learner corpus include the Longman Learners' Corpus, with 10 million English words written by English learners from 20 different L1 backgrounds, well-known for its use in dictionary and course book compilation that addresses "students' specific needs"[11], and the Cambridge Learner Corpus (CLC), a learner corpus collecting written English from 250,000 language learners all over the world, including

those produced by those taking the Cambridge ESOL English exams[12]. In addition, there are also a few L2 learner corpora collected only from Chinese speakers, including the Taiwanese Learner Corpus of English (Shih, 2000), the Chinese Learner English Corpus (CLEC; Gui & Yang, 2003), the Spoken & Written English Corpus of Chinese Learners (SWECCL; Wen, Wang & Liang, 2005), and the College Learners' Spoken English Corpus (COLSEC; Yang & Wei, 2005). The aforementioned L2 learner corpora usually feature learner language collected from post-secondary students.

In a different manner, L1 learner corpora mainly focus on the aspect of children language development (Behrens, 2008). Examples of major corpora in this regard include the Child Language Data Exchange System (CHILDES)[13], the Polytechnic of Wales (POW) Corpus[14], and the Lancaster Corpus of Children's Project Writing (LCPW)[15]. Among them, LCPW, which contains longitudinal data taken from 37 children aged between 9 and 11 in the UK, is the only one that focuses on written language. In addition, CHILDES also features a subcorpus of data provided by bilingual children.

To conclude, although learner corpus can provide valuable input to learner language features that can be used in the development of teaching resources (e.g., teaching materials, dictionaries, and online learning platforms) tailored to specific learner needs, its development is still limited largely by the difficulties in standardized data collection method and annotation schemes. Most of the existing learner corpora focused on L2 language learners, although a few of them were on children's L1. Whist the L1 and L2 learner corpora were not comparable due to inconsistent data annotation schemes and different student levels, the bilingual children data in CHILDES involve only spoken language (of 1 child in Hong Kong) and has limited scope of application

and implication. There seems to be no readily applicable corpus that can be used to study the learner language features in the biliterate setting in Hong Kong.

Thus the project will develop, based on standards developed for a new international initiative of multilingual student translation (MUST) corpus (Granger & Lefer, 2017), an error annotated learner corpus.

MUST is an international initiative, which aims to, based on the framework for developing ICLE, build a large multilingual student translation corpus with the collaborative efforts of researchers from different parts of the world (Granger & Lefer, 2017; 2020). The MUST corpus will include 25 languages and cover 50 language pairs[16]. Situated at the intersection of learner corpus research and corpus-based translation studies, the MUST corpus features the collection of rich contextual information about the learners' backgrounds and translation settings, as well as a shared annotation scheme of language errors for both learner language and translation research (ibid).

MUST provides a standardized corpus design, annotation scheme, metadata questionnaire, online interface for data input and annotation, to which the investigators of the proposed project has been granted access. These available resources have laid a solid groundwork for the proposed project. The corpus developed in this project constitutes of the Hong Kong subset of MUST, of which the Principal Investigator serves as a partner.

In addition, the project will tap into recent developments in corpora and translation education (Pan, 2019a, 2021a, 2021b, Pan & Laviosa, under review), as well as previous work of the investigators on Chinese/English language learning (Yan & Pan, 2016), the relationship

between learner variables and learner performance, including learning achievements and problems, in language learning and translation/interpreting training (Pan, 2012, 2014; Pan & Wang 2012; Pan & Yan, 2012, 2014; Yan, Pan & Wang 2010; Yan & Wang, 2012, 2015), corpus compilation (Chow & Wong, 2015; Pan & Wong, 2017), in particular translation/interpreting learner corpus design (Pan, 2012, Pan & Chan, 2013; Pan, 2017; Yan & Wang, 2014), the application of linguistic features for text quality assessment (Wong 2010), linguistic annotation of corpora data (Pan & Wong, 2015a, 2015b, 2017; Wong & Lee, 2013), computer tool development for semi-automatic annotation of linguistic features (Wong & Lee 2013; Wong et al. 2014; Chow & Wong 2015) and language and identity (Chan & Fong, 2016). The development and periodic reports on the MUST–HK corpus helped to testify the feasibility of a large-scale corpus for the study of language proficiency of tertiary students in Hong Kong (Pan & Wang, 2017, 2018; Pan, 2019b, Pan & Wong, 2021; Pan, Wong, Chan & Wang, 2021; Pan, Wong & Wang, 2021, under review a, under review b).

* Notes:

14. See the POW website: http://clu.uni.no/icame/manuals/POW.HTM.

15. See the LCPW website: http://www.lancaster.ac.uk/fass/projects/lever/.

16. More information can be obtained from the MUST website: https://uclouvain.be/en/research-institutes/ilc/cecl/must-partners.html.

(g) Methodology

This project aims to investigate the Chinese and English language proficiency of tertiary students in Hong Kong through the unique lenses of translation. Corpus compilation and annotation constituted of two major steps in the project. Employing instruments of

contextual/learner data collection and error annotation scheme developed for the MUST international initiative (Granger & Lefer, 2017; 2020), this project analysed the carefully collected Hong Kong subset of the MUST corpus using both quantitative and qualitative methods.

In particular, the project pivots on two main research questions:

(1) What are the high-frequency error types in written Chinese/English of tertiary students in Hong Kong?
(2) What are the relationships between the types of Chinese/English language features and relevant contextual/learner factors?

(h) Data collection and analysis

1) Corpus Compilation

The learner corpus, i.e., the Hong Kong subset of the MUST corpus, consists of translations and metadata provided by students from over eleven tertiary institutions in Hong Kong. Six main batches of data collection were performed during the project period (Sep 2018- Aug 2021).

Figure 1 displays the self-reported data of students participating in the latest batch of data collection.

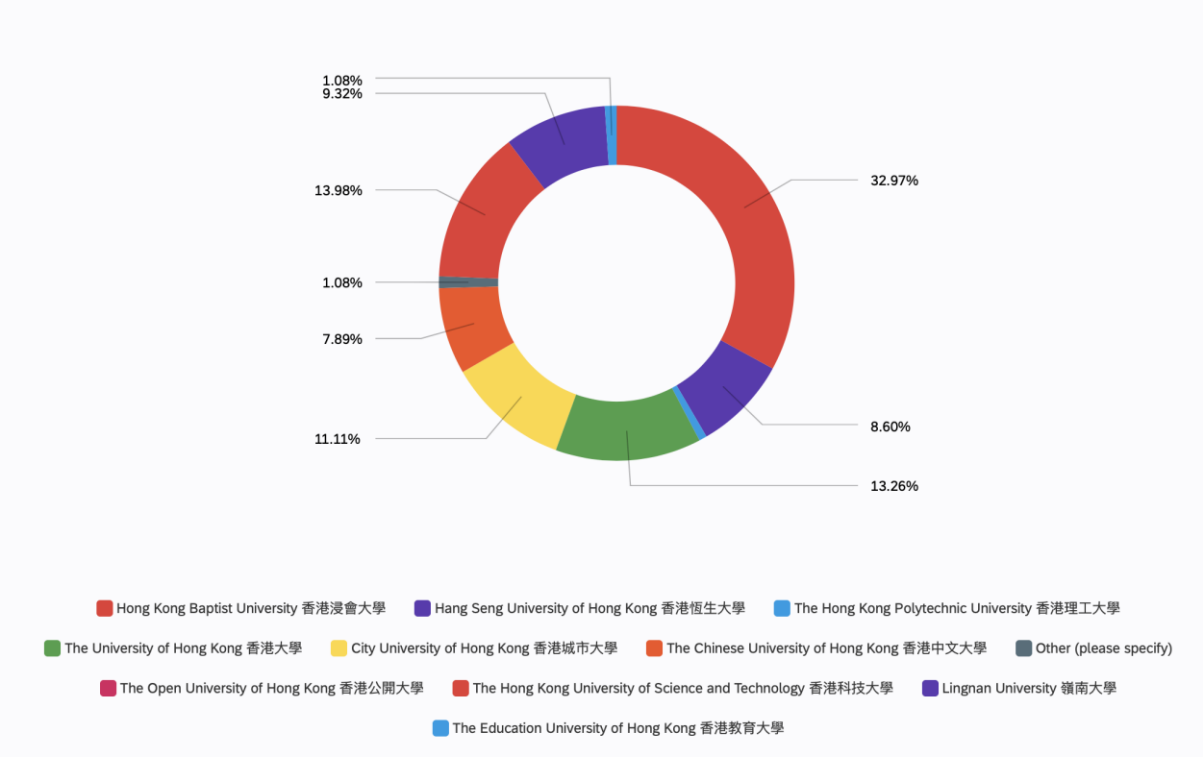

Figure 1. Student participants of the project (latest batch)

