IDENTIFICATION OF L2 INTERPRETESE: A CORPUS-BASED, INTERMODAL, AND MULTIDIMENSIONAL ANALYSIS

CUI XU

PhD

The Hong Kong Polytechnic University

2021

The Hong Kong Polytechnic University

Department of Chinese and Bilingual Studies

Identification of L2 Interpretese: A Corpus-based, Intermodal, and Multidimensional Analysis

Cui Xu

A thesis submitted in partial fulfilment of the requirements for the degree of Doctor of Philosophy

September, 2020

CERTIFICATE OF ORIGINALITY

I hereby declare that this thesis is my own work and that, to the best of my knowledge and belief, it reproduces no material previously published or written, nor material that has been accepted for the award of any other degree or diploma, except where due acknowledgement has been made in the text.

_____(Signed)

CUI XU (Name of student)

Abstract

Over the last two decades, there has been an upsurge of interest in the nature of translated language as a form of mediated communication (Baker, 1993). Abundant evidence has shown that translated language manifests some 'universal' lexical patterns that set it apart from non-translated or unmediated target originals, characterized by overall more simplified and conservative language use, an increased level of explicitness, and greater homogeneity among translated texts. As a special form of Translation (Pöchhacker, 2004), interpreting, by contrast, has received much less research attention in this regard due to, particularly, a much more daunting task of corpus construction and compilation. The current research follows the tradition initiated by Shlesinger (2008), and Shlesinger and Ordan (2012), by focusing interpreting per se as both spoken and mediated (i.e., having been translated) language with an aim to isolate interpreting-specific linguistic patterns. Special attention is diverted to simultaneous interpreting into a B language (abbreviated as SI), or L2 interpreting, and the linguistic patterns identified in comparison with non-mediated spoken language (abbreviated as NS) and L2 translation (abbreviated as WT) is thus defined as L2 interpretese.

Three major issues are explored, including: 1) What are the general variation patterns of 79 linguistic features under discussion in SI compared with NS and WT?; 2) Are the widely discussed translation-specific patterns also traceable in SI compared with NS and WT based on the current corpus data?; and 3) What are the general co-occurrence patterns of linguistic features under discussion in SI compared with NS and WT?

A parallel, intermodal, and (quasi-)comparable interpreting corpus named the LegCo+ corpus has been constructed, featured by one Cantonese component, namely, source speech (abbreviated as ST), and three English components (i.e., NS, SI, and WT) consisting of three pairs of corresponding subgenres, including two types of Q&A sessions, and Debates session. Two-phase data analyses have been carried out based on a selection of 79 linguistic features: an initial unidimensional analysis targeted at the first

two broad questions while a multidimensional analysis utilizing exploratory factor analysis (i.e., EFA) addressing the last one. The 79 linguistic features are based on two major sources: Biber's (1988) 67 linguistic features on register variation between spoken and written language; and the much-discussed linguistic features in previous studies on the nature of translated and interpreted language (e.g., Laviosa, 1998c; Sandrelli & Bendazzoli, 2005; Shlesinger & Ordan, 2012).

The unidimensional analysis is divided into two main sections: the first section deals with general variation patterns of the 79 linguistic features in SI, NS, and WT, with an aim to isolate SI-specific variation patterns; while the second section zooms in on two widely debated 'universal' linguistic patterns, that is, lexical simplification and explicitation (or increased explicitness as preferred in the current research), both across and within the three English varieties. Results of the unidimensional analysis indicate that: 1) Overall, there are great variations across SI, NS, and WT in terms of the distribution patterns of the 79 linguistic features. SI, in general, is characterized by either an overuse (i.e., more frequent use) or underuse (i.e., less frequent use) of linguistic features associated with more simplified, explicit, and potentially more conventional/conservative language production, compared with NS and WT, while there are other lexical patterns that cannot be readily interpreted, due to lack of information concerning the co-occurrence patterns of the identified linguistic features; 2) In terms of the 'universal' patterns of lexical simplification and increased explicitness, SI in general conforms to the overall patterns of lexical simplification and an increased level of explicitness in relation to unmediated, native spoken language (NS), expect for lexical density (an indicator for simplification) and 'that' adjective complements (an indicator for increased explicitness), which show the opposite trends. These patterns, however, are not always consistent when subgenre comparisons are carried out, indicating a possible genre influence (i.e., the influence of genre types) over the general variation patterns of interpreted language. As far as intermodal comparison is concerned, while SI shows consistent patterns of being more simplified than WT, characterized by lower STTR, higher top 10 vocabulary coverage, lower lexical density, and shorter average sentence length, the linguistic patterns

regarding increased explicitness are less clear-cut, and are not consistent across subgenres.

The multidimensional analysis also consists of two main parts: general co-occurrence patterns of linguistic features in SI, NS, and WT, with SI-specific patterns being highlighted along different dimensions; and the consistency of SI-specific co-occurrence patterns in terms of subgenre comparisons. The multidimensional analysis reports the following results: 1) Eight factors are extracted, accounting for about 40% of the total variance among SI, NS, and WT, but seven factor are kept in the end; 2) The seven factors are interpreted in functional terms as dimensions, based on the assumption that linguistic features co-occur to realize a shared communicative function (Biber, 1988); 3) SI exhibits specific linguistic co-occurrence patterns along all seven dimensions compared with NS and/or WT, albeit to varying extents. In many cases, SI shares more similarities than absolute differences with NS and/or WT. Dimension 1, 'Involved versus Informational Production', captures potential L2 *interpretese* as defined in the current research, which shows that SI is more marked in the use of linguistic features associated with involved and informal language production, while unmarked in the use of linguistic features indexing informational and integrated production. This pattern, however, is very genre-sensitive, as only one subgenre comparison (SI_A vs. NS_A) conforms to this pattern. Dimension 2, 'On-line Information Elaboration with Stancetaking Concerns', reveals the largest variance between SI and NS, but SI and WT show consistently negligible differences and the patterns are very homogeneous across subgenre comparisons, revealing possible shared co-occurrence patterns between translation and interpreting as forms of mediated language due to their more constrained nature in terms of on-line information elaboration, and also possible risk management behaviors of both interpreters and translators (Pym, 2008a). Dimension 3, 'Precise versus Simplified Description', reports distinctive intermodal differences, while the differences between SI and NS are much less noticeable. Along the remaining dimensions, SI shows more similarities with either NS or WT, but given the relatively intermediate dimension scores, the specific linguistic manifestations of co-occurrence patterns among the three English varieties are not very distinguishable. To sum up, interpreted language does showcase distinctive co-

occurrence patterns in relation to non-mediated spoken language and translated language from the same source along seven dimensions. However, sometimes these differences are genre-sensitive, and are not equally distinguishable.

This research is the first attempt to carry out a systematic study on linguistic variation across interpreted language (SI), non-mediated spoken language (NS), and translated language (WT), from both unidimensional and multidimensional perspectives. The inclusion of a multidimensional perspective, in particular, enriches the existing knowledge about the "multidimensional and multifaceted nature" (De Sutter & Lefer, 2020) of interpreted language, contributing to our knowledge of interpreting as a spoken form of mediated language. The focus on the L2 aspect of interpreted language makes up for the research lacuna due to the lopsided attention on native interpreting, shedding new light on the multi-constrained nature of (retour) interpreting. The fine-grained analysis on (sub-)genre variation on the general distribution or co-occurrence patterns of linguistic features of interpreted language has rich implications for future relevant studies (such as a multifactorial analysis). The findings on the linguistic patterns from either unidimensional or multidimensional perspectives also have implications for interpreter training and teaching.

Acknowledgements

This three-year-project can never be possible without the help and encouragement of many people to whom I am deeply indebted.

My deepest gratitude goes to my supervisor, Prof. Li Dechao, who inspired me to have carried out the current research project and offered me constructive guidance as well as enthusiastic encouragement during my PhD study. He always offered me firm support when I experienced doubts for my research design and taught me to think critically whenever possible. I also want to express my heartfelt thanks to my two external reviewers, Prof. Hu Kaibao and Prof. Yanxiu, who gave me inspiring advice for my dissertation revision, and my Board of Chair, Prof. Andrew Kay-fan Cheung, for his enduring support and encouragement. Sincere thanks also go to the academic staffs at the Department of Chinese and Bilingual Studies, Dr. Liu Kanglong, and Dr. Wu Zhiwei, for their invaluable suggestions and insightful ideas raised during my thesis confirmation.

I am truly grateful to Prof. Bart Defrancq from Ghent University, with whom I had the honor to work with during my research attachment programme, though our co-working time was greatly affected due to the Covid-19 pandemic that swiped across the globe. Nonetheless, Prof. Bart Defrancq was very generous in offering critical and detailed advice when reviewing parts of my dissertation chapters and helped broaden my view of interpreting research.

My sincere thanks are extended to Prof. Daniel Gile, Prof. Arnt Lykke Jakobsen, Prof. Sandra Halverson, Prof. Franz Pöchhacker, Prof. Haidee Kotze (Kruger), and Prof. Heidi Salaets, for their invaluable suggestions regarding my research at the 2018 CETRA Summer School, during which time I was still at the early stage of research design. Their comments and advice helped make my research more well-designed and feasible. Particularly I want to thank Prof. Haidee Kotze, who inspired me during CETRA to use the current variationist approach when I was still at a complete loss.

Special appreciation goes to Prof. Antony Pym, and Dr. Hu Bei, who shared generously with me one of the key literatures for my research. Since this piece of literature was manually written, they had to scan every page of it, which costed them a lot of time. I am truly grateful for their generous help.

I also want to thank many of my friends and colleagues at AG 518 since 2017, whose company and encouragement enlightened my life. Particularly I want to thank Dr. Song Shuxian, for she always expressed great concerns for my research and had detailed discussions with me. She also showed me great care and support during many difficult times. I also want to thank many other friends, Dr. Zhang Kaile and Li Ruitian for the memories we shared and all the difficulties we had gone through together. Their friendship will always be cherished.

My friends, Huang Haiyan from Ghent University and Joachim, whom I got to know when I was exchanging in Belgium, also helped and encouraged me a lot during that special period of time. I will never forget all the delicious meals Haiyan made for me, and the pep talks she gave. I am very grateful for their support, care, and friendship.

Mostly, I am deeply indebted to my loving family, my mom, dad, brother, and sister-in-law, without whose standing support and encouragement, love and understanding this dissertation can never be accomplished. My brother, who is also a PhD, is *de facto* the motivation for my doctoral journey, as he has firm belief in me, knowing that I would be courageous enough to continue the pursuit of knowledge after my master's study. I am deeply grateful for his encouragement. My family are my source of strength, and the love of my life.

Table of contents

List of abbreviations

NS: native speech

SI: simultaneous interpreting (into B)

WT: written translation (into B)

A: subgenre “Questions to the Prime Minister/Chief Executive”

B: subgenre “Questions to the Ministers/Secretaries”

C: subgenre “Debates”

NS_A: native speech _ “Questions to the Prime Minister”

NS_B: native speech _ “Questions to the Ministers”

NS_C: native speech _ “Debates”

SI_A: simultaneous interpreting _ “Questions to the Chief Executive”

SI_B: simultaneous interpreting _ “Questions to the Secretaries”

SI_C: simultaneous interpreting _ “Debates”

WT_A: written translation _ “Questions to the Chief Executive”

WT_B: written translation _ “Questions to the Secretaries”

WT_C: written translation _ “Debates”

TU: translation universals

List of Tables

List of Figures

List of appendices

Chapter 1 Introduction

1.1 Research motivations

The quest for the very nature of mediated language has been a great concern for translation scholars in the past few decades. Theorists of translation, by and large, assume that the language of translation, which is produced under particular sets of constraints, differs consistently and enormously from original language (or source language) (see also Baker, 1999; Toury, 1995). Toury (1995, p. 268), for example, argues that, "in translation, textual relations obtaining in the original are often modified, sometimes to the point of being totally ignored, in favour of (more) habitual options offered by a target repertoire". Based on Toury's argument, it seems that the language of translation shares greater similarity with unmediated target language than original language. Such an assumption, however, is not convincing, as translation is always constrained by "a fully articulated text in another language", under the influence of which the language of translation has long been regarded as a "deviant" representation of target language (Baker, 1999, p. 282). Constrained by sets of interwoven factors, particularly the highly cognitively challenging nature of bilingual processing, translated language manifests its own linguistic patterns that are believed to be distinguishable from both original and target languages.

Interpreting as "a form of Translation" (Pöchhacker, 2004, p. 11) is also a mediated language variety characterized by even more challenging cognitive processing. However, compared with its translated counterpart, only recently the nature of interpreted language has come to the spotlight. In comparison with translation, interpreting is mostly characterized by its "immediacy" (Pöchhacker, 2004), that is, "the source-language text is presented only once and thus cannot be reviewed or replayed", and "the target-language text is produced under time pressure, with little chance for correction and revision" (Kade, 1968; as cited from Pöchhacker, 2004, p. 10). Such feature of immediacy poses great cognitive constraints to interpreters, whose language production may exhibit great differences in relation to translated language produced by translators at their own pace.

This feature of immediacy is also partly attributed to the nature of interpreting as a form of spoken language, which often dies in the air once being uttered. It becomes even more prominent when simultaneous interpreting (SI) is under discussion, which is the focus in the current research. Although new forms of SI are emerging given their different intra-modal or inter-modal focuses (Pöchhacker, 2019), SI in the current research refers to the widely established "spoken language interpreting with the use of simultaneous interpreting equipment in a sound-proof booth" (Pöchhacker, 2004, p. 19). Therefore, it refers to voice-to-voice interpreting (or interpreting in its spoken mode) produced almost simultaneously (with several seconds of ear-voice-span) with source language speech. As simultaneous interpreters are highly paced and constrained by source speakers, while people speaking in their own languages "are free to speak their own mind and bypass possible production difficulties by rearranging the information and idea sequence, or by dropping or modifying information or using standard phrases" (Gile, 2009, p. 163), the languages interpreters produce are highly constrained and may manifest distinguishable linguistic patterns compared with unmediated spoken language.

Given this specific nature of SI being both spoken and mediated, it is assumed that interpreted language should be characterized by specific linguistic or lexical patterns that set it apart from the other two. Previous studies operationalize lexical patterns based on several selected linguistic features, such as list head coverage, high frequency words, lexical density, and standardized type-token ratio (STTR) for the study of simplification (Laviosa, 1998c), cohesive ties for the study of explicitation (Shlesinger, 1995), and idiomatic expressions for the study of conventionalization/normalization (Baker, 2004). The methods adopted are often frequency-based and unidimensional, in that the alleged 'universal' patterns for mediated languages are only examined along one single dimension. As Biber (1992) rightfully argues, the communicative possibilities offered by languages, which for sure include mediated language, is never unidimensional. The interplay of various constraints of language in mediation has strong implications for the "multidimensional and multifaceted" (De Sutter & Lefer, 2020) nature of mediated language variety. Following this line of thought, the current research carries out both

unidimensional and multidimensional analyses so as to inform the specific lexical patterns that isolate SI from unmediated native spoken language and/or written translation. Different from the majority of previous studies (e.g., Baker, 1995; Bernardini et al., 2016; Ferraresi et al., 2018; Laviosa, 1998c; Sandrelli & Bendazzoli, 2005), which almost unanimously focus on native translation and interpreting (i.e., translation or interpreting from a B language into an A language, or L1 translation/interpreting), this research project focuses on SI into a B language, or L2 interpreting. Out of many considerations for such a choice, one consideration concerns with the norm of interpreting practice in the Asian markets (such as mainland China, Hong Kong, Japan, and Korea), where interpreting into B is more frequently practiced, and is also the "facts of life" (Lim, 2003). This L2 (non-native) aspect of SI adds further complexity to the already complicated nature of mediated spoken language, making the topic under discussion even more intriguing.

In a nutshell, this research project sets out to explore and identify linguistic patterns specific to L2 interpreting (SI), that is, L2 interpretese, in relation to both native spoken language (NS) and L2 translation from the same source speeches (WT). The way for such an investigation is through the comparison of linguistic variation patterns across the three language varieties utilizing both unidimensional and multidimensional approaches. (Sub-)genre comparisons across the three language varieties have also been carried out to indicate if the identified SI-specific patterns are always consistent and genre-insensitive.

1.2 Defining L2 interpretese

The very term "interpretese" was first put forward by Miriam Shlesinger (2008) in her article titled "Towards a definition of *interpretese*: A corpus-based intermodal study". Despite the illuminating title, Shlesinger fails to provide a very clear definition of what *interpretese* really stands for. In a later research, Shlesinger and Ordan (2012, p. 55) clarify that the so-called *interpretese* refers to "the features of interpreted outputs, as distinct not only from their source or from 'similar' (non-translate, oral) texts in the same (target) language but also from written translations of the same (or 'similar') texts." The underlying motivation for such a definition, as argued by them (2012, p. 44), is to see

"whether interpreting is essentially 'the same' as translation, other than the fact that it happens to be oral; whether it is first and foremost a form of speech, with distinct spoken-like features that override its translation ontology". Based on this description, and their corpus data under study, at least three implications are drawn: First, since the *interpretese* under discussion opts for the default B-to-A working direction of SI, both mediated and unmediated language varieties are native, or L1 language varieties; Second, SI is approached as both Translated (i.e., mediated) and spoken language which requires not only comparable study, as is often the case in studies on distinctive features of written translation (see section 2.2.1), but also intermodal comparison which highlights the specific mode that interpreting is carried through; and Third, following a strict criterion, *interpretese* can only be identified when SI differs from both non-translated target language and written translations. However, their analysis indicates that they do not follow strictly the definition they have proposed, as the features they report are isolated either from comparable comparison between interpreted and non-interpreted target originals, or from intermodal comparison between interpreted and translated language. Moreover, in many cases, *interpretese* in their paper is used to refer to interpreting output, instead of the specific linguistic features or patterns of interpreting.

The very fact that the concept of *interpretese* is used interchangeably with other concepts such as output of interpreting or interpreting per se is not uncommon in corpus-based studies on the nature of interpreted language, or *interpretese* studies. He et al. (2016, p. 971), for example, equate *interpretese* with "interpreted language", with the underlying assumption that it is a dialect of language. The present author, however, decides to follow strictly *interpretese* as was originally defined. Since this research focuses in particular on L2 interpreting, modifications are made based on the definition of *interpretese* proposed by Shlesinger and Ordan (2012). The working definition for L2 interpretese is thus clarified as the following:

L2 interpretese refers to the specific linguistic patterns of L2 interpreting (SI) that isolate SI from both native/unmediated spoken language (NS) and L2 translation of the same

source (WT), based on a total number of 79 linguistic features selected according to previous studies on register variation (Biber, 1988) *and 'universal' features of translation and interpreting* (e.g., X. Hu et al., 2016; Laviosa, 1998c; Sandrelli & Bendazzoli, 2005).

These 79 linguistic features, as will be expounded in detail in Chapter three, are selected based on several sources. The first source is Biber's (1988) 67 linguistic features on register variation between spoken and written language, which is considered appropriate for the current research as one important comparison is intermodal comparison between interpreting and translation, which essentially are mediated spoken and written discourse. The second source is previous studies on 'universal' features of translation and interpreting (e.g., Baker, 1995; Bernardini et al., 2016; K. B. Hu & Tao, 2009; X. Hu et al., 2016; Kajzer-Wietrzny, 2012; Laviosa, 1998a, 1998c) focusing specifically on lexical simplification and explicitation. The third source is other features annotated automatically that are believed to be different across the three English varieties.

1.3 Research questions

The overarching goal of the current research is to identify linguistic patterns specific to L2 interpreting (SI), in comparison with native spoken language (NS) and L2 translation (WT), i.e., L2 *interpretese*. It attempts to address the following three main research questions (RQs):

RQ1: What are the general variation/distribution patterns of the 79 linguistic features in SI compared with NS and WT?