The participating students included undergraduate Chinese and English language learners participating in language courses at language centers, and Chinese and English language departments (including but not limited to translation students). Language centers usually aim at helping students with their effective writing in both the Chinese and English languages[17]. They usually provide institution-wide credit-bearing as well as non-credit-bearing courses to all students from different departments of the institution[18]. In some institutions, Chinese/English language departments also undertake the work of language centers in providing institution-wide language courses[19]. In this case, their institution-wide language courses will provide data for the project. Student participants taking these courses will be the primary group, coded as Chinese/English language general learners (CLGLs/ELGLs).

Chinese and English language departments (including translation programmes/departments) usually have specific criteria for recruitment of students on their Chinese and English grades in the Hong Kong Diploma of Secondary Education (HKDSE). Therefore, student participants from these departments will be the secondary group, coded as Chinese/English language major learners (CLMLs/ELMLs).

Figure 2 shows the percentage of the student majors.

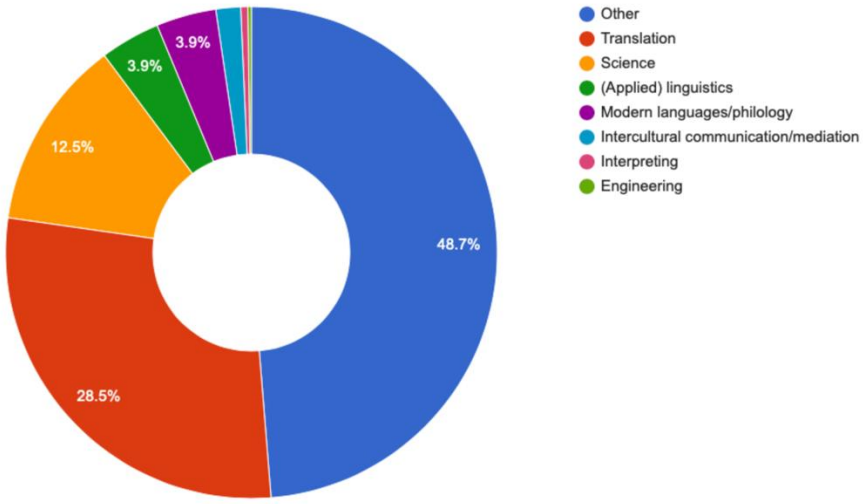

Figure 2. Current study background of the participating students

The participating students were invited to provide translations from both Chinese to English and English to Chinese. The texts for translation were on general topics, including selected excerpts taken from newspaper or magazine articles (Table 1). Each of the text was about 250-600 words in length, in line with the MUST specifications (Granger & Lefer, 2017). The investigators of the project assessed the text in terms of translation difficulty level based on their expertise of translation and basic criteria such as type/token ratio and the vocabulary range. The array of texts has the potential to enlist a wide range of language errors.

Table 1. The list of source texts used for the corpus.

Tears, fears & cheers: How did your workplace handle the post-election fallout?
Moonlight's Barry Jenkins on Oscar fiasco: 'It's messy, but kind of gorgeous'
Trump Delays a Tariff Deadline, Citing Progress in China Trade Talks
Green Book' Review: A Road Trip Through a Land of Racial Clichés
賈寶玉的大紅斗篷與林黛玉的染淚手帕《紅樓夢》後四十回的悲劇力量
Lowering bar for disadvantaged students has failed to redress imbalance in university admissions, regulator says
A company's meeting on its volunteering projects
好好過日子——時間沒有溜走
內地「碼農」的覺醒——抗議「996」還我加班費
The Guardian view on extinction: time to rebel
好好過日子——藥不能停
Overcome Procrastination
心寬，路更寬

Contextual information of the texts were coded by the principal investigator. Figure 3 shows the codes used for a sample source text.

```
source_text_author_speaker_status="Native speaker of the SL"
source_text_publication_status="Published" source_text_title_full_text="Tears, fears & cheers:
How did your workplace handle the post-election fallout?" source_text_author_name="Pat
DiDomenico " source_text_sampling="Sample" source_text_communicative_purpose="Mixed
communicative purpose" source_text_keyword2="workplace"
source_text_publication_format_subtype="Web"
source_text_reference_translation_f="file_I5IEYMM2SLBB.pdf"
source_text_reference_translation_publication="Unpublished at the time of data collection"
source_text_whole_text_f="file_B21525IKW0XD.pdf"
source_text_reference_translation_directionality="L2 to L1" source_text_author_type="Sole
author" source_text_language_type="General language" source_text_title_segment="Tears, fears &
cheers: How did your workplace handle the post-election fallout?"
source_text_target_audience="Specialized external audience"
source_text_general_language_type="Journalistic texts"
source_text_opinion_article_subgenre="Commentary/personal column"
source_text_whole_text_available="Yes" source_text_reference_translator_status="Member of
teaching/research staff" source_text_st_translation_status="Original text"
source_text_keyword3="emotions" source_text_publication_date="2016-11-18 23:59"
source_text_keyword1="presidential election" source_text_reference_translation="Available"
source_text_language="en" source_text_publication_url="https://www.businessmanagementdaily.com/
47609/tears-fears-cheers-how-did-your-workplace-handle-the-post-election-fallout"
source_text_publication_format="Electronic document/digital copy"
source_text_publication_reference="Not applicable" source_text_sample_position="Beginning
sample" source_text_mode="Written to be read" source_text_length_words="252"
source_text_journalistic_texts_genre="Opinion article"
source_text_reference_translation_target_language="zh_Hant_HK" >
```

Figure 3. Metadata coded for a sample source text.

Student translations were collected through the tailor-made Hybrid Parallel Text Aligner for the MUST corpus, i.e., Hypal4MUST (Granger & Lefer, 2017; Figure 4). Each translation took about 40 – 60 minutes.

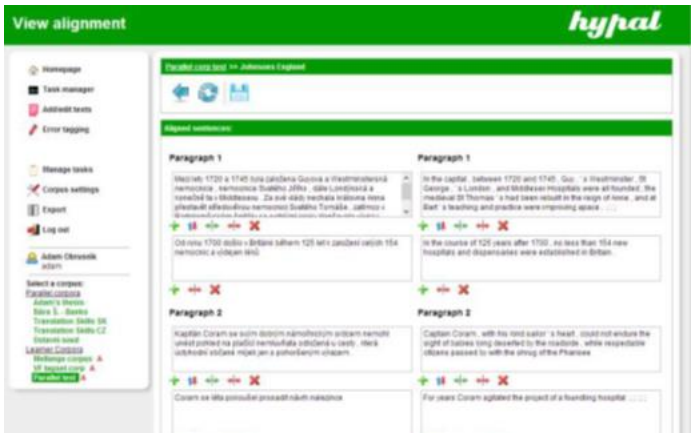

Figure 4. The Hypal Interface (Obrusnik, 2014, p. 68)

Apart from the translation data, metadata information of the corpus data (Figure 5), including student and task specific data (Granger & Lefer, 2017; Pan & Wang, 2017) were also collected

through a survey and uploaded to the same platform. Students took about 15 minutes on average to complete the information.


```
student_mother_tongue="zh_Hant_HK" task_student_tl_status="Mother tongue"
student_experience_with_cat="No" student_gender="Male" student_main_language_university_1="en"
student_main_language_university_2="na" student_main_language_university_3="na"
student_foreign_language_2="zh_Hans" student_foreign_language_3="es"
student_main_language_secondary_school_1="en" student_foreign_language_1="en"
student_sl_proficiency="Intermediate" task_student_duration="60" student_age="20"
student_translation_internships="No" student_semesters_studying_translation="2"
task_student_feedback_type_before_submission="Individual teacher's feedback"
task_student_tools_resources="Bilingual dictionaries"
task_student_feedback_before_submission="Yes" student_current_study_background="Translation"
student_tl_proficiency="Native" student_mother_tongues="1 mother tongue"
student_sl_countries_stays="No" student_prior_study_background="Other"
student_main_language_secondary_school_2="zh_Hant" student_sl_years_studying="17"
task_student_sl_status="Foreign language" student_main_language_primary_school_1="en"
student_main_language_primary_school_2="zh_Hant_HK" translation_brief_file_available="Yes"
marking="Marked" student_status="Trainee translator" translation_brief_additional_info="Not
applicable" level="Undergraduate" type_of_task="Examination" translation_brief_used="Yes"
translation_brief_audience_specified="Not specified" translation_brief_layout_and_guide="No
instructions given" translation_brief_requester_specified="Not specified" duration_minutes="60"
translation_brief_use_glossaries="No instructions given"
translation_brief_f="translation_brief_fileN75RTTMS1YMV.pdf" translation_brief_format="Hard copy"
translation_brief_format_specified="Specified" duration="Timed"
translation_brief_communicative_purpose_specified="Not specified" revision="Only one version of
the TT is expected" translation_brief_use_memories="No instructions given"
tools_and_resources="Allowed" target_language="zh_Hant_HK"
```


Figure 5. Metadata of a sample student and translation task information

The data collected were then processed for data cleansing and parallel text alignment to pair up the Chinese–English bilingual texts at the sentence level on the Hypal4MUST platform.

2) Corpus Annotation

Apart from POS tagging, the corpus was annotated with errors made in the student translations according to a standardized three-layer error annotation scheme of the MUST initiative, i.e. the Translation-oriented Annotation System (TAS 1.0), which was based on a broad range of leading error schemes in both language and translation studies worldwide (Granger & Lefer, 2017; 2020).

The current version of MUST annotation scheme comprises several major frameworks. The CELTraC error typology (Fictumová, Obrusnik, & Štěpánková, forthcoming), specifically

developed for the annotation of translation learner corpus was incorporated as one of the major frameworks. The typology took into consideration transfer and language errors on a two-layer system and was already incorporated to the Hypal interface (Granger & Lefer, 2017; Figure 6) used for the annotation in this project. Its primary annotation categories include content transfer, grammar, terminology and lexis, hygiene, and register and style (ibid).

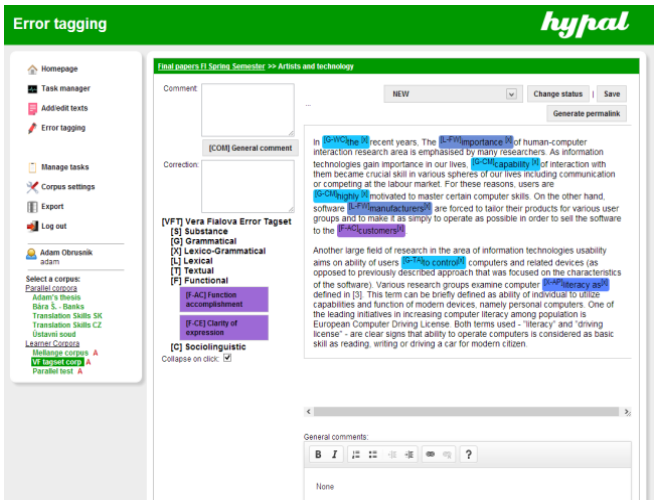

Figure 6. The Hypal Error tagging interface (Obrusnik, 2014, p. 68)

Another major framework was the Université catholique de Louvain Error Editor (UCLEE), a three-layer scheme used for the annotation of errors in FL student writing. The major error categories include form, grammar, lexis, punctuation, sentence, word, lexico-grammar, and infelicities (Granger & Lefer, 2017).

In addition, the annotation scheme also took into consideration partner discussions at the series MUST workshops (2016 – now). In the end, TAS 1.0 included the following categories:

Table 2. Annotation scheme (TAS 1.0, Granger & Lefer, 2017; 2020)

LAYER 1	LAYER 2	LAYER 3

ST-TT TRANSFER (TR)	Content Transfer (CT)	Omission (OMI) Addition (ADD) Distortion (DIS) Indecision (IND)
	Lexis (LE)	Translating untranslatable (TUN) Untranslated translatable (UNT) Term translated by non-term (TNT) Non-term translated by term (NTT)
	Discourse/Pragmatics (DP)	Connectors (CON) Theme-rheme (THR)
	Register and Culture (RC)	Register mismatch (REG) Cultural mismatch (CUL)
	Translation Brief (TB)	Inconsistency with glossary (GLO) Formatting (FOR)
LANGAUGE (LA)	Grammar (GR)	Inflectional morphology (INF) Tense/aspect (TNS) Voice (VOI) Word order (WOR) Determiner (DET) Pronoun (PRO) Preposition (PRE) Concord (CCD) Complementation (COM) Adjective (ADJ) Noun (NOUN)

		Verb (VRB)
		Adverb (ADV)
	Lexis and terminology (LT)	Single word non-term (SWN)
		Derivative (DER)
		Cognate (COG)
		Single word term (SWT)
		Multiword non-term (MWN)
		Compound (COP)
		Collocation (COL)
		Idiom (IDI)
		Multiword term (MWT)
	Cohesion (CO)	Pronoun reference (PRF)
		Linkword (LIN)
	Mechanics (ME)	Punctuation (PUN)
		Units, dates, numbers (UDN)
	Style and situational context (ST)	Heavy (HEA)
		Redundant (RED)
		Contextual variant (COV)
		Degree of (in)formality (FML)

The project team first piloted the annotation scheme on a small sample of the Hong Kong corpus to validate their suitability for the data collected. The Principal Investigator then trained the project research assistant(s) on the annotation scheme. Sample annotations and discussions were made within the project team to make sure all annotators understood the annotation scheme correctly and consistently. Then annotation was performed. When different annotators

worked on the same task, comparison and training were performed to help reach an initial inter-annotator reliability of up to 98%. Adjustments to the annotation scheme and annotation logs were recorded along the process (Pan & Wong, 2021). The annotation was performed on the Hypal4MUST platform.

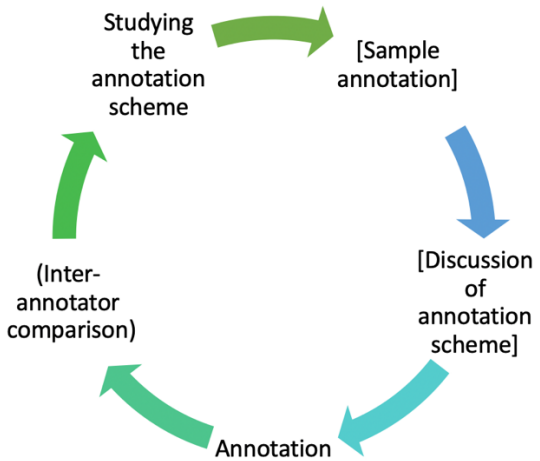

Figure 7. The annotation process employed for the project

3) Corpus analysis

The corpus data were then analysed through the Hypal4MUST platform and corpus analysis software Sketch Engine (Kilgarriff et al., 2014). The high frequency errors in the Chinese and English subsets were calculated respectively, and selected learner/contextual factors were chosen as parameters for cross-comparison among different subsets of the corpus.