> RQ 1.1. Are there any statistically distinctive linguistic features that are either overused or underused in SI compared with NS and WT?
>
> RQ 1.2 What linguistic patterns do these distinctive linguistic features indicate?

RQ2: Can the two widely acknowledged 'universal' patterns of mediated language (translation and interpreting), i.e., lexical simplification and increased explicitness, be

confirmed in the current research from both comparable (SI vs. NS) and intermodal perspectives (SI vs WT)?

RQ 2.1 Is SI more simplified than NS and/or WT? Is this simplification pattern consistent across genre comparisons?

RQ 2.2 Is SI more explicit than both NS and/or WT? Is this explicitness pattern consistent across genre comparisons?

RQ 3: What are the general co-occurrence patterns of the 79 linguistic features in SI compared with NS and WT?

RQ 3.1 Based on the general co-occurrence patterns of the 79 linguistic features, how many dimensions are identified?

RO 3.2 How does SI differ from NS and/or WT along these dimensions? Are these identified patterns consistent across genre comparisons?

1.4 Data and methodology

Essentially, the current research is corpus-based, which means that corpus data are the main resources for research data, and corpus linguistics the main research methodology. In terms of corpus data, a million-size intermodal (quasi-)comparable corpus named the LegCo+ corpus comprising three English components, i.e., SI, WT, and NS, and one Cantonese source, i.e., ST, have been constructed (see section 3.1). The inclusion of three English components has made it possible for the identification of L2 interpretese from both comparable and intermodal perspectives, following previous research traditions. An intermodal comparable corpus, as acclaimed by Shlesinger (1998, p. 3; original emphasis),

> […] would allow for the identification of patterns specific to interpreted texts (regardless of their source language) as pieces of *oral discourse*, in relation to comparable texts in the same language. It would also allow us to identify the patterns which single out interpreted texts as distinct *oral translational products* in a given language irrespective of their

source languages, through comparisons with comparable written translational products.

In terms of the methodology, corpus analysis will be carried out in two phases. The first phase, targeting at the first two research questions (RQ 1 and RQ 2), follows the traditional research trajectory by performing a unidimensional analysis which is essentially frequency-based (De Sutter & Lefer, 2020). The main purpose is to inform general variation patterns of the 79 linguistic features in SI, in relation to NS and WT, and to testify if the widely accepted 'universal' patterns are also applicable to SI into B. The second phase, addressing the third research question (RQ 3), adopts a multivariate technique based on Biber's (1988) multidimensional approach (MD) on register variation, with an aim to account for some of the identified patterns which lack ready explanations via a unidimensional perspective. The underlying assumption for the MD approach is that different linguistic features often show similar patterns of variation by co-occurring together, and "strong co-occurrence patterns of linguistic features mark underlying functional dimensions" (Biber, 1988, p. 13). Therefore, it can help unveil hidden patterns that cannot be readily identified or interpreted by univariate analysis.

1.5 Structure of the thesis

The thesis consists of six chapters.

Chapter one sets out to introduce the justifications and motivations for the study of L2 interpretese. Definition of L2 interpretese is provided based on the original definition of interpretese proposed by Shlesinger (2008) and Shlesinger and Ordan (2012). Research questions to be addressed are raised, after which the data and methodology to be adopted are briefly introduced. Afterwards, the structure of the dissertation is described, followed by a brief introduction of the main terminologies used in this research.

Chapter two reviews studies on the properties of spoken and written language, including distinctive features of speech compared with writing, distinctive features across various spoken registers, and individual features distinctive of spoken language. The reviewed

studies on spoken language/discourse have offered methodological implications for this project. Interpreting as mediated spoken language, which is often approached from the perspective of 'translation universals', is examined in detail, based on which the research gaps to be filled in the current research are identified.

Chapter three gives a detailed account of research data and methodology, focusing particularly on the massive work of corpus compilation and construction, including data selection, transcription, annotation and segment. A two-phase data analysis, i.e., unidimensional analysis typifying previous studies on the nature of translated and interpreted language, and a multidimensional analysis following Biber's groundbreaking work (1988) on register variation, is expounded.

Chapter four presents the results of a unidimensional analysis regarding the general variation patterns of L2 interpreting, or SI into a B language, in relation to both native speech and written translation. A close examination on two heatedly debated lexical patterns, i.e., lexical simplification and increased explicitness (or as used in previous studies, explicitation), both across and within the three language varieties is carried out in great detail to indicate the possible differences and/or similarities with regard to previous studies on L1 or native interpreting. Interpretations of the results are done against the background of the limitations of this method.

Chapter five moves on to a multidimensional analysis based on a multivariate statistical technique (i.e., exploratory factor analysis) to identify the linguistic variations of SI along different dimensions, compared to NS and WT. Consistency studies have also been carried out to examine whether the identified linguistic variation patterns of SI are consistent along different dimensions within subgenre comparisons. Results of this Chapter have important implications for the "multidimensional and multifaceted nature" (De Sutter & Lefer, 2020) of L2 interpreting as a multi-constrained language variety.

Chapter six, in the end, concludes with the main findings of the two-phase analysis, the implications and significance of the current research, the limitations with respect to

corpus design and construction, as well as the multidimensional (MD) approach adopted in this research. Finally, future directions are reflected upon.

1.5 Terminologies

To avoid confusion, a few terminologies adopted in this research need to be clarified, with some of them to be used interchangeably.

For a start, the pair "spoken/written" and "oral/literate". Following Shlesinger (1989), the author uses the term pair "spoken/written" to refer to the medium or mode of communication (or modality), while "oral/literate" to describe textual qualities. However, there are cases where "oral/literate" (especially "oral") is used to indicate the medium, such as oral translation (i.e., interpreting). When "oral/literate" is used in previous literature to express the same meaning as "spoken/written", the author will follow their usage.

Second, discourse, texts, and language variety, which in the current research will be used interchangeably to refer to "language in actual use".

Third, modality, mediation modes, and mode of delivery. Modality refers to the medium or mode of communication, that is, spoken or written, while mediation modes are used to refer specifically to translation and interpreting. Since translation is in essence a form of (mediated) written language, while interpreting a form of (mediated) spoken language, the two terms can also be used interchangeably in the current research. Mode of delivery, by contrast, refers to the production conditions under which the spoken language (including ST, SI, and NS) is produced, such as prepared, impromptu, or mixed.

The term "setting" is used to describe "the social context of interaction […] in which the activity is carried out" (Pöchhacker, 2016, p. 13). The present study focuses on two settings, i.e., the House of Commons in the UK Parliament, and the Legislative Council of Hong Kong, which can be generally referred to as legislative or parliamentary settings.

Another important distinction is made between register and genre. In his seminal work, Biber (1988) uses "register" a cover term, such as spoken and written registers. Genres, by contrast, are more specific, and they "characterize texts on the basis of external criteria [relating to the speaker's purpose and topic]", such as press reportage, press editorials, popular lore, bibliographies, official documents, general fictions, and so on. In this research, three comparable genres are included during corpus compilation and construction with an aim to identify possible genre influence for the identified linguistic patterns specific to SI. Chapter three provides detailed description.

A final distinction concerns with mediated language and constrained language. Mediated language is used both in a broad sense and in a narrow sense in this research. When used in a broad sense, it refers particularly to the status of the texts being non-native, and the status of the texts being translated. When used in a narrow sense, it refers specifically to the status of the texts being translated. For instance, both translation and interpreting are mediated language in that they are Translated; while L2 translation and L2 interpreting are bi-mediated in that that are both Translated and non-native language use. By contrast, native or target language originals are unmediated. Constrained language, in comparison, is used in a much broader sense, since communication or any language use "is always constrained in some way", including physical constraints, physiological constraints, psychological constraints, cognitive constraints, etc.. In view of this, it can be used to refer to all communicative events, including native language production. Used in a narrower sense, it can refer to any "communication taking place under conditions where one or several of the potential limiting factors play a greater than average role" (Lanstyák & Heltai, 2012, p. 100). When it comes to translation and interpreting, constrained language is mostly used to highlight the bilingual cognitive processing experienced by interpreters and translators (H. Kruger & Van Rooy, 2016a).

Chapter 2 Literature review

As defined in Chapter one, interpreting is first and foremost a translational activity, and since the interpreting phenomenon under study refers specifically to voice-to-voice simultaneous interpreting, there is no doubt that interpreting is also a special type of spoken language. In this case, interpreting may share both the linguistic properties of unmediated spoken language, as well as some features of language in mediation (such as translation). In this chapter, the author provides a review of previous studies on the properties of spoken language, and corpus-based studies on the distinctive features of translation and interpreting, i.e., translation universals and interpretese. Besides, since special attention has been devoted to L2 interpreting, research on the influence of working direction on (potential) features of interpretese will also be delved into.

2.1 The properties of spoken language

2.1.1 Distinctive features of speech in comparison with writing

Speech dies in the air. Once uttered, "the text is no longer available for editing" (Cook, 2014, p. 34). That is probably why, despite the primary status of spoken language from a developmental perspective, it has not attracted sufficient attention in the research community, especially before the early twentieth century. In the views of traditional grammarians, only writing deserves academic attention because it is the true language and it "can be collected, stored, examined, manipulated and analyzed in ways that were until very recently impossible for spoken language" (Chafe & Tannen, 1987, p. 383). However, with the rise of modern descriptive linguistics, many scholars such as Ferdinand de Saussure, Edward Sapir, and Leonard Bloomfield "went out of their way to emphasize the primacy of spoken as opposed to written language, relegating the latter to a derived and secondary status" (ibid.). Since then on, there have been quite a number of studies (Akinnaso, 1982; Bamford et al., 2013; Bernstein, 1964; Biber, 1988, 1995, 2006a, 2014; Blankenship, 1962, 1974; Carter & Sánchez-Macarro, 1998; Chafe, 1982; Chafe &

Tannen, 1987; DeVito, 1966, 1967; Drieman, 1962; Gibson et al., 1966; Redeker, 1984; R. Reppen, 1994; Tannen, 1982, 1985) dedicating to finding out the distinctive features of spoken language, often with reference to its written counterpart.

Among these studies, the earlier ones (e.g., Blankenship, 1974; DeVito, 1966, 1967; Drieman, 1962; Gibson et al., 1966) are basically experimental studies based on small samples, often focusing on one specific pair of registers (e.g., conversations vs. literate prose) or one single topic. For example, Gibson et al. (1966) investigate both speech and writing production of 45 student subjects on one certain topic, using linguistic parameters such as average sentence length, average number of syllables per 100 words, type-token ratio and Flesch Human-Interest score, to indicate the differences between speech and writing. Overall, spoken language is found to be linguistically less complex (indicated by readability parameters) and superior (in human interest terms) than writing, its average sentence length significantly shorter, type-token ratio (TTR) much lower, and it contains significantly lesser syllables per 100 words compared with writing. Blankenship's (1962) study, however, presents a different picture when the investigated registers are prepared lectures (speech) and publications (writing) produced by the same group of speakers or writers. Adapting Fries's system of linguistic indicators, Blankenship (1962) finds overall less linguistic variations between spoken and written language. As a matter of fact, individual variations are found to be much more prominent than modality variation, which brings out the question regarding the choice of data samples (e.g., which register? which topic? which subjects?) that were commonly asked in earlier studies (see Akinnaso, 1982 for a comprehensive overview). The implication drawn from Blankenship (1962) is actually a matter of the possible conditioning factors to be considered (such as register) when intermodal comparison between speech and writing is carried out. Redeker (1984) also acknowledges that, once modality and topics are manipulated, there are still other co-varying factors such as registers, participants' level of education, age and gender that may influence the linguistic properties of spoken and/or written language. It thus seems inadequate to approach linguistic variation between speech and writing from a single dimension (such as spoken or written modality) without taking into consideration other

possible constraining factors such as register.

To make up for such a deficiency, as well as to approach the differences and/or similarities between speech and writing from a functional instead of a mere modality-oriented perspective, Chafe (1982) puts forward two underlying dimensions that are supposed to characterize speech and writing, i.e., fragmentation vs. integration, and involvement vs. detachment, across four registers along the oral-literate[1] continuum, based on two assumptions that "speaking is faster than writing" (p.36), and that "speakers interact with their audiences [while] writers do not" (p.37). His analysis shows that spoken language in general is more fragmented than written language, characterized by either omission of connectives or more frequent use of coordinating conjunctions given the fact that speech is often "produced in spurts" (p.37). Meanwhile, it is also characterized by features of involvement, manifested by first person references, speaker's mental processes, monitoring of information flows (such as *well*, *I mean*, *you know*), emphatic particles, vagueness and hedges, and direct quotations. No obvious variations within either spoken or written registers have been reported, suggesting overall consistent patterns of the 'typical' spoken or written discourse.

Given the mixed results reported in earlier studies, Biber (1988) introduces a new and more robust methodology, i.e., the multidimensional approach, or the MD approach, to uncover as fully as possible the complex relationship between spoken and written languages across a wide variety of registers. To justify this new research approach, Biber (1986, p. 385) argues that "[t]he communicative possibilities offered by a language are complex, and there is no reason to expect a single dimension to be the central discriminator among all text types". Hence a multidimensional perspective is needed to explore the linguistic variations across different spoken and written registers. To achieve this goal, Biber (1988) collects 23 genres (or registers) based on the Lancaster-Oslo-

[1] As clarified in Chapter one, differences are made between the two pairs of terms, spoken/written and oral/literate, following Shlesinger (1989). Spoken/written refers specifically to the medium or mode of communication, while oral/literate is used to describe textual qualities. A spoken text can be characterized by oral features (such as daily conversation), or literate features (such as formal speech), and a written text can also exhibit literate (such as official documents) or oral (such as personal letters) features, as reported also in Biber (1988).

Bergen-Corpus of British English (the LOB corpus) and the London-Lund Corpus of Spoken English, in addition to professional and personal letters. These texts, in Biber's (1988, p. 67) words, cover "the full range of situational possibilities available". After text selection, a total number of 67 linguistic features are identified and extracted based on previous studies, and their frequencies counted using computer programs written by the author himself. In total, six factors or dimensions[2] are identified as relevant utilizing a multivariate statistical technique called exploratory factor analysis. These dimensions are then interpreted in functional terms based on the assumption that "a cluster of features co-occur frequently in texts because they are serving some common function in those texts" (p.91). Among the six dimensions, Dimension 1 ('Involved versus Informational Production'), Dimension 3 ('Explicit versus Situation-dependent Reference'), Dimension 5 ('Abstract versus Non-abstract Information') and Dimension 6 ('On-line Information Elaboration') are found to be more revealing than Dimension 2 ('Narrative versus Non-narrative Concerns') and Dimension 4 ('Overt Expression of Persuasion') in distinguishing spoken and written registers. This finding also casts light on the present study, as one of the comparisons to be carried out is between simultaneous interpreting (as a form of mediated spoken discourse) and written translation (as a form of mediated written discourse). In his more recent work, Biber (2014) finds that the oral-literate opposition extracted in Dimension 1 is the most consistent one in register variation, which is linguistically constructed as "clausal vs phrasal" (p.16).

Ever since Biber's (1988) pioneering work, a rising number of scholars[3] (Biber, 2006a; R. Reppen, 1994; R. Xiao, 2009) have followed suit, adopting the MD approach in their investigation of register variation in more specific domains, such as elementary schools (R. Reppen, 1994), university contexts (Biber, 2006a), corporate communication (J.

[2] The six dimensions identified in Biber's study (1988).
Dimension 1: Involved versus Informational Production
Dimension 2: Narrative versus Non-narrative Concerns
Dimension 3: Explicit versus Situation-dependent Reference
Dimension 4: Overt Expression of Persuasion
Dimension 5: Abstract versus Non-abstract Information
Dimension 6: On-line Information Elaboration

[3] For a more complete review of studies, please refer to Conrad & Biber (eds.). Variation in English: Multi-dimensional studies published in 2001, and Bamfold, Cavalieri, & Diani (eds.). Variation and change in spoken and written discourse published in 2013.

Bowker, 2013), etc.. Despite these diverse focuses, there is one 'universal' dimension that has been consistently identified, that is the 'oral versus literate', or 'clausal versus phrasal' dimension. This pattern is even found in register variation among different regional Englishes (Helt, 2001; Van Rooy et al., 2010; R. Xiao, 2009), and in other languages such as Somali and Korean (Biber & Hared, 1992; Biber & Kim, 1994). This has a strong implication for the shared underlying construct, or dimension, for various language varieties, such as the three language varieties under discussion in the current study.

2.1.2 Distinctive features among spoken registers

Biber's study (1988) demonstrates that spoken and written languages do not necessarily situate at the two extremes of the oral-literate pole, given the considerable diversity of the registers or genres covered. Rather, there is an oral-literate continuum along which one spoken register (e.g., conversations) may situate far apart from one written register (e.g., government reports), while another spoken register (e.g., prepared speeches) may share more similarities with the aforementioned written register. This inconsistency or register variation shows that there are no absolute differences between spoken and written languages, which may also help explain the inconclusive findings from previous scholarship (e.g., Blankenship, 1962). Bearing this fact in mind, nowadays more academic endeavors (Al-Surmi, 2012; Forchini, 2012; Friginal, 2009; Helt, 2001; Quaglio, 2009) have been made with respect to linguistic variations across spoken registers. Friginal (2009), for example, examines three spoken registers (including call center interactions, spontaneous telephone conversation, and face-to-face conversation) to isolate the specific characteristics of the spoken language used for outsourced call center transactions. The MD analysis extracts altogether three factors, or dimensions[4] interpreted in functional terms, capturing the variations among the three spoken registers, and the language of outsourced call center transactions is found to be more addressee-focused, polite, and elaborate (positive features in Dimension 1), more planned and procedural

[4] The three dimensions identified in Friginal's (2009, pp. 81–96) study are:
Dimension 1: "Addressee-focus, polite, and elaborate information vs. Involved and simplified narrative"
Dimension 2: "Planned, procedural talk"
Dimension 3: "Management information flow"

(positive features in Dimension 2), and is better at information flow management (positive features in Dimension 3), compared with telephone calls or face-to-face conversation, indicating strong linguistic variations across different spoken registers.

Other studies focus on television dialogue, movie language and natural face-to-face conversations (Al-Surmi, 2012; Forchini, 2011; Quaglio, 2009), often driven by a pedagogical aim. Al-Surmi (2012), for example, compares two types of television dialogue, i.e., soap operas and sitcoms, to natural conversations, and finds that sitcoms are closer to natural conversations than soap operas with respect to Biber's (1988) D1 ('Involved versus Informational Production'), D4 ('Overt Expression of Persuasion'), and D5 ('Abstract versus Non-abstract information'), while soap operas resemble more natural conversations in D2 ('Narrative versus Non-narrative concerns'). Both soap operas and sitcoms share D3 ('Explicit versus Situation-dependent Reference') with natural conversations. Studies by Quaglio (2009) and Forchini (2011) also reveal great similarities between movie language and face-to-face conversation. These findings shed light on foreign language teaching, especially English as Second Language (ESL), since previously television or movie languages were not considered as genuine spoken language, as was the case with translated (or mediated) language (Baker, 1999)

One conclusive finding that can be drawn from this brief review is that linguistic variations do exist even among spoken registers, with some registers or genres exhibiting more features of orality, while others more features of literacy. This finding is very enlightening to the current research, as the spoken data under investigation consist of proceedings of different sub-genres (see Chapter three), such as "Debates" and Questions and Answers sessions, despite the fact that they are all produced within similar legislative settings. An exploration of the possible linguistic variations among these different spoken genres (of the same mediation mode) may cast light on the overall consistency of SI-specific patterns.

2.1.3 Individual features distinctive of spoken language

In addition to the comparison studies (spoken vs. written registers, or spoken vs. spoken registers) operationalized as several or a large number of linguistic features as reviewed in section 2.1.1 and 2.1.2, other studies orient towards more fine-grained analysis by zooming in on certain individual feature(s) distinctive of spoken discourse, such as greater reliance on formulaic language or lexical bundles (Altenberg, 1998; Altenberg & Granger, 2001; Biber, 2006a, 2009; Biber et al., 1999, 2004; Ellis & Simpson-Vlach, 2010; Lin, 2013; Martinez & Schmitt, 2012; Renouf & Sinclair, 1991; Sinclair, 1991; see also Staples, 2015 for a summary), more use of stance features (Barbieri, 2008; Biber, 2006b; Biber et al., 1999; Lindemann & Mauranen, 2001; Staples & Biber, 2014; J. Swales & Burke, 2003), discourse markers (Aijmer, 2002; Lam, 2010; Muller, 2005; J. M. Swales & Malczewski, 2001) and vague language (Adolphs et al., 2007; Cheng, 2007; Evison et al., 2007; Fernández, 2013).

Studies on the use of formulaic language[5], or lexical bundles, multi-word units as used often in corpus-based linguistic studies (Biber et al., 2004), have been gaining momentum in the past two decades in psycholinguistics, language acquisition, and corpus linguistics. The main argument across different research perspectives is that fluent speech production relies heavily on the use of habitual collocations or formulaic expressions which are stored and can be retrieved as a whole[6], thus saving cognitive efforts for on-line speech production (Bolinger, 1975; Goldman-Eisler, 1958; Nattinger & DeCarrico, 1992; Pawley & Syder, 1983; Wray & Perkins, 2000). The emergence of spoken corpora has made it possible to put into test such a claim, although corpus linguists refer often to the term "lexical bundle" operationalized as "n-gram" with a frequency-driven focus. Biber et al.'s (2004) seminal paper compares the use of lexical bundles in university teaching (spoken

[5] Currently there has been no consistent definition as regards 'formulaic language'. Wray (2002), for example, has summarized over fifty related terms, including, for example, collocations, lexical phrases, multi-word units, lexical bundles, prefabs, formulaic language, formulaic expression, prefabricated chunks, etc.. Research with different focuses tends to use different terms, but there is one consensus, that is the heavy reliance of formulaic language in spoken and written discourse, spoken discourse in particular. In this research, the author will not give a detailed distinction among these terms, as this is not the main focus here. Instead she will refer to these terms as more or less the same.

[6] Technically speaking, there have been not enough psycholinguistic studies confirming that formulaic language is retrieved as a whole, but theoretically such a claim has been widely acknowledged.

discourse) and textbooks (written discourse) from a functional perspective. They report that lexical bundles for stance expressions and discourse organizations are more frequent in spoken discourse, while referential expressions are more frequent in written discourse. They conclude by suggesting that "lexical bundles should be regarded as a basic linguistic construct with important functions for the construction of discourse" (p.398). In other words, lexical bundles are basic building blocks of discourse in both speech and writing serving for different purposes. In a later study, Biber (2009) finds that the multi-word patterns typical of speech differ significantly from those typical of writing: there are more fixed sequences (e.g., I don't want to) in speech while more structural 'frames' followed by a 'slot' (e.g., a * of the) in writing. What this may indicate is that speech may require more fixed sequences or prefabricated chunks to reduce on-line production efforts while keeping the flow of speech, writing may have a different purpose, such as to make the discourse as exquisite as possible. Since interpreting is a cognitively challenging form of spoken discourse, exploration of the use of formulaic language may also shed new light on the linguistic manifestation of interpreted language.

Other individual distinctive features of spoken discourse, such as more frequent use of stance markers, discourse markers, and vague language, have also received much research attention. However, often they are examined from the perspective of (critical) discourse analysis rather than register variation, so no detailed review will be provided here. Nonetheless, the more frequent use of such features has also highlighted the distinctiveness of spoken language, which also constitutes an interesting research field for further interpreting studies.

Despite the overwhelming amount of studies on spoken registers using corpus-based approaches, many scholars, as rightly noted by Friginal (2009, p. 292), still perceive these studies "as somewhat deficient and limited in the overall description of the discourse of speakers because segmental and suprasegmental features of speech are not captured in traditional transcriptions". Facilitated by the advancement of transcribing techniques, a new line of research focusing on fluency and prosody features in spoken discourse, as

well as non-verbal behaviour (e.g., gesture, eye movement), has been gathering momentum. Following this line of research, many studies (Cheng et al., 2005; Ferragne, 2013; Gut, 2009) focus on prosodic features such as intonation, prominence and pitch movement. However, since these prosodic features require accurate annotation done by ready-made software for speech analysis such as Praat, the current research will focus on somewhat different paralinguistic features annotated manually in the transcription, such as filled pauses (uhm, er), false starts, truncated words, and repairs, to indicate the prosodic features of interpreting as mediated spoken discourse.

To summarize the above review on the linguistic properties of spoken discourse, we can safely draw the conclusion that spoken discourse in general possesses some distinctive features (either lexico-grammatical, semantic and discoursal, or phonetic and prosodic features) with respect to its written counterpart. Nonetheless, this distinctiveness is not always consistent, especially when the spoken and written registers under discussion share similar purposes or functions (such as prepared speech vs. official documents), which indicates the complex relationship between spoken and written discourses. Moreover, variations among spoken registers (and likewise, written registers) also exist, revealing the complexities of the nature of language in use. Biber (1988), in summarizing his seminal work, concludes that there are no single absolute differences between speech and writing, and differences between spoken and written registers need to be considered along dimensions of variance rather than a single dimension.

2.2 Interpreting as mediated spoken language

Interpreting, like speech, also dies in the air. However, different from natural speech which is monolingual and unmediated, interpreting is a mediated activity with two languages activated at the same time. Another mediated activity that is closely related to interpreting is written translation, which is essentially a form of mediated written discourse/language. This shared feature of “mediation” brings together translation and interpreting as two varieties of Translation (in a generic sense). Treading into the field of translation studies can shed light on interpreting studies since they “share epistemological,

methodological, institutional and wider sociological concerns" (Gile, 2004, p. 10). Indeed, many studies on interpreting follow the research roadmap of translation studies, especially on the research topic of the 'universal'[7] features of meditated language, under the cover name of "translationese" (Gellerstam, 1986), "third code" (Frawley, 1984) or "translation universals" (Baker, 1993).

2.2.1 Corpus-based studies on distinctive features of translation

2.2.1.1 Definition of translationese: from pejorative to neutral

Before the establishment of descriptive translation studies (henceforth DTS) initiated by Toury, which seeks for an objective description of translated language, the very term 'translationese' was often used in a very pejorative way, signaling the awkwardness of translation caused by overreliance on source language (structures). Nida (1969, p.496), for example, believes that translationese is the result of an exaggerated degree of formal correspondence. Newmark (1988) refers to 'translationese' 14 times to indicate inaccurate or bad translations. In the Chinese literature, translationese is often translated into "*fanyiqiang*" (Liu et al., 2009), or "*fanyizheng*" (Libo Huang, 2005), indicating a sense of unnaturalness or "disorder" of the translated texts.

Though such negative perceptions of translationese still exist or even prevail under certain context (such as translator training), the term 'translationese' has gradually acquired a neutral aura, thanks to the upsurge of research interest in the nature of translation, with a special reference to the linguistic features of translated language that set it apart from non-translated language (Baker, 1993). Gellerstam (1986) is the first one to use the very term 'translationese' in a neutral sense. In his definition, translationese is used "in reference to […] systematic influence on target language (TL) from source language (SL), or at least generalizations of some kind based on such influence" (p.88). He (2005) further clarifies translationese as "all forms of translation which can in some form be viewed as having

[7] In the present study, the author will use the term 'universal' consistently to refer to general tendencies instead of an absolute sense of universal. The author chose this term since it has been widely accepted in previous literature, and the meaning of this terms has also evolved from original absoluteness towards general tendencies or regularities.

been influenced by the original text, without the term implying any value judgement" (p.202). The underlying assumption is that translations as a whole can be regarded as a special kind of language constrained by source language influence. Similar views are also expressed by Duff (1981), Frawley (1984), Schäffner and Adab (2001), and Hansen and Teich (2001) with the consensus that translated language is "a third language", "a third code", "a hybrid text" constrained by and also distinct from both source and target languages. It seems clear that Gellerstam's use of "translationese" emphasizes more on source language influence, while the others highlight both the influence of source language and target language, which can be traced in the output of translation.

2.2.1.2 From translationese to translation universals

Despite the vivid discussions on the hybrid nature of translation, the real academic pursuit of the recurrent patterns of translational language has been *en vogue* after Baker (1993, 1995, 1996) call for the introduction of corpus linguistics (henceforth CL) to DTS, with a particular focus on discovering the 'universal' features of translation as mediated communication, "features which typically occur in translated texts rather than original utterances and which are not the result of interference from specific linguistic systems" (p.143). She uses the term 'translation universals' (henceforth TU) instead. Different from Gellerstam (1986) and other scholars (e.g., Duff, 1981; Frawley, 1984; Schäffner & Adab, 2001) whose concerns are either on the possible influence of source language on the translational output, or the hybrid nature of translation, Baker seeks to unravel the complexity of translated language beyond source text influence by stressing the target text-target language dimension. She proposes a new research method by carrying out comparable analysis between translated and non-translated texts of the same language via monolingual comparable corpora. The majority of translation scholars has answered the call of Baker (1993), with some scholars sticking to the term 'translationese' while the majority of others follow the TU fashion.

The rise of corpus linguistics has made it possible to detect the general tendencies of

translation, or TUs, based on large-scale machine-readable corpora. Following Baker's (1993) call, the research community have been engaged in fervent pursuit of several potential TUs, i.e., simplification[8] (Corpas Pastor, 2008; L. He et al., 2010; Jantunen, 2001, 2004; Laviosa, 1997, 1998a, 1998b, 1998c, 2002; Steiner, 2012; Williams, 2005), explicitation[9] (Baker, 2004, 2007; Hansen-Schirra et al., 2007; Klaudy & Károly, 2005; Lapshinova-Koltunski & Vela, 2015; Olohan, 2003, 2004; Olohan & Baker, 2000; Øverås, 1998; Pápai, 2004; Puurtinen, 2003, 2004; Steiner, 2008; Williams, 2005), normalization[10] (Baker, 2004, 2007; Bernardini & Ferraresi, 2011; Hansen-Schirra, 2011; Kenny, 1998, 2001; Mauranen, 2008; Olohan, 2004; Scott, 1999; Stewart, 2000; Williams, 2005), and levelling out[11] (Laviosa, 2002; Williams, 2005), based on previous small-scale studies (e.g., Blum-Kulka, 1986; Shlesinger, 1991; Toury, 1980; Vanderauwera, 1985). In Baker's (1993) view, these translation universals are "linked to the nature of the translation process itself rather than the confrontation of the linguistic systems" (p.243). In other words, these observed features of translation are believed to be translation-inherent, irrespective of the source or the target languages involved.

In terms of the operationalization of these 'universal' features or tendencies, scholars following the Bakerian research paradigm approach them from either grammatical, lexical, syntactic or semantic perspectives. Simplification, for example, is usually examined based on four lexico-grammatical features, including lexical density, list head coverage, standardized type-token ratio, and high frequency versus low frequency words (H. Kruger & Van Rooy, 2012; Laviosa, 1998a, 1998c). Explicitation, the tendency to spell things out (Baker, 1996), is often approached on the basis of a number of linguistic indicators, such as an increased use of connectives (Puurtinen, 2003, 2004), using specific

[8] In Baker's (1996) definition, simplification refers to the "tendency to simplify the language used in translation" (p.181-182).
[9] In Baker's (1996) definition, explicitation refers to the tendency to "spell things out rather than leave them implicit" (p.180).
[10] In Baker's (1996) definition, normalization refers to the "tendency to exaggerate features of the target language and to conform to its typical patterns", and "the higher the status of the source text and language is, the less the tendency to normalize" (p.183). Toury's (1995) 'law of growing standadization' also expresses a similar idea that "textual relations obtaining in the original are often modified […] in favor of (more) habitual options offered by a target culture" (p.268).
[11] In Baker's (1996) definition, levelling out refers to "the tendency of translated texts to gravitate towards the centre of a continuum", and it is "neither target-language nor source-language dependent" (p.184). Laviosa (2002) uses the term 'convergence' instead, which indicates the "relatively higher level of homogeneity of translated texts with regard to their own scores on given measures of universal features" (p.72).

words for general ones (Klaudy & Károly, 2005; Øverås, 1998; Perego, 2003), making explicit pronouns (Li-bo Huang, 2008; Olohan & Baker, 2000; Pápai, 2004), an increased use of reformulation markers (Baker, 2004, 2007; R. Xiao, 2011), longer sentence length (Olohan & Baker, 2000), explicitating background knowledge (Pym, 2011), etc.. Though conflicting findings have been reported, translations overall are found to be more simplified, more explicit, and more conventionalized than non-translations of the same language (H. Kruger, 2018).

Besides the four widely investigated universal features, many other TU hypotheses have attracted research attention, such as the unique items hypothesis[12] or underrepresentation (Cappelle, 2012; Eskola, 2004; Kujamäki, 2004; Mauranen, 2000, 2008; Rabadán et al., 2009; Tirkkonen-Condit, 2004; Vilinsky, 2012), source language interference or shining through effect[13] (Hansen-Schirra, 2011; Mauranen, 2004; Teich, 2003), the asymmetry hypothesis[14] (Becher, 2010; Klaudy, 2009; Klaudy & Károly, 2005), the gravitational pull hypothesis[15] (Halverson, 2003, 2007, 2009, 2010; Hareide, 2017a, 2017b), and more recently the literal translation hypothesis[16] (Chesterman, 2011, 2017; Dimitrova, 2005). Some of these hypotheses (e.g., the asymmetry hypothesis, the gravitational pull hypothesis) have been put forward to explain (and predict) the contradictory findings in previous literature on universal features of translation. This brings forth one of the most criticized aspect of the TU research, i.e., the insufficient explanations for the identified features of translation. For many translation scholars, the reported 'universal' features are

[12] The unique items hypothesis put forward by Tirkkonen-Condit (2004) is based on the assumption that "linguistic items or elements which lack linguistic counterparts in the source language in question" (p.177) tend to be underrepresented in translations. In other words, target language-specific items which lack straightforward source language equivalents tend to be untranslatable and will thus be underrepresented in translations.

[13] Source language shining through effect, according to Teich (2003). Refers to the phenomenon that "in a translation into a given language (TL), the translation may be oriented more towards the source language (SL), i.e. the SL shines though" (p.143). Toury's (1995) 'law of interference' also expresses a similar idea that "in translation, phenomena pertaining to the make-up of the source text tend to be transferred to the target text" (p.275).

[14] In Zanettin's (2013) definition, the asymmetry hypothesis refers to the phenomenon that "explicitations in one translation direction are more frequent than their corresponding implicitations in the opposite translation direction" (p.23).

[15] According the Halverson (2010), the gravitational pull hypothesis "suggests that both over- and under-representation of particular target-language items is possible. However, the likelihood of a particular translated outcome (e.g. over- or under-representation) will depend on the specific structure of the bilingual semantic network activated in any given instance. Specific configurations will predict specific translational outcomes." (p.352)

[16] In Chesterman's (2011, 2017) definition, the literal translation hypothesis is formulated as this: "*during the translation process, translators tend to proceed from more literal versions to less literal ones*" (p.241; original emphasis).

translation-inherent due to the unique translation process, but contradictory findings have indicated that there are many other potential factors contributing to the surface manifestations of the translational product, such as source language difference, genre difference, translators' risk-avoidance concerns, etc.. It seems that this monofactorial translation-inherent perspective is no longer sufficient to unravel the mysteries of translational features. A multifactorial perspective taking into consideration the possible interaction of various constraints, as will be introduced in the next section, is more promising.

Nearly all these enlightening studies on the nature of translated language are monolingual comparable studies. This approach has been criticized by many (Rabinovich et al., 2015; Rodriguez-Castro, 2011; Santos, 1995) who call for the inclusion of source texts, and try to identify translation universals based on parallel corpora consisting of both source texts and translations. To straighten out these two lines of research, Chesterman (2004, p. 259) proposes the so-called S-universals (S for source) and T-universals (T for target). S-universals attempt to "capture universal differences between translations and their source texts, i.e. characteristics of the way in which translators process the source text"; while T-universals aim to uncover "universal differences between translations and comparable non-translated texts, i.e. characteristics of the way translators use the target language". As been pointed out by Chesterman (2004), although Baker's (1993) use of 'translation universals' are *de facto* T-universals, she actually includes many examples of S-universals (such as simplification), which has caused much confusion in the corpus-based TU studies.

Much as the pursuit of universal features in translation is thought-provoking, there have been many critical voices in recent years, represented by scholars such as Chesterman (2004, 2017), House (2008), Becher (2010, 2011), Evert and Neumann (2017), regarding the testing, representativeness, universality, conceptualization and terminology, operationalization, and causality of these 'translation universals'. Of particular relevance to the present research is the criticism made by Evert and Neumann (2017) who question the robustness of the statistical techniques adopted in TU studies, as they state clearly that

> [t]he use of statistical techniques to draw inferences from the observed patterns in a corpus to the underlying population is still not very well established in translation studies. If a statistical analysis is carried out at all, it is often limited to uni-variate techniques, e.g. comparing the frequencies of individual linguistic features between translations and originals with Student's t-test or a similar method. (Evert & Neumann, 2017, pp. 1–2)

Bearing in mind all these criticisms, translation scholars in recent years have been exploring new directions, in terms of both research methods and research scopes, in research on distinctive features of translation.

2.2.1.3 New directions in translationese and TU research

As far as research methods are concerned, one way directs towards natural language processing, represented in particular by machine learning techniques, which are assumed to be able to distinguish automatically translated texts from non-translated target originals (Avner et al., 2016; Baroni & Bernardini, 2006; Bernardini & Baroni, 2005; Bernardini & Ferraresi, 2011; Ilisei et al., 2010; Ilisei & Inkpen, 2011; Popescu, 2011; Volansky et al., 2015), and the related research findings "bring clear evidence of the existence of translationese features even in high quality translations" (Baroni & Bernardini, 2006, p. 260). However, as cautioned by Volansky et al. (2015, p. 27), even though machine learning approach based on comparable corpora "can settle the ontological question, […] we are left with an epistemological unease". In other words, although machine learning approach helps uncover and verify translation-specific features distinct from non-translated texts, it cannot provide a ready answer to the mystery of "why", i.e., the reasons why they differ in these features.

To partly address Volansky et al.'s (2015) concern, recently there has been a new research approach gaining popularity, namely the variationist multifactorial approach, as opposed to the previous univariate and monofactorial analysis tradition. One aim of this newly emerging method is to straighten out the constraining factors underlying the identified translational features. This new method, which often resorts to multivariate techniques,

can be attributed to the awakening awareness of the multidimensional nature of translation, as well as the conceptual refinements of the widely discussed translation universals (see also H. Kruger & De Sutter, 2018). De Sutter et al. (2017) even see this trend as a methodological shift from monodimensional comparable corpus analysis to multidimensional empirical analysis. Under this methodological shift, new sophisticated statistical techniques, such as Factor Analysis (FA), Principle Component Analysis (PCA), Correspondence Analysis (CA), the Multifactorial Prediction and Deviation Analysis (MuPDAR), etc., have been utilized to investigate translation universals such as explicitation (H. Kruger, 2019; H. Kruger & De Sutter, 2018; H. Kruger & Van Rooy, 2012, 2016b; Zufferey & Cartoni, 2014), normalization (Delaere, 2015; Delaere et al., 2012; Delaere & De Sutter, 2013; H. Kruger & Van Rooy, 2012; Prieels et al., 2015), simplification (H. Kruger & Van Rooy, 2012), levelling out or convergence (H. Kruger & Van Rooy, 2012; Vandevoorde et al., 2016), source language shining through (Evert & Neumann, 2017), as well as to invoke general discussions on typical features of translation (X. Hu et al., 2016), often taking into consideration various registers or text types in order to examine the main effect or interaction effect between translation status and register type.