Notes:

17. See, for example, the HKBU language centre website: http://lc.hkbu.edu.hk/mission.php.

18. See, for example, the HKBU language centre website: http://lc.hkbu.edu.hk/course_credit.php.

19. See, for example, the HSMC Chinese department website: http://www.hsmc.edu.hk/index.php/component/departments/component/departments/?dep=hum&sid=24.

(i) Results and Discussion

The following parts report the main findings of the project. Further details of the project reports can be found in the project outputs/publications (Appendix 1).

1) Corpus statistics

Based on the calculation performed by Sketch Engine, the corpus consists of over 300,000 word tokens, with 195,448 in the Chinese subset and 131,295 in the English language subset, each with a type/token ratio of 2.5-3.5% (Table 3).

Table 3. Corpus statistics

	TOKENS	TYPES	TYPE/TOKEN RATIO
CHINESE	195,448	5,122	2.62%
ENGLISH	131,295	4,239	3.23%
TOTAL	326,743	9,361	2.86%

2) Most frequent error tags in the Chinese sub-corpus

Figure 8 shows the most frequent error tags in the Chinese sub-corpus. Distortion was, apparently, the highest frequency error type, which was mostly triggered by misunderstanding of the source language, as well as inaccurate target language expression. Nouns and verbs were the mostly common part-of-speeches where distortion occurred (Figure 9). When the top level language errors were taken into consideration, heavy structure (style and situational context),

multiword non-term collocation (lexis and terminology), and pronoun reference (cohesion) were the top three problems among the students. These problems become the urgent issues that both secondary and tertiary level of language teaching should focus on.

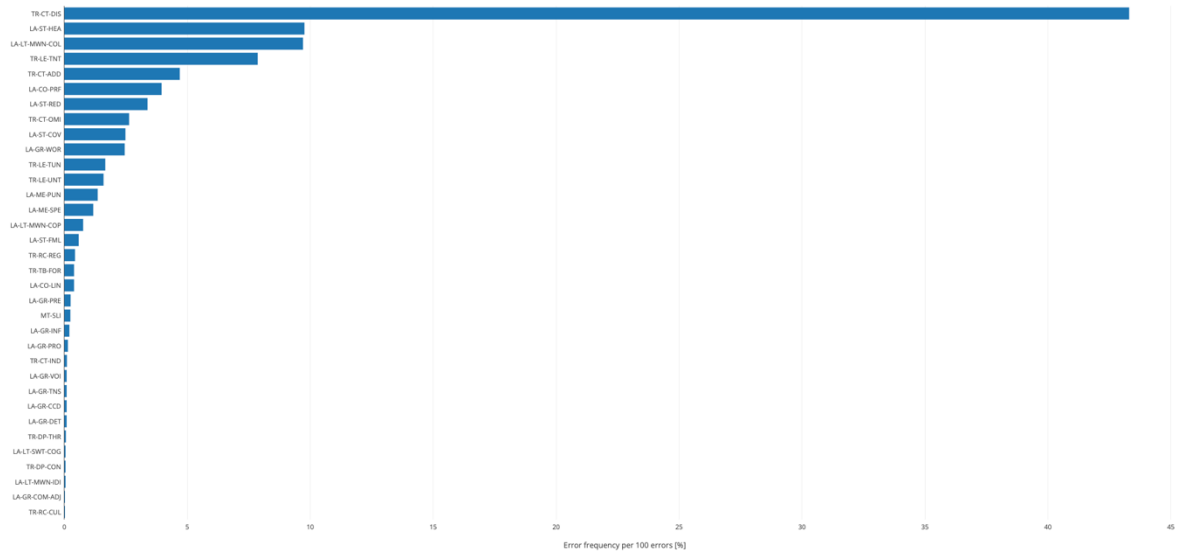

Figure 8. Most frequent error tags in the Chinese sub-corpus

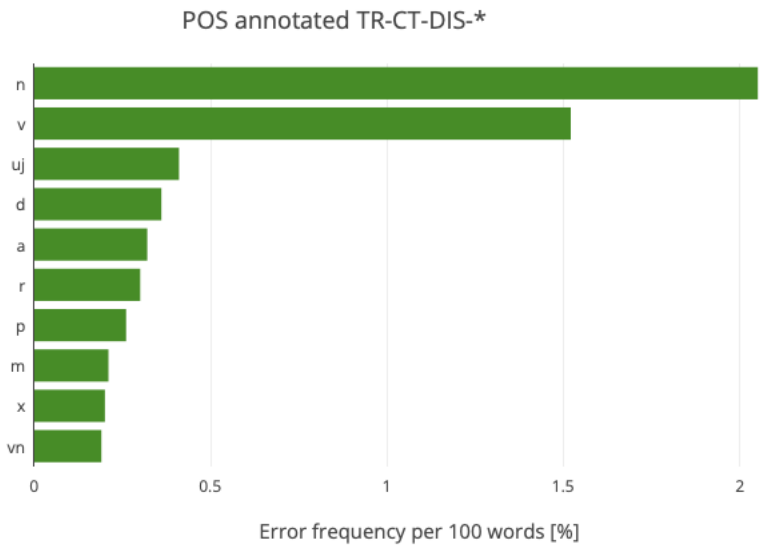

Figure 9. POS annotated distortion in the Chinese sub-corpus

3) Most frequent error tags in the English sub-corpus

Figure 10 shows the most frequent error tags in the English sub-corpus. Likewise, distortion was, apparently, the highest frequency error type, which was mostly triggered by inaccurate target language expression. Nouns and prepositions were the mostly common part-of-speeches where distortion occurred (Figure 11). The top level language errors were tense/aspect (grammar), spelling (mechanics), and punctuation (mechanics). These problems were the most urgent ones that need to be addressed in both secondary and tertiary level of language teaching.