Kruger and Van Rooy (2012) are among the earliest studies that cast doubt on the previous univariate-oriented studies on the nature of translation, though technically they do not adopt a multivariate technique but draw on analysis of variance to operationalize Baker's (1993) fourfold features of translation in a comparable corpus of translated English and original English produced in South Africa across six registers. Their results do not lend much support to the widely acknowledged universal features of translation as being more explicit, conservative, and simplified, except for two linguistic indicators, i.e., lexical density, and optional 'that'. They also find overall significant register variation between translated and original English, which disproves the hypothesis of "levelling out" of registers (Baker, 1996; Laviosa, 2002). In the same year, Delaere et al. (2012), echoing the call for a multivariate analysis in corpus-based translation studies, test the law of growing standardization utilizing a global multivariate approach, i.e., profile-based

correspondence analysis, to measure linguistic distances among different language varieties covering six text types and three language status (i.e., original Belgian Dutch, Belgian Dutch translated from English and French). A total number of 13 sets of variables (i.e., linguistic features of both standard and non-standard Belgian Dutch), or 'profiles' in their terminology, are examined to determine the profile-based chi-square distance among the nine language varieties. A two-dimensional plot indicates the general tendency of translated Belgian Dutch as being more standard than non-translated originals, and the identified differences are both source language and text type dependent.

Another representative study by Evert and Neumann (2017) probe into the impact of translation direction, an often-neglected variable, on the linguistic properties of translated texts. By exploiting a multivariate approach which combines a multivariate technique (i.e., principal component analysis), visualization, and supervised linear discriminant analysis (LDA), and based on a bidirectional parallel corpus of English and German texts, they detect a strong shining through effect in German translations (from English) than in English translations (from German). They propose the prestige effect discussed by Toury (2012) as a potential explanatory factor, but caution that further testing needs to be carried out. With respect to the multivariate approach, they claim that it "enables us to detect patterns of feature combinations which cannot be observed in conventional frequency-based analyses" (Evert & Neumann, 2017, p. 1), suggesting that "findings based on the (cumulative) interpretation of individual features may lead to spurious results that could be counteracted by other features not included in the study" (Evert & Neumann, 2017, p. 28). However, limitations have also been pointed out, as "the choice of features and texts heavily impacts the results". Nevertheless, as the authors state, "the multivariate approach […] is not only very useful for understanding the nature of translations, […] but is also very promising for various other areas of the study of language variation." (Evert & Neumann, 2017, p. 30)

More recently, a new sophisticated statistical technique named the Multifactorial Prediction and Deviation Analysis with Regressions (MuPDAR), developed by Gries and

Deshors (2014), has been utilized in the exploration of translational features as featured in Kruger and De Sutter (2018) and Kruger (2019). In their study, Kruger and De Sutter (2018) introduce this new statistical model to predict linguistic differences of the use of explicit or implicit 'that' among translated South African English (translated SAE), non-translated South African English (SAE), which is also a native variety of English but has a long contact relationship with Afrikaans, and British English (GBE, non-contact variety). Their analysis is based on several predictor variables relating to both complexity and conventionality with the aim to disentangle the different explanations for increased explicitness in translation. Their findings suggest that overall translated SAE resembles GBE more than SAE in terms of 'that' explicitation/omission pattern, but still "has its own distinctive strongly *that*-inclined fingerprint that sets it apart from the two non-translated varieties" (p.278). Based on their findings, they hypothesize that "translators are affected by risk aversion (using the most frequent and most formal option) and cognitive processing (due to bilingual activation), but not by CLI [cross-linguistic influence]; whereas non-translators are affected by CLI to a larger extent" (pp.280-281). In other words, despite the great similarities between translated English and non-translated English as both contact and non-contact varieties, translated English tends to be much more explicit in 'that' patterns due to a number of constraints absent from monolingual production. This finding is also consistent with the findings in Kruger and Van Rooy (2016a) and Kruger (2019).

The notion of 'contact variety' as used in Kruger and De Sutter (2018) reveals another trend in translation research, that is, widening research scopes or perspectives. Epitomized by the emergence of such concepts as "constrained language/communication" or "contact variety" (H. Kruger & De Sutter, 2018; H. Kruger & Van Rooy, 2016a, 2016b; Lanstyák & Heltai, 2012) from contact linguistics, related studies[17] tend to focus on the comparison between translated language and other contact varieties, non-native language (Gaspari & Bernardini, 2010; H. Kruger & Van Rooy, 2016a, 2016b), edited language (H.

[17] There are also studies which focus exclusively on contact language or contact variety, without reference to translated language, represented most by H. Kruger's studies.

Kruger, 2012, 2017), audio-visual translations (Prieels et al., 2015), and L2 or non-native language variety (H. Kruger & De Sutter, 2018; H. Kruger & Van Rooy, 2016b). The notions of "constrained language", "contact variety" or "contact language", based on the author's understanding, are not the same in that the former ('constrained language') stresses more the constraints involved in bilingual processing production, while the latter ('contact variety' or 'contact language') was originally proposed as a social or sociogeographical descriptor, but is used in these studies as a cognitive descriptor and it does not necessarily refer to bilingual processing production (H. Kruger & De Sutter, 2018). However, according to Lanstyák and Heltai (2012, p. 100), "both the language varieties spoken (and written) by bilingual communities and translated language are contact language varieties". As a matter of fact, they argue that when the bilingual translators are "off duty" engaging in monolingual communication, their communication "may also exhibit contact effects" (p.101). Kruger and Van Rooy (2016b) offer a straightforward explanation regarding the shared feature of all these different forms of language variety, that is "the transfer or cross-linguistic influence (CLI)" (p.119). In other words, all constrained language production, whether monolingual or bilingual, are in one way or another influenced by another language.

For translation studies, the underlying assumption is that translated language may share some features with other varieties of language produced by bilingual users, since all these varieties may be seen as contact varieties that are (in many ways) affected by psycholinguistic and social constraints of bilingual communication (Chesterman, 2004, 2015; Gaspari & Bernardini, 2010; Halverson, 2003, 2017; H. Kruger, 2012, 2019, 2019; H. Kruger & De Sutter, 2018; H. Kruger & Van Rooy, 2016a). This consensus has also echoed one of the previous criticisms made on the causality of the proposed 'universal' features of translation. That is, the so-called 'universal' features of translation are not unique to translation process *per se*, but rather are shared by general cognitive processing (Halverson, 2003; House, 2008; Szymor, 2018). An illustrative study by Kruger and Van Rooy (2016a) reveals great similarities between translated English and non-native indigenized variety of English as bilingual language production, since they both exhibit

preference for more explicit, more formal and normative choices due to both processing constraints and the "conscious construal of context and audience" (p.44). What we can learn from these studies is that, translations, after all, may not be as unique as we think. To unveil the multifaceted nature of translation as fully as possible, traditional comparable analysis between translated and non-translated texts is far from enough.

To summarize, the quest for translation universals or features of translation continues to gain momentum in the past decade thanks to the introduction of more sophisticated statistical methods and new research scopes and perspectives. Previous methodological doubts concerning issues of corpus (in)comparability (Bernardini & Zanettin, 2004), and theoretical criticisms about the universality, representativeness and causality of descriptive universals (Chesterman, 2004) have been addressed one by one, with more sound corpus design and more refined conceptualizations. However, there is one question that does concern scholars of interpreting: are these identified universal features of written translation also applicable to oral translation, that is interpreting?

2.2.2 Corpus-based studies on distinctive features of interpreting

Compared with the rigorous pursuit of distinctive features of translated language, studies on the nature of interpreted language are far lagging behind, due to particular obstacles in corpus-based interpreting studies (henceforth CIS) (Bendazzoli & Sandrelli, 2009; Sandrelli et al., 2010; Setton, 2011; Shlesinger, 1998; Straniero Sergio & Falbo, 2012), such as the unavailability of interpreting data due to confidentiality issues, the arduous and labor-intensive process of transcription, as well as the intrinsic evanescence of the spoken word.

In recent years, thanks to the rapid advancement of modern techniques such as automatic speech recognition and the increasing availability of online resources, such as parliamentary proceedings, some of the challenges have been coped with. This has greatly facilitated the development of CIS, which allows the research community to go beyond the 'black box' of simultaneous interpreters in a search for the typical patterns of

interpreted language.

2.2.2.1 Interpretese: A variant of translationese?

The notion of interpretese, as introduced in Chapter one, was first proposed by Shlesinger (2008), and further clarified in Shlesinger and Ordan (2012), referring to interpreting-specific linguistic features distinct from both written translations of the same or similar source texts and non-translated spoken originals. As a matter of fact, the same idea of identifying interpretese from both comparable (interpreted vs. non-interpreted) and intermodal (interpreted vs. translated) perspectives had long been discussed in her early work (1998), in which she called for corpus-based interpreting studies as an offshoot of corpus-based translation studies. Following this line of thought, Shlesinger (2008) first sets out investigating interpretese by comparing interpreted language with translated language to isolate modality-dependent features of interpreting. Focusing on several lexico-grammatical features[18] and based on the experimental outputs of six professional translator-interpreters rendering the same English text into Hebrew in both modalities, she observes strong modality-induced features that set (simultaneous) interpreting apart from translation, i.e., features of *interpretese* as she called. Interpreted language is found to be different from translated language in every aspect of the investigated features, exhibiting a strong influence of its oral modality. As a follow-up study, Shlesinger and Ordan (2012) expand their research to include both intermodal and comparable perspectives, with an aim to isolate the role of modality (being spoken or written) and ontology (being translated or non-translated) on the features of interpretese. Using a small comparable intermodal corpus, which contains interpreted texts, non-interpreted target originals, and translated texts of the same source, they examine altogether 29 linguistic features. Their results indicate that interpreted language exhibits more features of non-interpreted spoken originals. In other words, interpreting is found to be more spoken than translated, and viewed differently, it can be seen as "an extreme case of translation"

[18] The lexico-grammatical features investigated in Shlesinger (2008) include lexical variety (operationalized as type/token ratio), the verb system, the definite article, parts of speech (noun, adjective, verb, preposition, conjunction, adverb, participle, pronoun, negation, and copula) , possessives and lexical choices.

(Shlesinger & Ordan, 2012, p. 54) since all the features operationalized in translation studies, such as lower lexical density and lexical variety, are found to be more salient in interpreting.

Besides the systematic discussion on interpretese by Shlesinger, X. Y. Xiao (2015) also contributes to the investigation of interpretese, albeit with a different focus – to testify the equalizing effect as observed in Shlesinger (1989). Different from the use of the term "interpretese" in Shlesinger (2008) and Shlesinger and Ordan (2012), which basically regarded it as a variant of translationese, X. Y. Xiao (2015, p. 80) refers to interpretese as

> the output of interpreting, the rendition of a message in the target language produced consecutively or simultaneously in real time by an interpreter to represent the information and communicative effects of an original speech in a source language produced in a communicative setting.

Further still, she views interpretese as a genre in its own right, which is oral, translated, and covers a wide range of registers "from highly interactive conversations to formal speeches read from a written script" (X. Y. Xiao, 2015, p. 89). Thus compared with Shlesinger's (2008) definition, interpretese as defined in Xiao (2015) is a much more broad concept, as she equates interpretese with interpreting output with highlighted features.

Up till now, there has been no consensus regarding the conceptualization of interpretese, except for the few studies reviewed above. The vast majority of studies use interchangeably interpretese and interpreting (as well as translation) universals by focusing on the confirmation of the acclaimed universal features of (written) translation with reference to oral translation, i.e., interpreting. Facilitated by the emergence of interpreting corpora, such academic efforts in the search of interpreting-specific linguistic patterns have been made possible.

2.2.2.2 From interpretese to interpreting universals

The rigorous pursuit of the 'universal' features of translated language as being more simplified, explicit, conventional, and levelling-out has encouraged related studies with reference to interpreted language. Among the diverse TU hypotheses, the interpreting research community diverts more attention to the verification of simplification (Bendazzoli & Sandrelli, 2005; Bernardini et al., 2016; Dayter, 2018; Kajzer-Wietrzny, 2012, 2015; Lv & Liang, 2018), explicitation (Baumgarten et al., 2008; Dayter, 2018; Gumul, 2006, 2007, 2008, 2015, 2017; K. B. Hu & Tao, 2009; Kajzer-Wietrzny, 2012, 2018; Morselli, 2018; Shlesinger, 1989, 1995; Tang, 2018), normalization (K. B. Hu & Tao, 2010; Kajzer-Wietrzny, 2012; Shlesinger, 1991), source language interference (Dayter, 2018), and equalizing effect (Pym, 2007; Shlesinger, 1989; X. Y. Xiao, 2015) in interpreted texts. Given the nature of interpreted language as being not only spoken but mediated, interpreting has been approached from either a comparable perspective (interpreted vs. non-interpreted) (Bendazzoli & Sandrelli, 2005; K. B. Hu & Tao, 2009; Kajzer-Wietrzny, 2012), following the research trajectory of translation studies, an intermodal point of view (interpreted vs. translated; SI vs. CI) (Bernardini et al., 2016; Ferraresi et al., 2018; Gumul, 2007, 2012; H. He et al., 2016; Kajzer-Wietrzny, 2015; Morselli, 2018; Shlesinger, 2008), with an aim to isolate modality-specific features, a parallel angle (source vs. interpreted) (Shlesinger, 1989), or a mixed combination (e.g. K. B. Hu & Tao, 2010; Kajzer-Wietrzny, 2018; Lv & Liang, 2018; Shlesinger & Ordan, 2012; X. Y. Xiao, 2015).

One of the pioneering works on the general tendencies of interpreting is Shlesinger (1989), in which the equalizing (or "avoid the extremes") effect of SI on the positioning of the texts along the oral-literate continuum has been explored in great detail. Based on some experimental dataset consisting of four Hebrew texts and four English ones, along with their simultaneous interpretations (i.e., Hebrew<>English, and English<>Hebrew), and resorting to five parameters[19] identified in previous literature isolating orality/literacy

[19] The five parameters utilized in Shlesinger (1989, p. 11) are a. degree of planning; b. shared context of knowledge; c. lexis; d. nonverbal features; and e. degree of involvement.

features, Shlesinger (1989) finds a mixed, but overall equalizing tendency of SI which renders literate texts (both Hebrew and English) more oral in the interpreted versions, and oral Hebrew texts more literate[20]. Baker (1996, p. 184) refers to such a tendency as the universal of "levelling out" by summarizing that

> oral texts take on more literate features in simultaneous interpreting and literate texts become more oral. In other words, the process of translation tends to move texts more towards the centre of the oral-literate continuum, to locate them away from either extreme.

Such a generalization made by Baker (1996) has been criticized by Pym (2007, p. 13), who argues that Baker has simplified Shlesinger's research, "failing to mention the mixed findings for oral texts" as reported here, and "she renamed universal (proposing 'levelling out' instead of Shlesinger's 'equalizing' universal)", which is not exactly the same, since the equalizing effect emphasizes the role of translation in changing the linguistic patterns/features of the source texts, while the 'levelling-out' universal stresses more the patterns of translated/interpreted texts in the target language and culture.

In addition to the testing of the equalizing tendency in SI, Shlesinger (1989) also examines the role of mode of delivery on the explicitation universal, restricting the analysis to "the implications of shifts in the second parameter of orality" (p.172), i.e., shared context of knowledge. The preliminary findings offer counterevidence to explicitation, as there is overall a "greater degree of contextualization (which mutatis mutandis correlates with implicitation) in the interpretations of English oral-type texts as well as all of the literate texts, irrespective of language" (p.173; original emphasis). While Shlesinger (1998) fails to provide any explanation regarding the reduced level of explicitness in SI output, Pym (2007, p. 14) attributes it to the interpreter's risk-aversion strategy, claiming that

> because she [the interpreter] lacks knowledge of the context, the interpreter actually misses some of the cohesion patterns, therefore resulting in less explicitation in the rendition. When you are not sure of what is going on, you cannot risk underlining relations that are no more

[20] However, this is not borne out in the interpreted texts of oral English. Shlesinger (1989) attributed it to the inappropriate choice of features which turned out to be more typical of literate texts.

> than guesswork. A far better strategy in such situations is to say less, to use superordinates in case of doubt, and to stay close to the given cues, since even if you don't understand, there is a good chance the audience will. Hence the use of considerable lexical implicitation.

The quest for the equalizing effect of SI has not received much attention until over two decades later, when Xiao (2015) sets out examining the role of SI on the oral-literate continuum based on the language pair Chinese-English (bidirectional). Different from Shlesinger (1989) in which five parameters of orality/literacy distinction are selected and compared, Xiao (2015) adopts a methodological framework combining and adapting Biber's (1988) six dimensions as reviewed in section 2.1, as well as some of the dimensions in Shlesinger (1989), integrating altogether five dimensions[21] for English texts and four dimensions[22] for Chinese texts. By comparing the frequencies of 21 out of 67 linguistic features investigated in Biber (1988, 1995) between English/Chinese source texts and target texts, she reports mixed findings along the examined dimensions, but overall an equalizing trend. She then carries out factor analysis with an aim to identify the co-occurrence patterns of these features in the English data so as to confirm the already labelled dimensions (in functional terms) under examination. In her last step, she attempts to situate the investigated register, i.e., panel discussions, along the continuum of the five dimensions in comparison with other registers selected from Biber's (1988). Overall, an equalizing effect of SI has been confirmed, though there are mixed performances between Chinese and English along the five dimensions.

Xiao's (2015) research is no doubt the first attempt to combine the study of interpretese and interpreting universals with methods on register variation. Very much enlightening as

[21] The four dimensions for English texts in Xiao (2015, p. 101) are:
Dimension A: Involvedness (private verbs, contractions, first person pronouns, second person pronouns, and wh-questions);
Dimension B: Constrainedness (that complements, non-phrasal coordination, causative clauses, conditional clauses, and demonstratives);
Dimension C: Context-boundedness (time adverbials, place adverbials, wh-relative clauses, and phrasal coordination);
Dimension D: Abstractness (conjuncts, passives, and nominalizations);
Dimension E: Prosody (rate of delivery, pauses and disfluencies)

[22] The four dimensions for Chinese texts in Xiao (2015, p. 113) are:
Dimension A: Involvedness (private verbs, first person pronouns, second person pronouns, and wh-questions);
Dimension B: Constrainedness (causative clause, conditional clauses, and demonstratives)
Dimension C: Context-boundedness (time adverbials, place adverbials);
Dimension E: Prosody (rate of delivery, pauses, and disfluencies)

it is, this research is essentially deductive rather than inductive, in that the framework of dimensions has already been set in place, making redundant factor analysis carried out in the final part of her research. Besides, as acknowledged also by the author herself, the bulk of Xiao's research focuses only on the English data, leading to rather lopsided findings. More carefully designed research on the interpretese features needs to be done.

The two most researched universals in CIS are simplification and explicitation, approached often from either a source-text-oriented, parallel analysis or a target-text-oriented comparable perspective; while recently, an intermodal and even intermodal comparable perspective has been introduced. Simplification, "the process and/or result of making do with less words", was first observed in Blum-Kulka and Levenston (1983) regarding language learning and systematically investigated in Laviosa-Braithwaite (1996) and Laviosa (1997, 1998a, 1998c) in written translations. Research on simplification in interpreting can be traced back to Sandrelli and Bendazzoli (2005), who testify the simplification hypothesis in simultaneously interpreted language. Following Laviosa (1998c), Sandrelli and Bendazzoli (2005) investigate lexical density and list head coverage in four language combinations, i.e., Spanish/Italian to English, and Spanish/English to Italian, based on the European Parliament Interpreting Corpus (EPIC). Their results lend limited support to lexical simplification in interpreting, since only lexical density in interpreted Italian from English and list heads in interpreted English from Italian confirm the simplification trend. In terms of the possible reasons for the mixed patterns, Sandrelli and Bendazzoli (2005, p. 15) attribute them to the intrinsic constraints of SI in terms of

> the specific text production conditions, i.e. the pace of incoming speech is imposed by the source speaker and the interpreter has to assemble the target speech practically 'on-line', chunk by chunk, by selecting and re-arranging information to suit the norms of the target language. The parallel co-existence of source and target speeches and the time constraints under which interpreting is performed may explain why the patterns observed by Laviosa in relation to written texts do not apply.

Based on their findings and explanations, at least two implications can be drawn: 1)

conclusions drawn from comparison between translated texts and non-translated texts cannot be readily applied to interpreting; and 2) language pairs and working direction may have a direct influence on linguistic variation patterns.

Sandrelli and Bendazzoli's (2005) mixed findings have also been borne out in Russo et al. (2006) on interpreted Spanish, and Kajzer-Wietrzny (2012, 2015) on interpreted English from both comparable and intermodal perspectives. In her PhD dissertation, Kajzer-Wietrzny (2012) sets out to testify whether the widely acknowledged translation universals, i.e., simplification, explicitation, and normalization, are also applicable to interpreting. Based on a well-designed corpus named the Translation and Interpreting Corpus (TIC) composed of plenary sessions in the European Parliament, Kajzer-Wietrzny (2012) investigates three indicators for simplification, including lexical density, list heads, and high frequency words, following also Laviosa (1998c). Her comparable analysis offers counterevidence to the simplification hypothesis in interpreted language. In her later study (2015) focusing on an intermodal comparison between SI and WT using the same corpus (i.e., TIC), and the same indicators for simplification, Kajzer-Wietrzny (2015), once again, finds limited support to simplification in interpreting, while the opposite is true for translation, with respect to the variation patterns of the three linguistic indicators. In terms of lexical density, while translated English, conforming to the simplification hypothesis, shows lower lexical density than its non-translated originals, interpreted English shows quite the opposite. One possible reason for the higher lexical density in SI may be the interpreter's avoidance of redundancy due to time constraints, as Kajzer-Wietrzny (2015, p. 248) argues that "[i]ncreased lexical density may, therefore, be the result of condensation techniques used by interpreters to save time". Another explanation is attributed to an explicitating shift from referential to lexical cohesion (Gumul, 2006; Shlesinger, 1995). In terms of high frequency words, translators tend to use more high frequency words compared with native English writers; by contrast, interpreters opt for the opposite. As for list heads, which indicate degree of repetitiveness, interpreted English demonstrates lower tendency towards repetitiveness, except for the Spanish-to-English direction.

Despite these contradictory findings in interpreting studies, recent studies by Bernardini et al. (2016) and Ferraresi et al. (2018) lend support to lexical simplification in interpreting. Replicating Laviosa (1998b, 1998c), Bernardini et al. (2016) probe into simplification from an intermodal comparable perspective based on the newly expanded bidirectional (English<>Italian) corpus of interpreted and translated EU Parliament proceedings, i.e., EPTIC. Their results have generally confirmed lexical simplification in both translation and interpreting as languages in mediation, and they explain that "the mediation process reduces complexity in both modes of language production and both language directions, with interpreters simplifying the input more than translators" (Bernardini et al., 2016, p. 61). In conclusion, they concur with Shlesinger and Ordan's (2012, p. 54) view of interpreting as "an extreme case of translation" as far as simplification is concerned.

Probably the most discussed as well as highly controversial universal feature in translation studies is explicitation. Since interpreting in its spoken modality is perceptibly distinct from written translation, compounded by its intrinsic constraints related to time, linearity and (un)shared knowledge (Schjoldager, 1995; Shlesinger, 1995), it is claimed that increased expliciteness is highly unlikely in interpreting. Previous studies (Baumgarten et al., 2008; Defrancq et al., 2015; Gumul, 2006, 2008, 2017; Ishikawa, 1999; Niska, 1999; Shlesinger, 1995; Tang, 2014), however, have found consistent patterns of increased explicitness in interpreting, albeit with some contradictory findings (Kajzer-Wietrzny, 2012; Morselli, 2018; Shlesinger, 1989). Gumul (2007), for example, identifies fifteen types of explicitating shifts in both simultaneous and consecutive interpreting, five of which are statistically distinctive between the two modes, indicating that SI is less explicit than CI. In terms of the possible reasons, Gumul (2007) attributes them to "the time pressure and the need to allocating processing capacity resources to three competing concurrent operations: the Listening and Analysis Effort, the Production Effort, and the short-term Memory Effort (Gile, 1995, 1997)" (p.455). In other words, the lower level of explicitness in SI is contributable to its distinctive intrinsic constraints in relation to CI. Besides, certain explicitating shifts are also believed to be the results of

padding strategies adopted by simultaneous interpreters, and some are "purely textually motivated explicitation" (ibid.). In her earlier study, Gumul (2006) also reveals that the vast majority of the investigated explicitating shifts are sub-conscious and involuntary, "not attributable to the interpreters' conscious strategic behaviour" (p.171).

Aside from the more fine-grained analysis of explicitating shifts which are source-text-oriented, Kajzer-Wietrzny (2012) approaches explicitation from the perspective of translation universals. By examining the linguistic patterns of three indicators for expliciation, that is, optional 'that' after reporting verbs, linking adverbials, and apposition markers, in interpretations versus non-interpretations, Kajzer-Wietrzny (2012) fails to find any consistent patterns of more explicit language use in interpreted texts, except for optional 'that'. To testify these findings, Morselli (2018) investigates the same parameters based on EPTIC. His findings are generally in line with Kajzer-Wietrzny (2012), as there is "no clear evidence of more or less explicitness in interpreted/translated versus untranslated speech, and therefore no evidence for a universal tendency in its strictest sense" (Morselli, 2018, p. 10).

One final study worth mentioning is Dayter (2018), which presents similar mixed findings with respect to explicitation as well as simplification. Based on a newly constructed parallel aligned bidirectional corpus of Russian-English simultaneous interpreting (SIREN), Dayter (2018) touches upon three universals (i.e., simplification, explicitation and source language shining through) using scores for lexical variety, lexical density and POS proportionalities[23]. Her results show that, except for source language shining through effects, the English and the Russian data demonstrate opposite trends. While interpreted Russian conforms to lexical simplification and explicitation, interpreted English shows the opposite trends. With regard to the possible reasons, Dayter (2018) mentions one promising variable, namely, the working direction of simultaneous interpreting. While English-to-Russian interpreting has been exclusively carried out from B to A language direction, the Russian-to-English subcorpus "consists of up to a third of samples from

[23] The part-of-speeches under investigation include verbs, nouns, adverbs, adjectives, conjunctions, appositions, and pronouns. (Dayter, 2018, p. 254)

interpreters working into their B language" (p.257), which may be accountable for the contradictory findings. Besides, Dayter (2018, p. 257) also finds that simplification and explicitation are closely related, so she warns against "overinterpretation of shallow statistical indicators", arguing that the best way is "to take into account a range of variables from different language angles, as suggested for a multivariate analysis of variation […] and to keep the conclusions grounded by frequent checks back to the level of discourse" (ibid.).

The contradictory findings reviewed here on the 'universal' features or general tendencies of interpreting indicate that, conclusions drawn from studies on translated language cannot be readily applied to interpreted language, given some of the intrinsically different constraints faced by simultaneous interpreters versus translators. Though ontologically being a piece of translation, interpreting (SI in particular) is constrained by multiple input and output variables such as language pair, working mode, working direction, delivery speed of source speech, accent, interpreting experience, etc.. that distinguishes it from translation of written mode as well as non-translated or unmediated spoken language.

Among these variables, there is one – working direction – that fails to attract enough academic attention in interpretese research. The bulk of research focuses on interpreting from B to A, which is considered the norm in the European market. However, this is not true in Asian market where A-to-B working direction is the fact of life (Lim, 2005; Setton, 2011), and few studies have actually taken into account the influence of working direction on the linguistic patterns of simultaneous interpreting, as implied in Dayter's (2018) study. So in the following section, the author will give a brief review on how working direction may influence the lexical patterns of translation and interpreting.

2.2.2.3 The influence of working direction on potential interpretese

The issue of directionality, that is translating into a mother togue (i.e., B-to-A, native or 'passive' interpreting) or into a B language (i.e., 'retour' or 'active' interpreting), has long been a controversial topic in interpreting studies, especially in terms of interpreting

quality (Gile, 2005; Seleskovitch, 1999; Seleskovitch & Lederer, 1989) and interpreter training (Déjean Le Féal, 2005; Donovan, 2005; Iglesias Fernández, 2005; Lim, 2005; Padilla, 2005). Proponents of native interpreting (Bros-Brann, 1976; Donovan, 2003; Herbert, 1952; Seleskovitch, 1978, 1999; Seleskovitch & Lederer, 1989), mostly known as the Paris School, argue that "true interpretation […] can occur only into one's 'A' language" (Bros-Brann, 1976, p. 17), given the inherent difficulties experienced by simultaneous interpreters, such as the "dual listening process" and "the risk of linguistic interference" which is found to be more pronounced in retour interpreting (Déjean Le Féal, 2005, p. 170). Advocates of retour interpreting (Denissenko, 1989; Iglesias Fernández, 2005), mostly from the Russian School, however, defend that comprehension is most conducive to better production. This dichotomy aside, nowadays there have been many voices advocating a more balanced view on the issue of directionality, as many empirical studies (e.g., Seel, 2005; Tommola & Helevä, 1998) reveal far less disparities between into A and into B interpreting as claimed by theoreticians.

Scholars have reported mixed findings in perception studies carried out through questionnaires or interviews, with some interpreting practitioners favor more into A interpreting (Bartłomiejczyk, 2004; Chang & Schallert, 2007; Donovan, 2004; Martin, 2005; Nicodemus & Emmorey, 2013) while others prefer retour interpreting (Al-Salman & Al-Khanji, 2002; Lim, 2003, 2005; Nicodemus & Emmorey, 2013). Chang and Schallert (2007), for example, report that the majority of Chinese interpreters feel more stressed and less flexible when doing A-to-B (Chinese-English) interpreting, so they tend to use meaning-based strategy by generalizing, paraphrasing, or even omitting information that seems redundant to them, in order to guarantee their interpreting quality (see also e.g., Bartłomiejczyk, 2004, 2006; Jänis, 2002). This kind of strategic processing is particularly inspiring for the present research as different strategies adopted in into A and into B interpreting may ultimately lead to different linguistic manifestations of the interpreting output, or *interpretese*. Wu (2001, p. 84) explicitly points this out in terms of the summarizing skill of Chinese interpreters working into B, "when interpreting from a high-context and implicit source language like Mandarin into a low-context and explicit

target language like English, more words and longer delivery times are required". It seems that both working direction and language pair may play a role in the final surface manifestations of interpreting (see also Bartłomiejczyk, 2004, 2006).

Regardless of the mixed pictures in both theoretical assumptions and empirical studies, it is an undeniable fact that nowadays there is growing market demands for retour interpreting, even in the into-A-dominant European market. In contrast with this changing landscape of the interpreting market, there have been very few studies (Dayter, 2018; Gumul, 2017; Tang, 2018) delving into the influence of working direction on the surface manifestations of interpreting based on real-life interpreting performances (only Dayter, 2018). Most empirical (often experimental) studies, as briefly mentioned above, focus on the influence of language directionality on interpreter's strategic choices. Interpreters working from B-to-A are found to resort more often to additions, inferencing, and transcoding, while interpreters working from A-to-B use omissions, summarizing strategy, and paraphrasing more frequently (e.g., Bartłomiejczyk, 2004, 2006; Chang & Schallert, 2007; Jänis, 2002; Tang, 2018).

Among the limited number of studies directly related to the linguistic patterns of interpreted language, Gumul (2017) investigates the role of working direction on explicitation of simultaneous interpreting, albeit experimental studies. Her study reveals that explicitation is much more frequent in retour interpreting (Polish-English) than native interpreting (English-Polish), especially in terms of "adding connectives, reiteration, meaning specification, and disambiguating metaphors" (p.320). Follow-up retrospection reveals that in most cases interpreters explicate "due to adopting repair or preventive strategies" (p.321), which can be associated with the intrinsic difficulties of the interpreting process. In her recent study, Gumul (2020) explores the relationship between explicitation and simultaneous interpreting, and concludes that simultaneous interpreters explicitate with an aim to either mask processing problems encountered during SI or to explicitate for clearer semantic relationships. In a similar vein, Defrancq et al. (2015) also propose two possible drivers for the addition of explicitation markers, i.e., connectives,

which include a. explicitation proper, aiming to make semantic relationships between clauses more explicit than in the source texts, and b. to cover up or fill up gaps originating in interpreters' omission.

Another study by Tang (2018) looks into the explicitation patterns in consecutive interpreting (CI) in both B-to-A (English-Chinese) and A-to-B (Chinese-English) language directions carried out by student interpreters and professionals as well, adopting Systemic Functional Grammar (Halliday & Matthiessen, 2004). Her detailed analysis illustrates that language direction can affect both student and professional interpreters' explicitation patterns. Overall, professionals in both working directions make more explicitations than student interpreters, indicating a strong correlation between experience and explicitation patterns, and they do so through restructuring and paraphrasing in A-to-B interpreting while addition in B-to-A interpreting. Tang (2018) also provides an explanatory framework for the explicitation patterns made by student and professional interpreters in their CI output. Instead of adopting a 'universal' feature view, she offers both a cognition-oriented explanation (such as time management and gap filling) and a pragmatic point of view (such as optimizing listeners' comprehension) of the interpreter. What we can infer from these findings is that interpreting, be it consecutive or simultaneous, may exhibit linguistic features closely related to the interpreter's decision-making process. Without taking into consideration their strategic choices, it may be hard to explain why interpreting exhibits certain linguistic features different from either spoken language or written translation.

The take-away message from this brief review is that working direction can indeed affect the lexical patterns (or distribution of lexico-grammatical features) of simultaneous interpreting given the different strategies adopted in handling different constraints in the two directions (A-to-B or B-to-A). The bulk of related studies on interpretese, however, focuses exclusively on B-to-A interpreting, thus singling out the possible influence of directionality. As vividly demonstrated in Dayter's (2018) corpus-based study, contradictory findings in terms of simplification, explicitation, and normalization patterns

may be partly attributed to the different working directions (English <> Russian). And given the status quo of the interpreting service in both Hong Kong and Mainland China where A-to-B (Chinese-English) interpreting is the mainstream (Setton, 2011), this study focuses especially on simultaneous interpreting into B.

2.3 Research gaps

It has been well acknowledged that corpus-based interpreting studies, although quickly gaining momentum thanks to the increasing availability of interpreting corpora, are still lagging behind compared with their counterpart, i.e., translation studies, due to the much more challenging task of corpus construction and compilation of spoken data (e.g., Setton, 2011; Shlesinger, 1998). Nevertheless, the existing literature sheds light on the nature of interpreting as a distinct variety of constrained language due to particularly its intrinsic constraints relating to time, linearity and (un)shared knowledge (Shlesinger, 1995). The process and very nature of interpreting, as been observed in previous studies, is so complicated as to defy any conclusive findings in terms of its explicitness, linguistic sophistication and/or its use of (un)conventional language. On the one hand, interpreted language resembles native spoken language, evidenced in Shlesinger (2008) and Shlesinger and Ordan (2012), while on the other hand, it is a distinct variety of spoken language because of its mediated and constrained nature. In her recent study, Kajzer-Wietrzny (2018) identifies shared features between interpreted language and non-native or L2 language as constrained language. Intermodal studies (Bernardini et al., 2016; Shlesinger & Ordan, 2012) also report similarities as well as differences between interpreting and translation as two modes of mediation. For one thing, interpreting is found to be more 'spoken' than translated, in that it resembles more native spoken language in relation to translation. For another, it can be viewed as "an extreme case of translation" (Shlesinger & Ordan, 2012, p. 54), since some of the translation-specific patterns (such as simplification) seem to be more salient in interpreting. Such seemingly contradictory results are brought about due to a combination of constraining factors, some of which turn out to be the gaps in interpretese studies.

The first gap lies in the lack of diversity of corpora utilized, and the nature of the corpora under discussion. Among these studies, a lion's share resorts to the European Parliament Interpreting Corpus (EPIC), and the newly expanded European Parliament Translation and Interpreting Corpus (EPTIC). It is undeniable that both EPIC and EPTIC offer invaluable resources for the investigation of the linguistic patterns of interpreting, and the research carried out based on these corpora has deepened our understanding of the complex nature of (simultaneous) interpreting as a form of constrained language variety. However, if general linguistic patterns of interpreting are to be sought for, more diversified corpora need to be included. There is also one pitfall regarding the pseudo-comparability between interpreted texts and native spoken texts within these corpora and others such as TIC and Europarl, as has recently been pointed out by Defrancq (2018). The two groups, namely interpreters and the Members of the Parliament, share the same "discourse community" of European Parliament, in which they tend to influence each other's linguistic output, thus leading to a linguistic phenomenon what Defrancq (2018, p. 119) calls "linguistic convergence, i.e., the levelling out of specific features of both types of output, making the whole search for features typical of interpreting ultimately pointless". This homogeneity of the language use between interpreters and MPs may also help explain the contradictory findings in interpretese studies.

Secondly, the lack of a systematic analytical framework is more prominent in studies on interpreted language (or interpretese studies) compared with studies on translational language. The majority of corpus-based interpreting studies aims to testify the hypothesized 'universal' features of translated language by replicating the research methodology of translation studies. Questions have been raised as to whether the selected linguistic parameters or indicators designed specifically for written translations are suitable for spoken discourse (see also Bernardini et al., 2016). Besides, interpretese is usually approached from a unidimensional perspective, illustrated through frequency comparison among very few linguistic indicators, which often fails to reveal the hidden (linguistic) patterns of interpreted language in other dimensions. Among all the reviewed studies, Xiao's research (2015) turns out to be the only one applying a multidimensional

perspective to the analysis of linguistic patterns of simultaneous interpreting based on genre analysis. Different from hers, which focuses on the shift of the orality features of source texts and interpreted texts along the oral-literate continuum based on the different dimensions suggested by Biber (1988) and Shlesinger (1989), the present research aims to examine more general linguistic patterns specific to L2 interpreting, which brings out the other research gap, that is the lack of research attention on the influence of working direction on the linguistic manifestations of interpreting.

As explained in section 2.2.2.3, many interpreting theorists (e.g., Seleskovitch, 1978; Seleskovitch & Lederer, 1989) from the Paris School tend to avoid retour interpreting as they believe interpreting into a B language rather than a native language can add further difficulties to simultaneous interpreters, which may result in undesirable interpreting performance. Recently, however, many survey studies (e.g., Al-Salman & Al-Khanji, 2002; Lim, 2003, 2005) have demonstrated that for practicing interpreters, there are many other factors (such as language pair, preparedness, delivery speed and accent of the speaker) other than the factor of directionality that may pose more constraints during their interpreting process. Given the conflict between theory and practice, it seems intriguing to find out whether interpreting from A-to-B differs substantially compared with interpreting from B-to-A in terms of the linguistic patterns of the interpreting outputs. Besides, the issue of directionality is also closely related to the translation universals of "source language shining through effect", or "interference". The bulk of research focuses on the interference of B language on A language production. However, according to Dejean LeFéal (2005), it is often the more active language (i.e., A language) that affects the weaker. Therefore, "[t]he B language is […] more susceptible to interference than the A language" (Déjean Le Féal, 2005, p. 170).

Last but not least, as one of the genetically distinct language pairs, Cantonese/English, in comparison with Mandarin/English language pair, has received scare academic attention, despite the ready availability of the interpreting data provided by the Hong Kong LegCo, and language pair does seem to influence lexical patterns of SI as reported in many studies

(e.g., Ferraresi et al., 2018; Russo et al., 2006; Sandrelli & Bendazzoli, 2005). Ferraresi et al. (2018), for example, find that, compared with mediation mode (i.e., translation and interpreting), the influence of source languages involved "seems to be stronger than that of the former" (Ferraresi et al., 2018, p. 718). It is thus reasonable to see the possible linguistic manifestations of interpreted language carried out in another language pair which involves ontologically different languages.