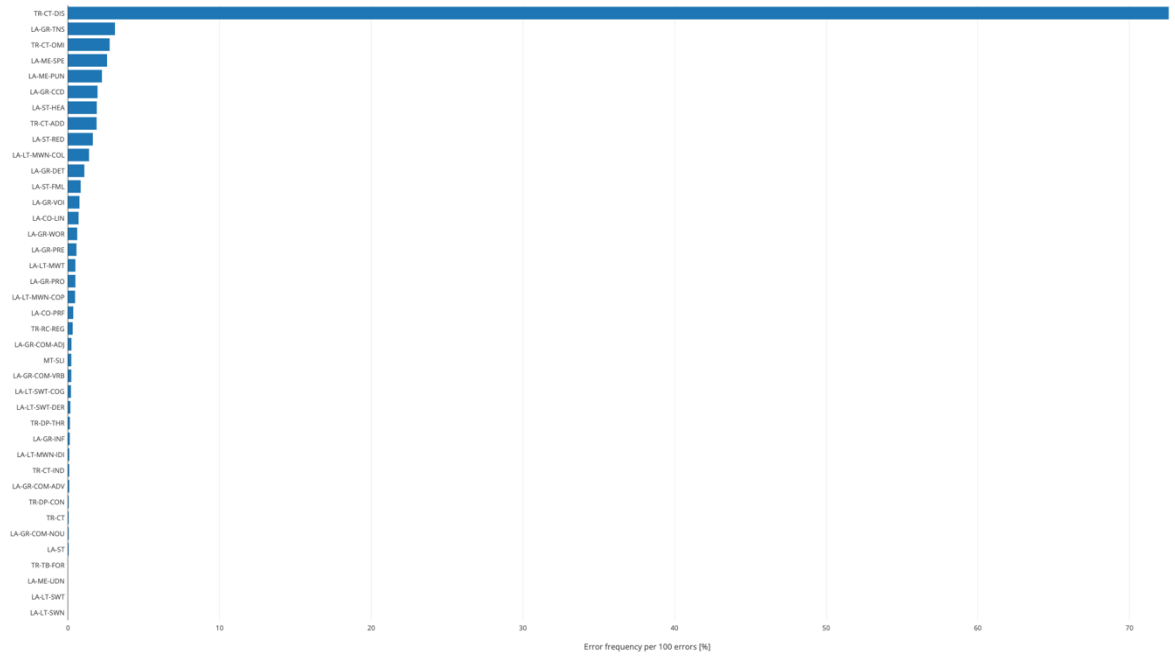

Figure 10. Most frequent error tags in the English sub-corpus

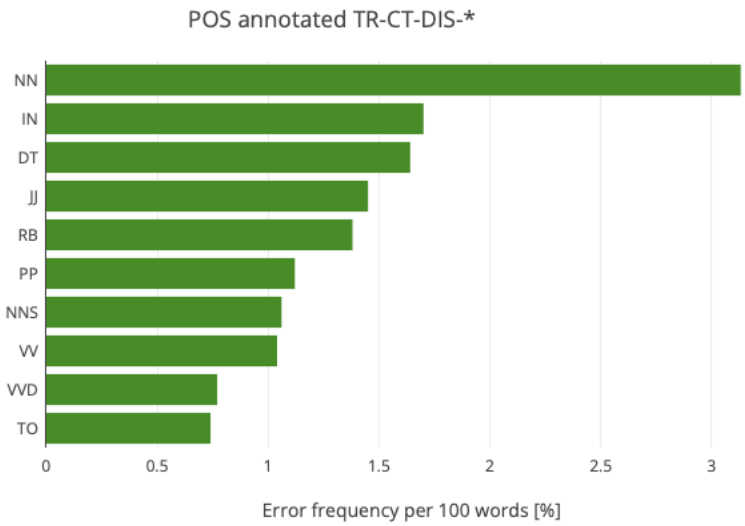

Figure 11. POS annotated distortion in the English sub-corpus

4) Gender and students' Chinese/English language features

The typical 3- and 4-grams of the students' outputs were computed using Sketch Engine. According to Figure 12 and 13, the results indicate that male and female students have slightly different preferences of word cluster use in both the Chinese and English outputs: female students tend to employ more regular phrases than male students do.

	Word	Frequency ?
1	的 關 係	122
2	生物 多 樣性	107
3	生物 多樣 性	105
4	野生 動物 的	103
5	而 不 是	91
6	綠色 新 政	91
7	多 樣性 的	88
8	了 一 個	88
9	生物 多 樣性 的	88
10	一 個 世紀	88
11	在 辦 公室	87
12	我們 的 情緒	85

	Word	Frequency ?
1	的 關 係	47
2	關 係 較	26
3	生物 多 樣性	25
4	封 電子 郵件	25
5	在 辦 公室	24
6	的 關 係 較	24
7	總統 選 舉	22
8	的 總統 選	22
9	最 令 人	22
10	的 總統 選 舉	22
11	我們 的 情緒	22
12	野生 動物 的	22

Figure 12. Top 3- and 4-grams of the student outputs (female vs. male) in the Chinese corpus

	Word	Frequency ?	
1	Dream of Red	165	•••
2	of Red Mansions	161	•••
3	Dream of Red Mansions	153	•••
4	of the Red	135	•••
5	Dream of the	133	•••
6	the Red Chamber	133	•••
7	Dream of the Red	133	•••
8	of the Red Chamber	132	•••
9	told me to	130	•••
10	a blood test	105	•••
11	for two years	104	•••
12	for three months	103	•••

	Word	Frequency ?
1	Dream of Red	54
2	of Red Mansions	53
3	Dream of Red Mansions	53
4	Dream of the	37
5	Dream of the Red	37
6	of the Red	37
7	the Red Chamber	31
8	of the Red Chamber	31
9	told me to	30
10	stone is the	25
11	A Dream of	25
12	A Dream of Red	24

Figure 13. Top 3- and 4-grams of the student outputs (female vs. male) in the English corpus

5) MOI and students Chinese/English language features

Likewise, the typical 3- and 4-grams of the students' outputs were compared between CMI and EMI students: EMI students seem to employ slightly more regular phrases than CMI students do in general.

	Word	Frequency ?
1	的 關 係	85
2	生物 多 樣性	67
3	而 不 是	63
4	生物 多樣 性	62
5	野生 動物 的	61
6	多 樣性 的	56
7	生物 多 樣性 的	56
8	封 電子 郵件	52
9	一 個 世紀	52
10	用 了 一	51
11	和 食物 鏈	51
12	我們 的 情緒	50

	Word	Frequency ?
1	的 關 係	84
2	生物 多樣 性	67
3	野生 動物 的	65
4	生物 多 樣性	65
5	在 辦 公室	62
6	總統 選 舉	60
7	了 一 個	59
8	一 個 世紀	58
9	的 總統 選	58
10	的 總統 選 舉	58
11	我們 的 情緒	57
12	綠色 新 政	56

Figure 14. Top 3- and 4-grams of the student outputs (CMI vs. EMI in secondary school) in the Chinese corpus

	Word	Frequency ?
1	Dream of Red	124
2	of Red Mansions	120
3	Dream of Red Mansions	116
4	of the Red	93
5	Dream of the Red	91
6	Dream of the	91
7	the Red Chamber	87
8	of the Red Chamber	86
9	told me to	75
10	for three months	69
11	for two years	67
12	a blood test	65

	Word	Frequency ?
1	Dream of Red	95
2	of Red Mansions	94
3	Dream of Red Mansions	90
4	told me to	86
5	of the Red	83
6	Dream of the	83
7	Dream of the Red	83
8	of the Red Chamber	81
9	the Red Chamber	81
10	for two years	62
11	a blood test	60
12	my cholesterol level	58

Figure 15. Top 3- and 4-grams of the student outputs (CMI vs. EMI in secondary school) in the English corpus

6) Previous study background and Chinese/English language features

When students' previous study background (translation vs. non-translation) is taken into consideration, translations students seem to employ slightly more regular phrases than non-translation students do in general.

	Word	Frequency ?		Word	Frequency ?
1	最 佳 電影	12	1	的 關 係	165
2	的 製 作 人	11	2	生物 多 樣性	124
3	的 製 作	11	3	生物 多樣 性	121
4	製 作 人	11	4	野生 動物 的	117
5	美國 總 統	10	5	在 辦 公室	106
6	最 年輕 的	10	6	封 電子 郵件	105
7	的 最 佳	9	7	總統 選 舉	104
8	總 統 特朗普	9	8	而 不 是	103
9	野生 動物 的	9	9	我們 的 情緒	103
10	和 食物 鏈	8	10	一 個 世紀	103
11	海洋 和 食物	8	11	多 樣性 的	102
12	學生 辦 公室	8	12	的 總統 選 舉	101

Figure 16. Top 3- and 4-grams of the student outputs (Translation vs. Non-translation) in the Chinese corpus

	Word	Frequency ?
1	told me to	11
2	what to do	9
3	for two years	9
4	the time has	9
5	If you can	8
6	me to take	8
7	my cholesterol level	8
8	the blood test	8
9	one by one	8
10	for three months	8
11	I went to	7
12	He told me to	7

	Word	Frequency ?
1	Dream of Red	217
2	of Red Mansions	212
3	Dream of Red Mansions	204
4	of the Red	173
5	Dream of the Red	171
6	Dream of the	171
7	the Red Chamber	165
8	of the Red Chamber	164
9	told me to	150
10	for two years	120
11	a blood test	118
12	for three months	115

Figure 17. Top 3- and 4-grams of the student outputs (Translation vs. Non-translation) in the English corpus

7) Language proficiency and Chinese/English language feature

Last but not least, students self-perceived target language proficiency (Native, Advanced vs. Intermediate) seems to lead to different preferred word clusters as well: those of "native" target language performed slightly better than "advanced" and "intermediate" in producing more regular phrases.