In a nutshell, the present research strives to fill the afore-mentioned gaps by examining the lexical patterns of L2 interpreting involving a genetically different language pair from both unidimensional and multidimensional perspectives. The final goal, as in any research practice, is "to raise awareness about what interpreting is and what processes (linguistic, pragmatic, practical or cognitive) are engaged during an interpretation" (Cencini, 2002, p. 1).

Chapter 3 Data and methodology

Identification of overall linguistic patterns of interpreting is only possible with a large-scale machine-readable corpus. In this chapter, a new parallel, intermodal and (quasi-)comparable corpus named the LegCo+ is introduced for such a research purpose. Details for corpus construction will be explicitated, including the principles for corpus design, and the different steps taken (i.e., transcription, segmentation and annotation) for corpus compilation. After the introduction of the new corpus, two-phase data analyses, i.e., an initial unidimensional analysis and a follow-up multidimensional analysis utilizing a multivariate statistical technique on linguistic variation, will be expounded in detail.

3.1 Corpus linguistics as a research approach

Since the appearance of the first machine-readable corpus – the Brown Corpus – in the field of linguistic studies in the 1960s, corpus linguistics as a new research approach has been gathering momentum. However, it was not until the 1980s, thanks to the widespread availability of electronic corpora and computational tools, that corpus linguistics has established itself as a new research paradigm widely applied in linguistics, computational linguistics, foreign language teaching, lexicography, etc. (Biber & Reppen, 2015; K. B. Hu et al., 2007; Laviosa, 1998b; Laviosa-Braithwaite, 1996). The main reason for the popularity of corpus linguistics can be attributed to its distinctive characteristic that "it is possible to actually 'represent' a domain of language use with a corpus of texts, and possible to empirically describe linguistic patterns of use through analysis of that corpus" (Biber & Reppen, 2015, p. 1), which was almost impossible in the pre-corpus period.

The introduction of corpus linguistics to translation and interpreting studies (TIS), however, is rather late, since for a long time translated language had been regarded as a distorted language inferior to standard language use and worth no academic attention (Baker, 1993, 1996; K. B. Hu, 2012; K. B. Hu et al., 2007). Thankfully, the field of translation studies experienced a revolutionary change spearheaded by Even-Zohar (1978) and Gideon Toury (1980, 1995), shifting the focus of translation studies from

prescriptivism to descriptivism. The establishment of descriptive translation studies or DTS (Toury, 1995) as a sub-branch of translation studies has reshaped the center of translation research from a ST-oriented (source-text oriented) perspective to a TT-oriented (target-text oriented) view. Translations are no longer regarded as secondary and inferior to original texts, but rather they are the "facts" of the target system. Systematic description should be carried out to reveal the underlying norms of recurrent patterns for translational behaviour. The emergence of electronic corpora has made possible such description. In the meantime, there has been an upsurge of interest in the nature of translated language as a form of mediated and constrained communication (Baker, 1993), and the introduction of corpus linguistics to translation studies can help reveal systematically how translated language differ from unmediated target language. Bernardini (2015) has nicely summarized three factors as the catalysts of the marriage between CL and TS, or corpus-based translation studies (CTS or CBTS). In addition to the emergence of electronic corpora, as well as the shifting focus of translation studies from ST to TT, the quantitative focus offered by CL has also attracted the attention of translation scholars. The present research, likewise, relies heavily on corpus and corpus tools, which makes corpus linguistics the major research method adopted in this research. In the following sections, the author introduces a specialized corpus newly constructed, namely, the LegCo+ corpus, for the current research. Details about corpus design and compilation are provided in the following sections.

3.1.1 Corpus design

3.1.1.1 Principles

The overarching principle for corpus design and compilation, as argued by Bernardini et al. (2018, p. 13), "[….] ultimately depends on, and at the same time constrains, what it will be used for". In other words, one's research goal determines the way a corpus is designed and constructed. The general aim of this project is to explore linguistic variation patterns among simultaneous interpreting, native speech, and written translation in order

to isolate linguistic patterns specific to simultaneous interpreting as both spoken and mediated discourse (Shlesinger & Ordan, 2012). Special attention has been paid to interpreting into a B language, or L2/retour interpreting, the reasons of which have been expounded in Chapter two. Therefore, at least three components have been included to sort out features of L2 *interpretese*. In addition, source speeches in Cantonese are also included and will be referred to for the explanation of certain linguistic patterns identified.

In light of this aim, the LegCo+ corpus is designed to be parallel, intermodal, and comparable. A parallel corpus, as defined in Baker (1995, p. 230), "consists of original, source language-texts in language A and their translated versions in language B". In other words, a parallel corpus contains both source texts/speeches and their translations (in a generic sense). A comparable corpus is composed of "two separate collections of texts in the same language: one corpus consists of original texts in the language in question and the other consists of translations in that language from a given source language or languages" (ibid., p.234). That is, a comparable corpus consists of both mediated and unmediated texts/speeches in the same language. Besides, based on this definition, Baker (1995) does not take into account the issue of directionality, which can be attributed to the fact that translation into A/native language is the default translation direction. A special type of comparable corpus, namely, intermodal comparable corpora, has been proposed by Shlesinger (2008) as an extension of Baker's (1995) categorization. To put it simply, it refers to "corpora containing parallel or comparable outputs of translation and interpreting" (Bernardini et al., 2016, p. 62). However, such corpora are more challenging to construct "due to the shortage of texts that are both translated and interpreted in authentic settings" (ibid.). Thanks to a unique setting, i.e., "social-spatial contexts of interaction in which interpreting events take place" (Grbic, 2015, p. 371), in Hong Kong – the Legislative Council or HK LegCo – construction of such a corpus has been made possible.

3.1.1.2 A unique setting – The Legislative Council of Hong Kong

The overwhelming body of existing interpreting corpora, as reviewed in Chapter two, is

based on the proceedings of the parliamentary settings in the European Parliament. Marzocchi and Pöchhacker (2015, p. 298) have rightly pointed out the reasons behind, "[t]he legal or political requirements for openness of proceedings and the availability of audio streaming make parliamentary settings a welcome source of discourse data for corpus-based research". Such openness and availability are also true in terms of the meeting sessions of the Legislative Council of Hong Kong, or the HK LegCo, the unicameral legislature in Hong Kong where different political functions, such as approving, rejecting or modifying laws and regulations, scrutinizing budgets, and monitoring the work of the government, are realized. To ensure openness and transparency of governmental work, all the LegCo meeting sessions, including council meetings and committee meetings, are broadcast live, with sign language interpreting, simultaneous interpreting into both Mandarin Chinese (or Putonghua) and English provided for those with special needs and those who cannot understand Cantonese floor speeches, thereby making the Hong Kong LegCo a unique setting for the investigation of interpreting events. The proceedings of the Cantonese floor speeches are recorded verbatim in the Official Record of Proceedings of the Legislative Council (i.e., Hansard) as the floor version, and then translated into English and (Mandarin) Chinese versions separately[24].

Given the diversity of the LegCo proceedings, the author has randomly selected 16 meeting sessions from the council meetings serving for different political purposes, including *The Chief Executive's Questions and Answers sessions*, *Questions to the Secretaries*, and *Debates on motions and bills*, during the year period 2015-2017. The reasons for such a choice are as follows. First, the corpus under investigation is designed to be of more general issues rather than very specific ones. Among the various council and committee meetings, *The Chief Executive's Questions and Answers session* and *Questions to the Secretaries* often cover a wide range of topics that are of concern to the

[24] According to the official website of the Legislative Council of the Hong Kong Special Administrative Region of the People's Republic of China, https://www.legco.gov.hk/general/english/counmtg/cm1620.htm, "The proceedings of the meetings are recorded verbatim in the Official Record of Proceedings of the Legislative Council (Hansard). The records of proceedings of the Council are first presented in the original language as delivered by Members and officials at Council meetings (Floor version). They will then be translated into the English and Chinese versions separately."

general public. A decision was therefore made to include some of these proceedings. Second, as both *Debates* and *Q&As* are typical legislative or parliamentary discourses, and they may exhibit distinct linguistic patterns given their different situational purposes, the author is intrigued to find out whether there are linguistic variations among them within the same setting, i.e., the Hong Kong LegCo. Besides, despite the spoken nature of all the selected proceedings, *Debates* and *Q&As* differ both in terms of preparedness of speeches as well as the degree of interactions (monologue vs. dialogue). Based on the visual input, Members engaging in (mostly) monologue debates seem to be more prepared than those in dialogic Q&As, except for the written Q&A sessions in *Questions to the Secretaries* proceedings. Information about the interpreters' preparedness of the speeches, however, is not sufficient. This may further highlight the complexities of the potential patterns of linguistic variations. Detailed analysis needs to be done with respect to the correlation between the subgenres under discussion and their linguistic manifestations and, if possible, between preparedness of speech (both original and interpreted) and linguistic variation.

The LegCo proceedings aside, 27 video clips of unmediated native English speeches covering the same time span from another setting, i.e., the House of Commons in the UK Parliament, have been selected as a reference. The reason for such a choice over others (such as parliamentary speeches in the U.S. Congress or the European Parliament) is due to the well-known fact that Hong Kong was once colonized by the Great Britain for over one and a half centuries. During this long period time of colonization, LegCo was first established in 1843 under the Charter of British Colony, and it follows many of the traditions of the UK Parliament, irrespective of some changes it has undergone since Hong Kong's return to China back in 1997. Proceedings in the UK Parliament, therefore, are believed to be the most comparable ones compared to others. Being a bicameral legislative body, the UK Parliament has two legislative bodies, i.e., the House of Commons and the House of Lords, and it is the House of Commons that wields real power. Proceedings in the House of Commons are also diverse and serve different political functions, such as *Debates on motions and bills*, and *Oral Questions to the Prime Minister*

and Ministers. To achieve maximum comparability between the components of native speech and interpreted speech, decisions were made to randomly select meeting sessions of similar types (*Questions to the Prime Minister, Questions to the Ministers,* and *Debates on motions and bills*) and of a similar number of running words during the period 2015-2017. Therefore, the components in the LegCo+ corpus are comparable in terms of setting, corpus size, genres, time period, and also power relations among the speakers (or speaker roles), as shown in Table 3.1. However, the author has to point out that, despite the overall resemblance of the topics for the LegCo proceedings and parliamentary proceedings in the UK Parliament, they are essentially region-dependent, and the procedural languages between the two differ to some extent, which may lead to a compromise of the comparability between SI and NS. Nevertheless, they are the best possible resources the author can resort to, and they meet most of the criteria suggested by Baker (1995, p. 234), namely, they "should cover a similar domain, variety of language and time span, and be of comparable length".

Table 3.1 Outline of the LegCo+ corpus

	ST	**SI**	**WT**	**NS**
Corpus size	400,000 (characters)	235,156 (tokens)	301,292 (tokens)	228,174 (tokens)
Language	Cantonese	English	English	English
Mediation status	Unmediated (native)	Mediated (non-native)	Mediated (non-native)	Unmediated (native)
Total time length	28h5m	28h5m	N/A	22h49m
Genres	Genre A: *Questions and Answers to Prime Minster/Chief Executive* Genre B: *Questions to the Ministers/Secretaries*; and, Genre C: *Debates*			
Setting	The Hong Kong Legislative Council			The UK Parliament
Participants power relations	Chief Executive, President, Legislative Members (including Secretaries) representing different political parties			Prime Minister, Speaker, Parliamentary Members (including Ministers) representing different political parties
Time period	2015-2017			

Table 3.1 outlines the four components or subsets in the LegCo+ corpus, three of which are English components (i.e., SI, WT, and NS) and will be the focus of analysis in this

research, while the Cantonese component (i.e., ST) will be used mainly as a reference, and will be further annotated in a later research stage. In addition, the corpus size of each subset in Table 3.1 reveals further information. To begin with, the LegCo+ corpus is a million-size corpus, which is considerably large for an interpreting corpus, considering the intrinsic obstacles involved in corpus construction. Second, given the comparability of the corpus sizes of subsets SI and NS, and considering the shorter time length of NS compared with SI, it can be safely said that native English is generally delivered at a much higher speech rate than interpreted English. Third, since SI and WT are produced from the same source speeches, the smaller size of SI indicates that spoken language is less verbose (i.e., using less words) and perhaps more simplified than written language, which is unexpected given the well-evidenced redundant nature of spoken language (Chernov, 1994). But it may also demonstrate the unique nature of interpreting in relation to translation. Fourth, both SI and WT are bi-mediated at least since they are both translated (in a generic sense) and non-native/L2, which has complicated the comparison with native language use. More prominent linguistic patterns of SI in this project are expected, and great cautions will be taken when interpreting the results.

3.1.2 Corpus compilation

In their description of the first intermodal interpreting corpora, i.e., the European Parliament Translation and Interpreting Corpus (EPTIC), Bernardini et al. (2018) outline four basic steps, including transcribing the data, PoS tagging and lemmatization, alignment, and making the corpus ready for searching. Hu and Tao (2013), and Hu et al. (2016) introduce five steps for their construction of the Chinese-English Consecutive Interpreting Corpus, known as CECIC, which basically overlap with Bernardini et al. (2018), except that they include one additional step of digitalizing video and tape recordings, and one additional step of editing and word-segmenting the texts. In this study, based on the research objective as well as the easy accessibility of online video proceedings, the author focuses mainly on three steps, including data transcription, data segmentation and data annotation.

3.1.2.1 Data transcription

One of the major challenges for constructing spoken corpora, interpreting corpora included, is transcription, as it involves "a transition from a spoken mode to the written mode" and decisions have to be made regarding "the amount of detail" to be included in the corpus given the multi-modal nature of spoken data (Adolphs & Knight, 2010, p. 44). As argued by Shlesinger (1998, p. 2), "the difficulty lies not only in the act of transcription per se, but in the fact that certain elements of spoken communication are both so subtle and so subjective as to defy description." Cencini (2002, p. 2) also points out that "[a]ny transcription is (inevitably) a partial mirroring of an interaction, which cannot give an exhaustive representation of an event […] As a result, the feasibility of a study on interpreting depends on the features present in a transcription". Setton (2011, p. 52) proposes three possible ways for the transcription of simultaneous interpreting, including 1) synchronized interlinear transcription – with selected prosodic features (pauses, pitch or intensity stress etc.) and optional word-for-word gloss; 2) a parallel tabular presentation by aligned segments, either roughly time-aligned or matched by content; and 3) a 'fluent' and 'clean' transcript, punctuated and with speech errors and hesitations eliminated. Nevertheless, the choice of transcription systems finally depends on "the priorities researchers have and the solutions they must find to a series of problems" (Bernardini et al., 2018, p.24). In other words, the transcription system being adopted is tied to the research objective and the research questions to be addressed.

Adolphs and Knight (2010, p. 45) converge on this view, stating that "[t]he level of detail of transcription reflects the basic needs of the type of research that they are intended to inform". They suggest to "identify the spoken features of interest at the outset, and to tailor the focus of the transcription accordingly" (Adolphs & Knight, 2010, p. 44). As this research aims to examine the use of linguistic features in the three English components (i.e., SI, NS, WT), all the video recordings were transcribed orthographically with additional paralinguistic features such as repetitions, filled pauses ('ehm', 'er', 'ah', transcribed as "uh" in present study), false starts, repairs, which are included in the

transcription in attempt to stay as close as possible to the natural occurring communication. The decision to include the aforementioned paralinguistic features instead of prosodic features such as stress, tempo, rhythm, etc. was made out of two main considerations: the first one was that annotation of prosodic features is extremely time-consuming and requires more tailored tools such as Praat for the exploration of phonetic phenomenon, which is beyond the current research scope; the second consideration went to a reference to previous studies in which paralinguistic information mark-up are suggested to be included to "investigate differences between interpreting and written translation as well as features of interpreted language" (K. B. Hu, 2016, p. 200)

In the present research, as outlined in Table 1, three spoken components in the LegCo+ corpus need to be transcribed, i.e., Cantonese source speech (ST), English simultaneous interpreting (SI), and native English speech (NS). Necessarily, two transcription systems have been adopted for Cantonese and English respectively, with some shared transcription symbols for certain features such as filled pauses, false start, inaudible segment, intonation, and interruptions for the purpose of convenience. A number of transcribers were enrolled and trained for transcribing, including local undergraduate English majors (studying at the Hong Kong Polytechnic University) from Hong Kong, undergraduate English majors (studying at the Hong Kong Polytechnic University) from Mainland China, and several undergraduate students at Jinan University in Guangzhou. Transcription of the Cantonese source speech was done by the Hong Kong local undergraduate students majoring in English. Transcription of the English simultaneous interpreting was partially done by Mainland undergraduate students majoring in translation and interpreting at the Hong Kong Polytechnic University, and partially done by the author herself. Transcription of the English native speech was partly carried out by undergraduate students majoring in translation and interpreting at Jinan University in Guangzhou, and partly by the author. All the transcriptions have been proofread by the author for at least two times to ensure transcription accuracy.

Initially, the transcription system followed Tang's (2014) study (see Appendix 1). In a

later stage, however, certain modifications were made to cater to the research objective. Table 3.2 provides the revised version of transcription codes adopted in the transcription of the English components (SI and NS), and it should be noted that some transcription symbols are also applicable to the transcription of Cantonese source speech.

Table 3.2 Transcription codes for the spoken English components (SI and NS) and partly for the Cantonese components (ST)

Notation	Meaning	Example
<spelling*->	Truncated words, false starts, and mispronunciations	[…] and why the **<pan-democra*->** pan-democrats are so happy? […] so I have not **<mismatregi*->** misrepresenting your ideas.
<uh>	Stammer or hesitations as filled pauses	**<uh>** The election committee can vote **<uh>** **<uh>** can vote two at the same time.
<rep/ spelling /rep>	Self-repairs	[…] but you **<rep/ can /rep>** are able to come up with a better option […]
<inaudible>**	Inaudible segment	And if they've built up enough experience, they will be more **<**inaudible>** with the procedures […]
<sound>	Indicating non-verbal events heard on the tape, such as page flipping sound, noting, coughing	So on the one the twenty-seventh of August, **<page-flipping>** <uh> **<page-flipping>** Hong Kong Macao Office thought that some decision should be made […]
…	Indicating interruptions	I'm not saying that. I'm not **…** Well please refrain from making a debate and Ms Mo, please be seated.
.	Intonation symbol for full stop	Well, I think we have to draw this Q&A session to a close**.**
?	Intonation symbol for questions and doubts	Would you agree what you are doing is bringing shame to us and to the country as a whole**?**
!	Intonation symbol for exclamations and strong emotions	Mr CHU, you talked about the Queen's Pier ten years ago. How time flies!

In this revised transcription system, as compared to the original one attached in Appendix 1, additional transcription codes have been added, such as interruptions symbolled by "…" (in the ST version, it's "……" under the pinyin writing system), inaudible segments

without the specific time frame, self-corrections (or repairs) indicated by "<rep/ spelling /rep>". These changes aside, decisions were also made to spell out all the numbers, abbreviations, repetitions, etc. in the way they were pronounced. By so doing, the author hopes to deliver as close as possible the actual speech production to enhance the accuracy of the proportion and distribution of linguistic features.

All the transcriptions were initially done in .xlsx format, with Cantonese source speech, Mandarin Chinese translation, English interpretation, and English translation aligned paragraph by paragraph. They were then converted to .doc format to check the possible transcription inaccuracies such as basic spelling mistakes, wrong annotations of paralinguistic features, and so on. Speaker information (e.g., "PRESIDENT (In Cantonese):") at the beginning of the actual utterances was also deleted to further improve accuracy for later calculation of running tokens. After all initial work had been done, all the transcribed full texts were converted to .txt version for segmentation and further coding.