	Word	Frequency ?
1	的 關 係	167
2	生物 多 樣性	128
3	生物 多樣 性	123
4	野生 動物 的	122
5	在 辦 公室	111
6	而 不 是	107
7	我們 的 情緒	107
8	封 電子 郵件	107
9	一 個 世紀	106
10	多 樣性 的	106
11	總統 選 舉	105
12	生物 多 樣性 的	105

	Word	Frequency ?
1	額 外 的	17
2	貧 窮 學生	14
3	學生 辦 公室	12
4	外 的 壓力	11
5	學生 填 寫 額	11
6	寫 額 外	11
7	學生 填 寫	11
8	小 時 的	11
9	填 寫 額	11
10	填 寫 額 外	11
11	的 學生 填	11
12	大學 的 入學	11

	Word	Frequency ?
1	額 外 的	28
2	學生 辦 公室	23
3	外 的 表格	17
4	小 時 的	17
5	額 外 的 表格	15
6	貧 困 學生	14
7	過 於 保守	14
8	寫 額 外	14
9	填 寫 額	14
10	填 寫 額 外	14
11	聖 約翰 大學	14
12	大學 的 入學	13

Figure 18. Top 3- and 4-grams of the student outputs (Native, Advanced vs. Intermediate) in the Chinese corpus

	Word	Frequency ?
1	Dream of Red	11
2	the past ten	8
3	of Red Mansions	7
4	Dream of Red Mansions	7
5	the past ten years	7
6	past ten years	7
7	but the time	6
8	in the past	6
9	time has passed	6
10	in the past ten	6
11	spiritual stone is	5
12	one by one	5

	Word	Frequency ?
1	of the Red	117
2	Dream of the Red	117
3	Dream of the	117
4	of the Red Chamber	114
5	the Red Chamber	114
6	Dream of Red	95
7	told me to	94
8	of Red Mansions	94
9	Dream of Red Mansions	86
10	for two years	68
11	a blood test	67
12	for three months	62

	Word	Frequency ?
1	of Red Mansions	113
2	Dream of Red	113
3	Dream of Red Mansions	113
4	A Dream of	66
5	A Dream of Red	65
6	told me to	63
7	for three months	60
8	of the Red	59
9	for two years	58
10	Dream of the	57
11	Dream of the Red	57
12	the Red Chamber	54

Figure 19. Top 3- and 4-grams of the student outputs (Native, Advanced vs. Intermediate) in the English corpus

(j) Conclusions and Recommendations

This project aims to investigate the Chinese and English language proficiency of tertiary students in Hong Kong through the unique lenses of translation. It identified the high-frequency error types in written Chinese/English of tertiary students in translation Hong Kong, and the relationship between Chinese/English language features and relevant contextual/learner factors.

The main outputs of the study included (also refer to Appendix 1):

- An over 300,000-word learner corpus with Chinese–English and English–Chinese translations contributed by tertiary-level bilingual students in Hong Kong, annotated with internationally standardized language error types;
- Batteries of Chinese and English language errors of tertiary bilingual tertiary learners in Hong Kong (see section (h) and (i));
- An online symposium (cum project workshop) promoting the project findings and with focused discussion on the latest development of corpora and translation education by renowned scholars from Belgium, Czech, Italy, UK, US, Spain, and Macau, Hong Kong and the Chinese Mainland;
- An edited volume on the latest development of copra and translation in relevance to the project (Pan & Laviosa, under review);
- A book chapter and journal article (Pan, Wong & Wang, under review a, under review b) on the project findings, and at least one more article is under preparation;
- A total of eight conference papers/talks; and
- An online platform that showcases students' errors in the translational Chinese/English written language and pedagogical solutions to these student errors (https://ctn/hkbu.edu.hk/hktilc/hkmust, available soon).

The over 300,000-word error-annotated translation learner corpus developed in the project can provide rich research and teaching resources. Granger (1998) puts forward an significant factor of consideration in compiling a learner corpus:

> One factor which has a direct influence on the size of learner corpora is the degree of control exerted on the variables … and this in turn depends on the analyst's objectives … If the researcher is an SLA [Second Language Acquisition] specialist who wants to assess the part played by individual learner variables such as age, sex or task type, or if he[/she] wants to be in a position to carry out both cross-sectional and longitudinal studies, then he[/she] should give priority to the quality rather than the quantity of the data. (p.11)

Since the project corpus included learner, task and source text metadata of more than 30 types, and was annotated a three-layer annotation scheme of over 40 error types, it is considered large in scale for the current study. As a matter of fact, the corpus is considered so far the largest annotated subset in the MUST international corpus. Also, with data collected from over eleven institutions across Hong Kong, the corpus can be regarded representative of current language proficiencies of the target student population.

Based on results obtained from this large-size annotated corpus, the project identified distortion as the highest frequency translational error of both Chinese-English and English-Chinese students in Hong Kong, which is similar to the results obtained by Izquierdo et al. (2021) on the English to Spanish subset, who employed the same MUST TAS 1.0 annotation scheme on the translation of multiword expressions (22,184 words annotated). At the present stage, there are only a couple of annotated corpora among MUST partners (with the MUST HK subset being so far the largest annotated subset). The next step will be to compare results with

annotated MUST subsets developed by partners of other language combinations (if possible of similar size) in the future. At the latest MUST Workshop, similar interests have been identified and opportunities for performing comparative studies will be explored.

As far as written language errors were concerned, the top three problems in students' English-Chinese translation was heavy structure (style and situational context), multiword non-term collocation (lexis and terminology), and pronoun reference (cohesion), and the top three in Chinese-English were tense/aspect (grammar), spelling (mechanics), and punctuation (mechanics). These were identified as the most urgent problems that need to be addressed in both secondary and tertiary level of language teaching.

The error batteries of both Chinese and English problems of students at tertiary institutions in Hong Kong can be further divided into translational and language ones. The study supports the idea that translation can serve as a unique lens to study students' language proficiency than written tasks, as the results can also shed light on the source language comprehension and influence.

The error batteries can be further employed in teaching and assessment. Students can be trained on the different error types with examples from the annotated learner corpus developed in the project. Language and translation teachers can make use of the annotation scheme for assessment purposes as well, so that students can obtain information based on the frequency of annotations made on their translations.

Also, the project-related online symposium helped to gather worldwide translation and learner corpora developers and researchers, including Sylviane GRANGER, one of the MUST

international corpus initiator and TAS 1.0 developer used in the project, Sara LAVIOSA, the author of the most cited book *Corpus-base Translation Studies*, Adam OBRUSNIK, developer of the Hypal interface used in the project, Mark DAVIS, developer of English-Corpora.org the project referenced on, and many other very relevant names in the field of corpora and translation education (please refer to Appendix 2 and 3 for the profile of the invited speakers and their abstracts). The two-day symposium, consisting of 15 presentations and 5 roundtable discussions, attracted over 400 participants (scholars, students, practitioners and public audience), and over 100 participants at each single session. The discussion at the symposium and its follow-up publication will certainly bring further theoretical and applicable input to the project and its sustainable development.

In addition, the project indicates that learner factors such as gender, MOI at secondary schools, previous study background and self-perceived language proficiency may lead to different language features produced by students in both the Chinese and English translational outputs. Based on these findings, we recommend that tailor-made exercises should be developed to help enhance identified deficiencies in students written Chinese and English language respectively. A project website with online platform that showcases students' errors in the translational Chinese/English written language and pedagogical solutions to these student errors is thus developed to tackle such needs (https://ctn/hkbu.edu.hk/hktilc/hkmust, available soon).

To conclude, the project, with its rich annotated data and student/context information collected, can provide valuable insight into the language proficiency, and most importantly, deficiencies of students in Hong Kong. Based on the initial findings presented here, more in-depth analyses will be carried out to find out the specific differences among learners of different language needs, and hopefully, longitudinal variances among learners based on pedagogical

interventions. The project corpus will also be extended with the inclusion of more data. In addition, the project can be expanded in the near future to cover comparisons with existing and future learner corpora of other language combinations, especially those developed by other regional MUST partners who employ the same annotations scheme.

The project team would like to take the opportunity to thank once again the support by SCOLAR's R&D Research and Development Projects 2018-19, without which, such a large annotated corpus of learner data and high-impact project-related symposium would not be possible. Despite the difficulties caused by social unrest and Covid-19, we believe the project has generated quite satisfying outcomes.

Acknowledgements to the funder will be made in all future publications relating to the project (including the papers and book under review and under preparation). It is hoped that the project will serve as a start rather than an end of learner corpus research on the language proficiency of students in Hong Kong, and the project team will continue to explore the data collected for more possible applications beyond the project period.

(k) Bibliography

Al Khafaji, A. H. A. (2007). Translanguage. *Meta, 52*(3), 436-476.

Behrens, H. (2008). *Corpora in language acquisition research: History, methods, perspectives*. Amsterdam: John Benjamins.

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Chan, S. K., & Fong, G. C. F. (2016). Hong Kong speak: Cantonese and Rupert Chan's translated theatre. In C. Rojas & B. Andrea (Eds.), *The Oxford handbook of modern Chinese literatures*. London: Oxford University Press.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Cambridge: MIT Press.

Chow, I. C., & Wong, B. T. M. (2015). The mega-sized, multi-genre Chinese-English parallel corpus for computer-aided translation. In *the International Conference on New Horizons in Translation Technology*. Hong Kong, China, 24 April.

Cook, G. (2010). *Translation in language teaching: An argument for reassessment*. Oxford: Oxford University Press.

Corder, S. P. (1967/1983). The significance of learners' errors. In B. W. Robinett & J. Schachter (Eds.), *Second language learning: Contrastive analysis, error analysis, and related aspects* (pp. 163-172). Ann Arbor: The University of Michigan Press.

Evans, S. (2013). The long march to biliteracy and trilingualism: Language policy in Hong Kong education since the handover. *Annual Review of Applied Linguistics, 33,* 302-324.

Fan, M. Y. (2001). An Investigation into the vocabulary needs of university students in Hong Kong. *Asian Journal of English Language Teaching*, *11*, 69-85.

Fictumová, J., Obrusnik, A. & Štěpánková, K. (forthcoming). Teaching specialized translation: Error-tagged translation learner corpora.

Granger, S. (1998). *Learner English on computer*. London: Longman.

Granger, S. (2002). A bird's eye view of learner corpus research. In S. Granger, J. Hung & S. Petch-Tyson (Eds.), *Computer learner corpora, second language acquisition and foreign language teaching* (pp. 3-33). Amsterdam: John Benjamins.

Granger, S. (2003). The International Corpus of Learner English: A new resource for foreign language learning and teaching and second language acquisition research. *TESOL Quarterly, 37*(3), 538-546.

Granger, S., Dagneaux, E., Meunier, F., & Paquot, M. (Eds.). (2009). *International corpus of learner English: Version 2*. Louvain-La-Neuve: Presses Universitaires de Louvain.

Granger, S., & Lefer, M. A. (2017). *General report of the MUST kickoff meeting*. Louvain-la-Neuve: Centre for English Corpus Linguistics, Université catholique de Louvain.

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

Gui, S. C., & Yang, H. Z. (Eds.). (2003). *CLEC—Chinese learner English corpus*. Shanghai: Shanghai Foreign Language Education Press.

Horner, B., & Lu, M. Z. (2012). (Re) Writing English: Putting English in translation. In C. Leung & B. V. Street (Eds.), *English: A changing medium for education*. Bristol: Multilingual Matters.

Hymes, D. H. (1972). On communicative competence. In J. B. Pride, & J. Holmes. (Eds.), *Sociolinguistics: Selected readings* (pp. 269-293). Harmondsworth: Penguin.

Izquierdo, M., Sanz, Z., Zubillaga, N., & Manterola, E. (2021). Basque in student translations: What MUST tell us. Paper presented at MUST Workshop. Université catholique de Louvain, Beligum (virtual conference).

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., and Suchomel, V. (2014). The Sketch Engine: ten years on. *Lexicography*, *1*, 7-36.

Laviosa, S. (2014). *Translation and language education: Pedagogic approaches explored*. London: Routledge.

Lin, A. M. Y. (2015). Conceptualising the potential role of L1 in CLIL. *Language, Culture and Curriculum*, *28*(1), 74-89.

Lin, A. M. Y., & Man, E. Y. F. (2009). *Bilingual education: Southeast Asian perspectives*. Hong Kong: Hong Kong University Press.

Lin, L. H. F., & Morrison, B. (2010). The impact of the medium of instruction in Hong Kong secondary schools on tertiary students' vocabulary. *Journal of English for Academic Purposes, 9*(4), 255-266.

Lo, Y. Y., & Lo, E. S. C. (2014). A meta-analysis of the effectiveness of English-medium education in Hong Kong. *Review of Educational Research, 84*(1), 47-73.

Malmkjær, K. (1998). Introduction: Translation and language teaching. In K., Malmkjær. (Ed.), *Translation and language teaching: Language teaching and translation* (pp. 1-11). Manchester: St. Jerome.

McEnery, T., Xiao, R., & Tono, Y. (2006). *Corpus-based language studies: An advanced resource book.* New York: Routledge.

Naimushin, B. (2002). Translation in foreign language teaching: The fifth skill. *Modern English Teacher, 11*(4), 46-49.

Ngan, H. Y. W. (2009). Developing biliteracy through studying the bilingual representation phenomenon in translation texts. *Babel, 55*(1), 40-57.

Obrusnik, A. (2014). Hypal: A user-friendly tool for automatic parallel text alignment and error tagging. In *11th International Conference Teaching and Language Corpora*, Lancaster, 20-23 July 2014, 67-69.

PACTE. (2003). Building a translation competence model. In A . Fábio (Ed.), *Triangulating translation: Perspectives in process oriented research* (pp. 43–66). Amsterdam: John Benjamins.

Pan, J. (2012). *Problem analysis and the learning of interpreting: Perceptions, evaluation and corpus analysis of students' interpreting work (PhD dissertation).* City University of Hong Kong, Hong Kong.

Pan, J. (2014). Repetition and self-correction in students' interpreting performance: Corpus evidence of the "why" and "how". In the *4th Using Corpora in Contrastive and Translation Studies Conference*, Lancaster, United Kingdom, 24– 26 Jul.

Pan, J. (2017). A Corpus-based Study of College Students' Translation Performance: The Construction and Initial Findings of the HK-CL(CE/EC)TC." In S. Granger, & M. A. Lefer, *General Report of the MUST Kickoff Meeting* (pp. 147–168). Louvain-la-Neuve: Centre for English Corpus Linguistics, Université catholique de Louvain.

Pan, J. (2019 a). Researching translator and interpreter training: Convergences & divergences. Invited keynote speech presented at the Guangdong-Hong Kong-Macau Postgraduate Academic Exchanges in Foreign Languages and Translation, Sun Yat-Sen University, Zhuhai, 11-13 May.

Pan, J. (2019 b). Employing learner corpora in the study of translator and interpreter training: Implications from lexical cohesion. Paper presented at The Ewha GSTI Conference 2019: Science and Technology in Translation and Interpreting, Seoul, Korean, 9 November. In *Ewha GSTI Conference Proceedings*, p. 39.

Pan, J. (2021 a). Researching translator and interpreter training: convergences and divergences. Invited talk by School of Foreign Language Studies, Zhejiang Scie-Tech University (virtual seminar, 31 Mar).

Pan, J. (2021 b). Translator and (or versus?) interpreter training – topics, methods, and empirical findings Researching translator and interpreter training: convergences and divergences. Invited talk by Division of Humanities & Social Sciences of BNU-HKBU United International College (virtual seminar, 6 May).