3.1.2.2 Data segmentation

In all, the LegCo setting comprises 16 full texts for Cantonese source speech, 16 for English simultaneous interpreting, 16 for English translation, covering three subgenres (*CE Q&A*, *Questions to the Secretaries*, with both written questions and answers as well as impromptu follow-up questions included, and *Debates on motions and bills*), while the UK Parliamentary setting includes 27 full texts of English native speech and three corresponding subgenres (*Oral Questions to the Prime Minister*, *Oral Questions to the Ministers,* and *Debates on motions and bills*). As the duration of legislative sessions in the LegCo setting varies from one and a half to three hours, while that of parliamentary sessions in the UK Parliament varies from half an hour to two hours, the number of running words for each session is extremely unequal. Therefore, it was decided to segment each full text (English components only) into several parts with the number of running words ranging from 1,300 to 2,000. In their study on register and features of

translated language, Kruger and Van Rooy (2012)[25] also combined shorter texts into longer text units to avoid analysis problems associated with very short texts. Likewise, the same approach was adopted in Hu et al. (2016)[26]. The current research follows a similar idea. In total, 477 text segments (138 for NS, 149 for SI, and 189 for WT) were collected. Since data segmentation is based mainly on running tokens, one segment usually contains speeches of more than one speaker (or interpreter), which may help flatten out the influence of idiosyncrasies of different speakers.

Besides text segmentation, corresponding subgenres in NS, SI, and WT were also codified under the three broad genre categories presented in Table 3.1 to facilitate data analysis (see Table 3.3.). A-C in NS represent *Oral Questions to the Prime Minister* (A), *Oral Questions to the Ministers* (B), and *Debates* (C). A-C in SI represent *Chief Executive's Questions and Answers* (A), *Questions to the Secretaries* (B), and *Debates* (C). While A-C in WT stand for *Chief Executive's Questions and Answers* (A), *Questions to the Secretaries* (B), and *Debates* (C). The same coding "A", "B", and "C" is used to indicate the comparability of the corresponding subgenres in the three language varieties. Given the different discourse functions represented by these corresponding subgenres, linguistic variations are expected. In the following data analysis in Chapter four, genre variation as well as genre influence will be examined.

Table 3.3 Information about data segmentation and coding of text genres

English components	**Sub-genres (number of text segments)**			**Total**
NS	NS_A (49)	NS_B (29)	NS_C (60)	138
SI	SI_A (51)	SI_B (27)	SI_C (71)	149
WT	WT_A (65)	WT_B (30)	WT_C (95)	190
TOTAL				477

[25] The corpus text length in their study ranges from about 50 words to about 20,000 words.
[26] In Hu et al. (2016, p. 6), they acknowledge that translated texts in the News genre differ significantly both in length and subject, with some very short texts, so they have to combine them into texts of about 2,000 words.

3.1.2.3 Data annotation

For many corpus linguists, once transcription is done, mark-up (Randi Reppen, 2010) or coding (Adolphs & Knight, 2010) is added to enrich the information presented in the corpus. Reppen (2010) distinguishes two types of mark-up for spoken corpora, i.e., document mark-up and annotations. Document mark-up refers to standard marking such as SGML or XML, as well as header information at the basic level. Annotations, in comparison, cover a wider range, but mostly include part-of-speech tagging (PoS tagging), which assigns each word its grammatical categories (e.g., nouns, adjectives, pronouns, etc.). According to Hu (2016), annotations describe "the nature or properties of the texts in the corpus", thereby adding "an extra layer of information, which can be counted, sorted, and compared" (Hyland, 2015, p. 301). In this study, the author does not draw a strict distinction between mark-up, coding and annotation, but will refer to "annotation" as a cover term.

In terms of annotation for a translation and interpreting corpus, researchers often focus on two types: metadata annotation, or header information, and linguistic annotation. Metadata annotation provides information about the data per se, i.e., "data about data" (Adolphs & Knight, 2010, p. 42). Burnard (2005) argues that "without metadata the investigator has nothing but disconnected words of unknowable provenance or authenticity" (p.31). For an interpreting corpus, metadata often cover information related to source speaker (such as name, gender, mother tongue, speaker role), communication event (such as duration, text length, delivery rate, mode of delivery), and the interpreter (such as gender, working experience, mother tongue, delivery rate, preparedness, and Ear-Voice span). For the LegCo+ corpus under investigation, metadata about the Cantonese source speakers in the ST subcorpus and the English native speakers in the NS subcorpus, including name, gender, political function, speaker's role, mother tongue, are added. As to metadata of the communicative events, only partial information is available, such as duration, text length, and delivery rate. While mode of delivery (i.e., impromptu, mixed, and prepared) is an important parameter, and is often found to exert an influence on the

linguistic patterns of the interpreting output (e.g., Dayter, 2018; Kajzer-Wietrzny, 2012), it can only be deduced from the visual input, or the video recordings, as done by Sandrelli and Bendazzoli (2005) and Bernardini et al. (2016) for the EPIC and EPTIC corpus, and the annotation may not be reliable. Therefore, annotation of mode of delivery will be referred to with great caution when certain linguistic patterns are to be accounted for. As far as interpreters and translators are concerned, they belong to two different institutions. While interpreters are recruited from the Official Languages Division (i.e., OLD) at the Civil Service Bureau, with some being in-house civil servants while others freelancers, translators are in-house staffs at the Translation and Interpretation Division from the Hong Kong LegCo, responsible for the translation of documentary works such as the Policy Addresses of the Chief Executive. Therefore, translation and simultaneous interpretation are often carried out independently. However, other metadata regarding some of the details of translators and interpreters, such as their age, working experience, working mode, their working conditions are not readily available.

Automatic linguistic annotation usually generates two sets of tagging, i.e., PoS tags and lemmas, accompanying the actual transcripts (Bernardini, Collard, et al., 2018). A part-of-speech tagged corpus is a valuable resource and can contribute to our understanding of the nature of translated texts by allowing researchers to carry out detailed analyses of the distributional patterns of the morphosyntactic categories. A number of PoS taggers is available, especially for English, such as Treetagger (Schmid, 1994), Stanford Tagger (Toutanova et al., 2003), TagAnt (Anthony, 2015), and some web-based programs such as Wmatrix (Rayson, 2009). In this research, Nini's (2014) Multidimensional Analysis Tagger (MAT) was utilized for tagging both part-of-speech of the three English components (i.e., SI, WT, and NS), and the 67 linguistic features analyzed in Biber's (1988) model. The MAT is a program designed to replicate "Biber's (1988) tagger for the multidimensional functional analysis of English texts" (Multidimensional Analysis Tagger (v.1.2) – Manual, p.1). It also offers an option for the researcher to fully tag their corpus data through the Stanford Tagger (2013) included in this program. Altogether, the Multidimensional Analysis Tagger generates two tagged files, one is the fully tagged

version based on the Penn Treebank tagset, and the other is the tagged version of Biber's (1988) 67 lexico-grammatical features. It should be noted that all the listed taggers are originally designed for written texts, so "the expected performance of taggers and lemmatizers on spoken corpora is likely to be much worse than in written corpus projects" (Bernardini, Collard, et al., 2018, p. 32), since the spoken features cannot be readily identified and annotated and thus jeopardize the tagging accuracy. To minimize that risk, the author temporarily removed the paralinguistic annotations as illustrated in section 3.1.2.1 to improve the tagging accuracy. As for the Cantonese source speeches, currently they are not PoS tagged due to the lack of ready-made tagging software. Besides, unlike the EPTIC corpus, the LegCo+ corpus has not been lemmatized at the moment, as lemmatization is not the concern of the current research. Further efforts will be made in this regard.

3.2 The unidimensional approach to linguistic variation

The unidimensional or univariate approach can be said to be the dominating method for data analysis in the studies of the nature of mediated language, especially of interpreted language, as reviewed in Chapter two. It is often illustrated through frequency comparison among very few linguistic indicators between texts of different mediation status (mediated vs. unmediated). Once a statistical difference has been identified in terms of the selected linguistic indicators, the mediated texts are then claimed to be different from the unmediated texts, exhibiting certain lexical patterns known as 'translation universals'. Although this unidimensional perspective, as critiqued in Chapter two, fails to unveil the hidden lexical patterns in other dimensions in which texts of different mediation status may share similarities or demonstrate differences, it can provide a general picture about the overall variation patterns of the linguistic features in question, and promising, albeit mixed, results have also been reported in previous studies revealing the mediated nature of interpreting (Bernardini et al., 2016; K. B. Hu & Tao, 2013; Kajzer-Wietrzny, 2012, 2015; Russo et al., 2006; Sandrelli & Bendazzoli, 2005). To make a direct comparison with previous studies, the author decided to opt for a unidimensional analysis first before

delving into the multidimensional approach. However, in contrast to traditional unidimensional analyses, in which a certain 'universal' hypothesis (e.g., simplification) is operationalized with very few linguistic indicators (e.g., lexical density, list head coverage, high frequency words) to be confirmed/refuted, this research makes no presumption and takes into account as many as 79 linguistic features in order to see their general variation patterns among texts of different mediation status or different language varieties (i.e., NS, SI, and WT). The decision to include as many as 79 linguistic features is made with a view to the subsequent multidimensional analysis, the details of which (including the selection of the 79 linguistic features) will be elaborated in section 3.3.

Before the univariate analysis, which is realized through statistical tests, a number of tests need to be done to assess the suitability of the statistical methods to be used. First, tests of normality are performed across the three language varieties to see if these 79 linguistic features are normally distributed. Very few studies, to the best knowledge of the present author, would carry out normality tests before data analysis, and often they would choose the non-parametric Mann-Whitney U test for statistical significance test as the data under discussion (i.e., frequencies of linguistic features) were often categorical. The author believes it is more scientifically adequate if normality tests are carried out first before a decision is made with respect to the choice of tests for statistical significance. It was therefore decided that both statistical (i.e., the Kolmogorov-Smirnov with Liffiefors significance correction and Shapiro-Wilk test) and visual tests (i.e., the Quartile-Quartile plot) of normality would be carried out for NS, SI, and WT (see Appendix 2). The null hypothesis is that the population is normally distributed. If p value is smaller than 0.05, then it is highly likely that the population is not normally distributed. The results (see Appendix 3) show that 21 out of 79 linguistic features in the NS subset, 14 out of 79 in the SI subset, and 10 out of 79 in the WT subset, follow normal distribution. When examined together (i.e., tests of normality for both NS and SI as one dataset, and for SI and WT as one dataset), however, only 7 features (i.e., VPRT, BEMA, PRIV, SPAU, TOP10, LD, and SW) are normally distributed in the dataset of NS and SI, and 6 (i.e., VPRT, BEMA, DT, IN, LD, and SW) in SI and WT dataset. An example of both statistical

and visual tests for normality in the dataset of NS is given in Table 3.4 and Figure 4.1, which shows average word length (AWL) is normally distributed in NS, while type-token ratio (TTR) and amplifiers (AMP) are non-normally distributed.

Table 3.4 Tests of normality for AWL, TTR, and AMP in the NS dataset

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AWL	.073	138	.073	.986	138	.195
TTR	.117	138	.000	.894	138	.000
AMP	.081	138	.027	.963	138	.001

a. Liffiefors Significance Correction

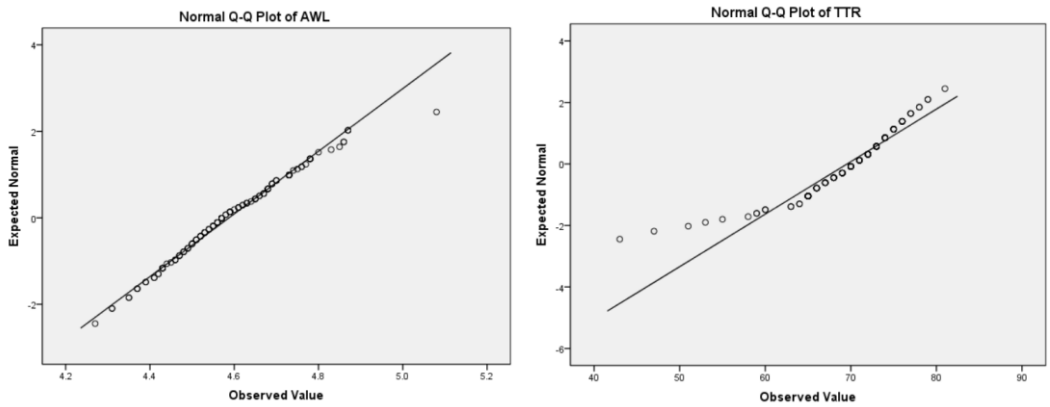

Figure 3.1 Normal quartile-quartile plot (QQ plot) of AWL and TTR

Besides tests of normality, tests of homogeneity of variances (i.e., Levene's test) were also carried out to assess whether the variance among these normally distributed linguistic features in the datasets is approximately equal, another necessary assumption for parametric test, such as ANOVA. The null hypothesis is the population/dataset variance are equal (also called homoscedasticity). A p value less than 0.05 indicates statistically significant difference between the variances in the population/dataset. Table 3.5 and 3.6 indicates that only 3 linguistic features (PRIV, SPAU, and LD) show no statistically significant variance in the NS/SI dataset, and 2 linguistic features (DT and LD) in the SI/WT dataset. Given the small number of linguistic features which follow normal distribution, it was decided that all 79 linguistic features will be compared using non-parametric Mann-Whitney U test to indicate the statistical difference between NS and SI

(comparable comparison), and between SI and WT (intermodal comparison). Results of the statistical tests are reported in Chapter four.

Table 3.5 Test of homogeneity of variances for the NS/SI dataset

	Levene statistic	df1	df2	Sig.
VPRT	16.405	1	285	.000
BEMA	25.049	1	285	.000
PRIV	2.059	1	285	.152
SPAU	.169	1	285	.681
TOP10	4.489	1	285	.035
LD	.056	1	285	.014
SW	22.402	1	285	.000

Table 3.6 Test of homogeneity of variances for the SI/WT dataset

	Levene statistic	df1	df2	Sig.
VPRT	11.878	1	337	.001
BEMA	20.469	1	337	.000
DT	.448	1	337	.504
IN	8.613	1	337	.004
LD	1.055	1	337	.305
SW	7.180	1	337	.008

The univariate analysis, as explained above, aims to offer an overall picture regarding how SI shares similarities with or demonstrates difference from NS and WT, but it cannot tell us if the statistically distinctive features are correlated with each other. In other words, it fails to unveil the following questions: Do the identified statistically distinctive features co-occur systematically to realize a certain function as interpreted by translation scholars, such as making explicit the message, making the target text more standardized and acceptable? Is there only one single dimension, as represented all together by the identified linguistic features, in which SI differs from the other language varieties (i.e., NS and WT)? How should we explain the similarities among the three language varieties in which case certain linguistic features show no statistical significance? etc.. A multivariate or multi-dimensional analysis is appropriate to address such questions.

3.3 The multidimensional approach to linguistic variation

The multidimensional approach or multi-feature approach (also known as the MD/MF approach) was first introduced in Biber (1986) and then developed more fully in Biber

(1988) in his analysis of linguistic variation between spoken and written registers. The idea for a multidimensional exploration is based on the belief that a univariate dimension fails to uncover the underlying dimensions discriminating spoken and written registers, since "[t]he communicative possibilities offered by a language are complex, and there is no reason to expect a single dimension to be the central discriminator among all text types" (Biber, 1986, p. 385). "Multidimensional" in Biber's (1992c, p. 132) view assumes that "multiple parameters of variation will be operative in any discourse domain", which suits a study on translation and interpreting well as both activities are governed by multiple constraints, and a unidimensional perspective can only present limited variations among different language varieties.

Previous studies on linguistic features of mediated language reviewed in Chapter two have revealed that currently there is a "lack of a unified analytic model, which has caused much confusion in discussion of linguistic features at various levels, in different genres, or in different languages" (X. Hu et al., 2016, p. 4). The "unified analytic model" in X. Hu et al. (2016) refers to a systematic methodology adopted, as many studies draw hasty conclusions about the 'universal' features or tendencies of translated/interpreted language based on a small number of linguistic features, which seems very problematic since "individual features hardly ever function in terms of a single property. It is much more likely to assume that one feature contributes to several properties" (Evert & Neumann, 2017, p. 2). Later studies (e.g., Dayter, 2018; H. Kruger & Van Rooy, 2016a) on the four-fold translation universals (Baker, 1993) also reveal that these TUs are closely related to each other. To avoid an overinterpretation of shallow statistics, as highlighted by Dayter (2018, p. 257), there is a need "to design studies that take into account a range of variables from different language levels, as suggested for a multivariate analysis of variation […]; and to keep the conclusions grounded by frequent checks back to the level of discourse". The present study aims to take that step towards a multivariate analysis to reveal the underlying dimensions which univariate analyses fail to uncover. Before looking into the specific steps taken for the multidimensional analysis in this research, some of the key features of the MD approach are introduced, and their indications for the present study

discussed.

3.3.1 Key features of the MD approach

In his methodological overview, Biber (1992c, p. 332) summarizes eight general characteristics of the multi-dimensional (MD) approach to genre variation. Since some of these characteristics overlap with those of corpus-based studies (such as "corpus-based", "computer-based"), the author only outlines some of the key features that stand out from other corpus-based approaches adopted in studies on distinctive features of translation and interpreting.

To begin with, the MD approach is explicitly featured by its multi-dimensional perspective, instead of a unidimensional one as often assumed by many corpus-based studies. As translation and interpreting have been increasingly acknowledged as multifaceted activities, a multi-dimensional perspective is conducive to uncovering the underlying constraints of translational activities. The key to a multi-dimensional analysis, as will be explained later, is the use of multivariate statistical techniques such as cluster analysis and factor analysis, which helps identify the underlying constructs based on a number of quantifiable variables.

Secondly, the MD approach adopts what Biber (1992c) calls "variationist and comparative perspectives", given that "different kinds of text differ linguistically and functionally, so that analysis of any one or two text varieties is not adequate for conclusions concerning a discourse domain (e.g., speech and writing in English)" (p.332). This has important implications for studies on 'universal' features of translation, whose research tradition is to compare translated texts (as one text variety) with non-translated target originals (as another text variety). Adapting variationist and comparative perspectives would save us from making hasty generalizations.

Thirdly, the MD approach combines quantitative analysis with functional interpretations. This approach is quantitative in nature, thanks to the power of multivariate techniques

which provides a large number of quantitative data. It is interpreted in functional terms, based on the assumption that "statistical linguistic co-occurrence patterns reflect underlying shared communicative functions" (ibid.). X. Hu et al. (2016, p. 28) have successfully identified a "translational" dimension with such functional characteristics as "reduced information load", "overrepresentation of the most frequent words", "less preference for reduced forms", "overrepresentation of function words", "extension of sentences and paragraphs", "overrepresentation of relative structures and markers of logical cohesion", and "underrepresentation of some particular items". It seems plausible that this combination of quantitative analysis with functional interpretation may also shed light on the specific lexical patterns of simultaneous interpreting.

Besides, the MD approach "synthesizes macroscopic and microscopic approaches" (ibid.). The macroscopic approach investigates "the overall parameters of linguistic variation" based on the analysis of the overall distribution of linguistic features across texts and genres, while the microscopic approach looks into the distribution of linguistic features in individual texts. Applied to the study of interpretese, this may offer us a detailed picture as regards the linguistic variation within subgenres among the three language varieties as described in section 3.1.2.2, as well as across the three varieties as a whole.

In a nutshell, the MD approach proposed by Biber (1986, 1988) can serve as a powerful analytical model for the identification of linguistic variation among interpreted texts, translated texts, and non-interpreted/unmediated originals. In the following sections, the author will outline the major steps taken for a multidimensional analysis.