Pan, J., & Chan, K. (2013). Investigating the routes to professional translators / interpreters: The construction and development of the HK-CL(CE/EC)TIC. In the *2nd Business Translation Forum of China*, Beijing, PRC, 25-26 May.

Pan, J., and Laviosa, S. (Eds.) (under review). *Corpora and Translation Education: : Advances and Challenges*.

Pan, J., & Wang, H. H. (2012). Investigating the nature of the semi-natural interpretation: A case study. In M. A. Jiménez Ivars & M. J. Blasco Mayor (Eds.), *Interpreting Brian Harris: Recent developments in translatology* (pp. 77–94). Switzerland: Perter Lang.

Pan, J., & Wang, H. H. (2017). The Development of Textual Competence in Student Translators: A corpus-based study of problems of coherence and cohesion. In *Translation in Transition 3 (TT3)*, Ghent, Belgium, 13-14 Jul.

Pan, J., & Wang, H. H. (2018). Learner factors relating to errors of coherence and cohesion in translation: Some preliminary findings. Paper presented at the MUST (Multilingual Student Translation) Workshop, Université catholique de Louvain, Belgium, 11 September.

Pan, J., & Wong, B. T. M. (2015a). Pragmatic markers in interpreted political discourse: A corpus-driven study. In the *International Conference on Corpus Linguistics and Technology Advancement (CoLTA)*, Hong Kong, 16-18 Dec.

Pan, J., & Wong, B. T. M. (2015b). Investigating pragmatic markers in interpreted political speeches from Chinese to English. In the *International Conference "Found in translation – translations are the children of their times"*, Bucharest, Romania, 10-11 Sep.

Pan, J., & Wong, B. T. M. (2017). Developing pragmatic competence in political retour interpreting: A corpus-driven study on the use of pragmatic markers. In the *Teaching Translation and Interpreting Conference*, Łódź, Poland, 15-16 Sep.

Pan, J., Wong, B. T. M., Chan, S., & Wang, H. H. (2021, 25 June). Investigating the Chinese and English Language Proficiency of Tertiary Students in Hong Kong: Perspectives from the Hong Kong Subset of the Multilingual Student Translation Corpus. Invited

paper at the 25th Anniversary Conference of the Standing Committee on Language Education and Research (SCOLAR). In *Programme Book* (page 4). The Hong Kong Convention and Exhibition Centre, Hong Kong.

Pan, J., Wong, B. T. M., and Wang, H. H. (under review a). Making a way through the jungle: Exploring learner data in translation. In Pan, J., and Laviosa, S. (Eds.). *Corpora and Translation Education: : Advances and Challenges*.

Pan, J., Wong, B. T. M., and Wang, H. H. (under review b). Navigating learner data in translator and interpreter training.

Pan, J., Wong, B. T. M., and Wang, H. H. (2021, 5-6 June). Making a way through the jungle: Exploring learner data in translator and interpreter training. Plenary paper at the International Symposium on Corpora and Translation Education. In *Programme Book* (page 15-17). Hong Kong Baptist University, Hong Kong (virtual conference).

Pan, J. & Wong, T. K. (2021, 18 Nov). Distortion in student translations: Annotation of the Hong Kong subset of the MUST corpus. Paper presented at MUST Workshop. Université catholique de Louvain, Beligum (virtual conference).

Pan, J., & Yan, X. J. (2012). Learner variables and problems perceived by students: An investigation of a college interpreting program in China. *Perspectives: Studies in Translatology, 20*(2), 199–218.

Pan, J., & Yan, X. J. (2014). Inaccurate pronunciation in students' interpreting performance: Evidence from a learner corpus. In *the 11th Teaching and Language Corpora Conference*, Lancaster, United Kingdom, 20–23 Jul.

Poulisse, N. (1999). *Slips of the tongue: Speech errors in first and second language production.* Amsterdam/Philadelphia: John Benjamins.

Ricardo-Osorio, J. G. (2008). A study of foreign language learning outcomes assessment in U.S. undergraduate education. *Foreign Language Annals, 41*(4), 590-610.

Richards, J. C., & Rodgers, T. S. (2001). *Approaches and methods in language teaching*. Cambridge: Cambridge University Press.

Selinker, L. (1972/1983). Interlanguage. In B. W., Robinett & J., Schachter (Eds.), *Second language learning: Contrastive analysis, error analysis, and related aspects* (pp. 173-196). Ann Arbor: The University of Michigan Press.

Selinker, L. (1992). *Rediscovering interlanguage*. London: Longman.

Shih, R. H. H. (2000). Compiling Taiwanese learner corpus of English. *Computational Linguistics and Chinese Language Processing*, *5*(2), 87-100.

Sidiropoulou, M. (2015). Translanguaging aspects of modality: Teaching perspectives through parallel data. *Translation and Translanguaging in Multilingual Contexts*, *1*(1), 27-48.

Steward, D., Bernardini, S., & Aston, G. (2004). Introduction: Ten years of TaLC. In G., Aston, S., Bernardini, & D., Steward. (Eds.), *Corpora and language learners* (pp. 1-20). Amsterdam: John Benjamins.

Sun, Y. Y., & Cheng, L. Y. (2013). Assessing second/foreign language competence using translation: The case of the college English test in China. In D. Tsagari & G. Floros, (2013), 235-252.

Tsagari, D., & Floros, G. (Eds.). (2013). *Translation in language teaching and assessment*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Tsang, W. K. (2008). Evaluation research on the implementation of the medium of instruction guidance for secondary schools. *HKIED Research Newsletter*, *24*, 1-7.

Wen, Q. F., Wang, L. F., & Liang, M. C. (Eds.). (2005). *SWECCL—Spoken and written English corpus of Chinese learners*. Beijing: Foreign Language Teaching and Research Press.

Wong, B. T. M. (2010). Semantic evaluation of machine translation. In *the 7th International Conference on Language Resource and Evaluation (LREC)* (pp. 2884–2888). Valletta, Malta, 19–21 May.

Wong, B. T. M., Chow, I. C., Webster, J., & Yan, H. B. (2014). The Halliday Centre Tagger: An online platform for semi-automatic text annotation and analysis. In *the 9th International Conference on Language Resources and Evaluation (LREC)* (pp. 1664–1667). Reykjavik, Iceland, 26–31 May.

Wong, B. T. M., & Lee, S. Y. M. (2013). Annotating legitimate disagreement in corpus construction. In the *11th Workshop on Asian Language Resources (ALR)* (pp. 51–57). Nagoya, Japan, 14 October.

Yan, X. J., & Pan, J. (2016). Backgrounds, attitudes and software application of tertiary-level Putonghua learners in Hong Kong: A focus group interview study (in Chinese). *Journal of International Chinese Studies, 7*(1), 176–188.

Yan, X. J., Pan, J., & Wang, H. H. (2010). Learner factors, self-perceived language ability and interpreting learning: An investigation of Hong Kong tertiary interpreting classes. *The Interpreter and Translator Trainer, 4*(2), 173–196.

Yan, X. J., & Wang, H. H. (2012). Second language writing anxiety and translation: Performance in a Hong Kong tertiary translation class. *The Interpreter and Translator Trainer, 6*(2), 171–194.

Yan, X. J., & Wang, H. H. (2014). The construction and application of an error annotated learner translation corpus in translation classes. In *11th International Conference Teaching and Language Corpora*, Lancaster, UK, 20-23 Jul.

Yan, X. J., & Wang, H. H. (2015). The interplay between software usage, motivation and gender differences: A survey based on a Putonghua classroom in Hong Kong. *Overseas Chinese Education, 76*(3), 368–376.

Yang, H. Z., & Wei, N. X. (Eds.). (2005). *COLSEC—College learners' spoken English corpus*. Shanghai: Shanghai Foreign Language Education Press.