3.3.2 Major steps taken for a multidimensional analysis

Biber (1988, pp. 63–64) outlines three basic methodological steps taken for the analysis of text variations. The preliminary step deals with the selection of the linguistic features to be investigated, the collection of texts, and the calculation of the frequencies of these linguistic features. This is followed by two quantitative steps: factor analysis and factor

scores calculation. The factor analysis, as briefly mentioned, can reduce a large number of variables into several latent factors by clustering these "linguistic features into groups of features that co-occur with a high frequency in texts" (Biber, 1988, p. 64). Through the analysis of the most widely shared functions of these co-occurring features constituting each factor, textual dimensions can be specified. The final step identifies factor scores with the operational representation of textual dimensions. By calculating factor scores for each text in each factor or dimension, we can compare genre variations across different text (segments). In the following sections, the specific steps taken for a multidimensional analysis of the linguistic features of interpreted language are outlined.

3.3.2.1 Selection, retrieval, and standardization of linguistic features

In Biber's (1988) study, 67 linguistic features are selected for comparison between written and spoken registers. These 67 linguistic features were chosen based on previous studies on the differences (and similarities) between spoken and written language. X. Hu et al. (2016), following Biber's (1988) advice to include as many linguistic features as possible, focus on 96 linguistic features, 67 of which replicate Biber's (1988) while the remaining are based on previous studies on translation universals. In this project, the author follows both Biber (1988) and X. Hu et al. (2016), and includes altogether 79 linguistic features (see Appendix 3). The added features are mainly based on previous studies on both spoken discourse and mediated discourse. Overall, the selected 79 linguistic features can be grouped into the following three types:

1. Biber's (1988, pp. 223–245) 67 linguistic features in 16 categories (A-P)
2. Textual features discussed in previous studies on translation and interpreting features (Q), including *standardized type-token ratio* (STTR), *average sentence length* (ASL), *top10 vocabulary coverage* (TOP10)[27], *lexical density* (LD),

[27] Previous studies on lexical patterns of interpreting, as reviewed in Chapter two, pay special attention to lexical simplification in interpreting, operationalized by four parameters as outlined in Laviosa (1998a), i.e. list head coverage (the first hundred words in the wordlist), lexical density (the proportion of content words to total running words), proportion of high frequency words to low frequency words (often operationalized as the most frequent 200 words), and the proportion of lemma (though seldom investigated in interpreting). As the text size of each text segment is between 1,300 to 2,000 tokens, it seems inappropriate to examine list head coverage and high frequency words in this case. It is therefore decided to investigate the ten most frequent words in the wordlist (TOP10), following X. Hu et al. (2016). The suggested parameters of lexical simplification will be used when comparing the three language varieties

shorter words (<= 3 letters) (SW), *longer words* (>=7 letters) (LW), and *coordinating conjunctions* (CC).

3. Other features which are believed to be different among the three language varieties, including *determiner 'the'* (DT), *preposition or subordinating conjunction* (IN), *possessive endings* (POS), *particles* (RP), and *WH-pronouns* (WP).

After the selection of the linguistic features to be investigated, a number of corpus tools are utilized to extract their frequencies automatically. Nini's (2014) Multidimensional Analysis Tagger (MAT) can not only annotate the 67 linguistic features, along with other part-of-speeches, but also provide the frequency data. The remaining features were extracted using WordSmith v.6 (Scott, 2012). As far as lexical density is concerned, the author follows Laviosa's (1998a, p.565) calculation method "by subtracting the number of function words in a text from the number of running words (which gives the number of lexical words) and then dividing the result by the number of running words" and is expressed as a percentage. Prior to statistical analysis, the raw frequencies were normalized for each text segment to frequency per 100 words. Besides, other descriptive statistics are also provided (see Appendix 4), though they may turn out to be less informational as the majority of the data is non-normally distributed, including the mean frequency, the minimum and maximum frequencies, the "range" or the difference between the minimum and maximum values, and the standard deviation.

3.3.2.2 Statistical analysis

The main statistical techniques used for a multi-dimensional analysis, in Biber's studies (1986, 1988, 1992a, 1992b, 1995), include factor analysis and cluster analysis. Factor analysis, as briefly described before, is a multivariate statistical technique used to reduce a large number of observed variables into a few unobserved ones called factors, or dimensions in functional terms. In other words, it aims to uncover the latent variables or

as a whole.

constructs based on the interdependence among the observed variables. For each factor, it "represents an area of high shared variance in the data, a grouping of linguistic features that co-occur with a high frequency" (Biber, 1988, p. 79). Since this study aims to uncover potential patterns of interpreted English that set it apart from unmediated, native English and translated English, factor analysis seems to be a good choice to examine whether latent dimensions exist among the three language varieties.

There are two types of factor analysis, i.e., exploratory factor analysis (EFA) and confirmatory factor analysis (CFA). As the names have suggested, EFA is exploratory in nature in that no prior theoretical assumptions have been made, so it is basically inductive, drawing conclusions based on data analysis. CFA, on the other hand, compares several existing theoretical models and confirms which one fits best based on the indicator of goodness-of-fit, so it is basically a "top-down" approach. Biber (1988, pp. 81–82) focuses on the first type as he argues that "the use of factor analysis in linguistics is usually exploratory (rather than confirmatory)" . While X. Hu et al. (2016, p. 11) claim to have combined both EFA and CFA "to see whether or not these features can be grouped to represent a 'translational' dimension", confirmatory factor analysis is nowhere to be seen. As the current research is also exploratory in nature, exploratory factor analysis was chosen.

In terms of exploratory factor analysis, a step-by-step procedure will be taken (see also Biber, 1988, pp. 81–91; X. Hu et al., 2016, p. 18), which include: a. assessing the suitability of the research data for factor analysis; b. choosing the method for factor extraction; c. deciding the number of factors to be extracted; d. choosing the method for factor rotation; e. extracting factor loadings for measured variables in each factor; and f. labelling and interpreting the extracted factors in dimensions.

In the first step, a pair of tests, i.e., Kaiser-Meier-Olkin Measure of Sampling Adequacy or KMO MSA, and Bartlett's Test of Sphericity, are performed to determine the suitability of the research data for factor analysis on the correlation matrix. The first test is a statistic that indicates the proportion of variance in the observed variables that might be caused

by underlying factors. A value less than 0.50 often indicates that a matrix is unacceptable for factoring. The second one tests the null hypothesis that the correlation matrix is an identity matrix, which indicates that the variables are unrelated and therefore unsuitable for further analysis. Ideally the chi-square score generated in Bartlett's Test should be statistically significant ($p<0.05$). Table 3.7 shows that the current research data (i.e., frequencies of 79 linguistic features) are suitable for factor analysis, with the KMO MAS score (.768) sitting between "middling" (.70's) and "meritorious" (.80's) based on Kaiser and Rice's (1974) criteria, and the p value ($p = .000$) for the Bartlett's test smaller than 0.001.

Table 3.7 Suitability test for factor analysis

KMO and Bartlett's Test

Kaiser-Meyer-Olkin Measure of Sampling Adequacy.		.789
Bartlett's Test of Sphericity	Approx. Chi-Square	21017.968
	df	3081
	Sig.	.000

In the next step, the method for factor extraction is determined. There are several methods available for factor extraction, each with its specific advantages and disadvantages, such as principal components analysis, unweighted least-squares method, maximum-likelihood method, principal axis factoring, alpha, and image factoring. Some of them work best for normally distributed data, such as maximum-likelihood method, while others for non-normally distributed data, such as principal axis factoring. Biber (1988) chooses principal factor analysis (PFA), namely principal axis factoring (PAF), for factor extraction, since "[t]his procedure extracts the maximum amount of shared variance among the variables for each other" (Biber, 1988, p. 82) while seeking for the least number of factors, and the solutions produced are found to be "more accurate and have been preferred in recent social science research" (ibid.). The present author decides to follow Biber (1988) using principal axis factoring to extract the factors.

Once the extraction method is decided, "the best number of factors in a solution must be determined" (Biber, 1988, p. 82). There are also several methods to reach this decision.

Biber (ibid.) advocates the use of a scree plot, "a plot of eigenvalues, which are direct indices of the amount of variance accounted for by each factor" and "will normally show a characteristic break indicating the point at which additional factors contribute little to the overall analysis". Other commonly used methods include Kaiser's (1960) eigenvalue-greater-than-one rule, and parallel analysis. In this research, a parallel analysis utilizing the parallel analysis engine (Donavan et al., 2007), facilitated by Cattell's (1966) scree plot will be used to determine the number of factors to be kept.

After factor extraction, the rotation of the factors must be done, because a lion's share of the observed linguistic features will load on the first factor, thus hiding the constructs underlying other factors. In a rotated solution, however, "each factor is characterized by the few linguistic features that are most representative of a particular amount of shared variance" (Biber, 1988, p. 84). There are two broad types of rotation method: orthogonal rotation in which the axis is maintained at 90 degrees and thus the factors to be extracted are uncorrelated to each other, represented by Varimax, and oblique rotation, which allows correlation among the underlying factors, represented by Promax. Biber (1988, p. 85) chooses the Promax method since "it permits minor correlations among the factors", while X. Hu et al. (2016, p. 19) select the Varimax method as it is "the most commonly used method". Given the different factors extracted must be interpreted in functional terms, the oblique rotation method would make it much difficult to interpret the factors as dimensions, as there are great chances these underlying dimensions may overlap with each other. To avoid this situation, the author decides to choose Varimax for the rotation.

The rotated factor matrix then shows the weight of each linguistic features loaded on each one of the factors, which is called factor loading. A factor loading "indicates the extent to which one can generalize from a factor to a particular linguistic feature, or the extent to which a given feature is representative of the dimension underlying a factor" (Biber, 1988, p. 85). There are both positive and negative factor loadings. A positive value indicates that the concerned linguistic features occur often, while the negative value suggests their absence. If the absolute value of the factor loading of a certain linguistic feature is larger than 0.30, then according to Biber (1988), it is regarded as statistically significant.

However, no consistent conclusion has been reached regarding the threshold for statistical significance of the absolute value of a factor loading (Fidell & Tabachnick, 2007; Field, 2005; Hair et al., 1998). Hair et al. (1998) suggests the significance cut-offs of a factor loading should be decided based on the sample size. The smaller the sample size is, the higher threshold for the factor loading. The suggested 0.30 cut-off is adequate for a sample size of 300. (ibid., p.112) As there are 477 text segments in this research, the suggested 0.30 cut-off is sufficient enough. The larger the absolute value of a factor loading of certain linguistic feature, the stronger "the co-occurrence relationship between the feature in question and the factor as a whole" (Biber, 1988, p. 85).

Finally, a factorial structure featuring several factors characterized by the co-occurrence patterns of several linguistic features with either positive or negative weights will be generated. Based on the underlying assumption that "strong co-occurrence patterns of linguistic features mark underlying functional dimensions" (Biber, 1988, p. 13), micro-analysis of the functional dimensions will be determined based on the factor or dimension scores, "operational representatives of the hypothesized dimensions", and "the similarities and differences among genres (the textual 'relations') can be analyzed with respect to these scores to support or refute hypothesized interpretations" (ibid., p.92). In this case, differences as well as similarities across and within the subsets of simultaneous interpreting, native speech, and written translation are expected to be seen along different dimensions. The results are reported in Chapter five.

Chapter 4 General linguistic patterns of L2 interpretese: A unidimensional analysis

In this chapter, as explained in Chapter three, following the methodology adopted in previous studies on the lexical patterns of translation and interpreting, the author has carried out a unidimensional analysis to investigate linguistic variation across three language varieties, i.e., native speech (NS), simultaneous interpreting (SI), and written translation (WT). The overarching goal is to offer a general picture of the statistically significant linguistic features of SI into a B language, or L2 *interpretese* as tentatively named in this research, compared to unmediated spoken discourse and written translation, based on nonparametric statistical tests. Two widely investigated universal hypotheses in corpus-based translation and interpreting studies, namely, lexical simplification and explicitation (or increased explicitness[28]), are zoomed in to see whether the lexical patterns identified in native interpreting apply to or are even more prominent in L2 interpreting, and whether such lexical patterns are consistent across subgenres in similar settings. All the statistical tests were performed utilizing IBM SPSS 20.

4.1 Overall linguistic patterns of L2 *interpretese*

As defined in Chapter one, L2 interpretese refers to the distinctive features of L2 interpreting with respect to unmediated native speech and L2 translation from the same source. Based on this definition, two pairs of comparison have been carried out from both interlingual comparable (SI vs. NS) and intermodal (SI vs. WT) perspectives. The former comparison aims to reveal distinctive features specific to L2 interpreting (non-native, mediated), while the latter attempts to uncover features specific to the mode of 'oral

[28] As reviewed in Chapter two, the two notions "explicitation" and "explicitness" do not refer exactly to the same thing. While "explicitation" has its roots in the "Explicitation Hypothesis" proposed by Blum-Kulka (1986), which is essentially process-oriented and involves comparison between source texts and translated texts, the notion of "explicitness" is "a property of lexicogrammatical or cohesive structures and configurations in one text" (Hansen-Schirra et al., 2007, p. 243). Chesterman (2004, 2017), in fact, refers to the explicitation hypothesis as a S-universal. Despite the conceptual differences, many existing studies have mixed the two, and often seek to verify the explicitation hypothesis without referring to the source texts. In this study, the author will, in most cases, stick to the notion of "explicitness", while bring up the concept of "explicitation" only when it is used in previous studies.

translation'. Non-parametric Mann-Whitney U tests were performed based on the normality tests carried out in Chapter three.

Based on the statistical tests, 54 out of the 79 linguistic features show statistically significant difference between SI and NS, with 23 overused[29] while the remaining 31 underused in SI compared with NS. On a general note, the results indicate that SI differs noticeably from NS in terms of the variation patterns of the 79 linguistic features under discussion (see Appendix 3), as shown in Table 4.1 and 4.2 (see Appendix 5 for a full overview).

Table 4.1 23 linguistic features overused in SI compared with NS.

Linguistic Features	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
CAUS	8771.500	18362.500	-2.154	.031
COND	5267.500	14858.500	-7.140	.000
CONJ	6743.500	16334.500	-5.047	.000
GER	7546.000	17137.000	-3.894	.000
HDG	8457.000	18048.000	-3.185	.001
NEMD	6666.000	16257.000	-5.149	.000
NOMZ	6356.000	15947.000	-5.587	.000
OSUB	8081.000	17672.000	-3.139	.002
POMD	8098.000	17689.000	-3.108	.002
SPP2	5583.000	15174.000	-6.688	.000
SYNE	8122.500	17713.500	-3.097	.002
THVC	7856.000	17447.000	-3.453	.000
XX0	4622.500	14213.500	-8.055	.000
BYPA	8660.000	18251.000	-2.347	.019
PASS	7771.000	17362.000	-3.573	.000
WHQU	8536.000	18127.000	-2.609	.009
CC	7739.500	17330.500	-3.618	.000
IN	7120.000	16711.000	-4.500	.000
RP	5314.500	14905.500	-7.082	.000
TOP10	8659.500	18250.500	-2.308	.021
DT	8687.500	18278.500	-2.268	.023
LD	8741.000	18332.000	-2.192	0.28
SW	5569.000	15160.000	-6.708	.000

Table 4.2 31 linguistic features underused in SI compared with NS.

[29] In this research, the terms "overused" and "underused" do not carry any value judgement. They are used to indicate features which are "more frequently used" or "less frequently used" in a certain language variety. The "overuse" or "underuse" of a certain linguistic feature were decided based on the mean rank reported in the Mann-Whitney U test.

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
TTR	6443.000	17618.000	-5.473	.000
AMP	2841.500	14016.500	-10.594	.000
DEMO	4316.000	15491.000	-8.492	.000
DPAR	7167.500	16758.500	-4.436	.000
EX	4691.500	14282.500	-7.959	.000
FPP1	4940.500	16115.500	-7.602	.000
JJ	5406.000	16581.000	-6.940	.000
PIN	5721.000	16896.000	-6.491	.000
PIT	8808.000	19983.000	-2.097	.036
PLACE	6110.500	17285.500	-5.945	.000
PRED	7262.500	18437.500	-4.297	.000
RB	8883.000	20058.000	-1.990	.047
THAC	4081.500	15256.500	-9.023	.000
TIME	7469.500	18644.500	-4.003	.000
TOBJ	2023.000	13198.000	-11.784	.000
TPP3	6395.500	17570.500	-5.531	.000
TSUB	3011.500	14186.500	-10.483	.000
BEMA	8026.500	19201.500	-3.209	.001
PASTP	9122.500	20297.500	-2.354	.019
PIRE	7116.500	18291.500	-5.118	.000
PRIV	5791.000	16966.000	-6.392	.000
SERE	2441.000	13616.000	-11.620	.000
SMP	9130.000	20305.000	-2.068	.039
THATD	7020.000	18195.000	-4.643	.000
WHOBJ	6942.000	18117.000	-5.976	.000
WHSUB	6569.000	17744.000	-5.298	.000
WZPRES	8782.000	19957.000	-2.135	.033
WP	4621.500	15796.500	-8.057	.000
STTR	4237.500	15412.500	-8.604	.000
ASL	3983.000	15158.000	-8.965	.000
LW	7816.500	18991.500	-3.508	.000

Table 4.1 and 4.2 show mixed pictures regarding the distribution patterns of statistically distinctive linguistic features in SI compared with NS. On the one hand, both Mann-Whitney U test and One-way ANOVA in Table 4.1 reported statistically distinctive features which suggest that SI is more lexically dense or informative than NS, characterized by an overuse of nominal features such as gerunds (GER, U=7546.000, p=.000), nominalizations (NOMZ, U=6356.000, p=.000*)*, by-passives (BYPA, U=8660.000, p=.019), agentless passives (PASS, U=7771.000, p=.000), and lexical density (LD, U=8741.000, p=.028). The overuse of passive structures, according to Hu

and Tao (2013)[30], can be interpreted as a tendency towards normalization, considering the different linguistic preferences for passive structures in Chinese (Cantonese included) and English; while high frequency of nominal features, argued by Kruger and Van Rooy (2016a, p. 40), may suggest an "increased referential explicitness as well as formality". On the other, the U test in both Table 4.1 and 4.2 reported distinctive features of SI associated with repetitive language use and lower level of lexical variety, such as more frequent use of top 10 vocabulary coverage (TOP10, U=8659.500, p=.021), and less frequent use of type-token ratio (TTR, U=6443.000, p=.000) and standardized type-token ratio (STTR, U=4237.500, p=.000).

In addition to linguistic variations at the lexical level, Table 4.2 reveals also syntactic variations between SI and NS, which has been under-explored in previous studies (except for the syntactic feature 'average sentence length'). The Mann-Whitney U test reported statistically underused features in SI indexing syntactic elaboration (Biber, 1988), such as 'that' relative clause on object position (TOBJ, U=2023.000, p=.000), 'that' relative clauses on subject position (TSUB, U=3011.500, p=.000), past participial clauses (PASTP, U=9122.000, p=.019), pied-piping relative clauses (PIRE, U=7116.500, p=.000), *WH* relative clauses on object position (WHOBJ, U=6942.000, p=.000), *WH* relative clauses on subject position (WHSUB, U=6569.000, p=.000), present participial *WHIZ* deletion relative (WZPRES, U=8782.000, p=.033), and average sentence length (ASL, U=3983.000, p=.000). In other words, in terms of syntactic elaboration and complexity realized through subordinating clauses, SI is syntactically simpler, less elaborated, and perhaps less informative than NS. Upon first reflection, this finding seems to contradict with previous analysis, which reports higher information density in interpreted texts, featured by an overuse of nominal features and higher lexical density. A closer examination may reflect the syntactically more simplified nature of interpreting as a form

[30] In Hu and Tao's (2013) study, they compared the frequency of passive structures in interpreted English, with that in original spoken English and translated English, based on the Chinese-English Conference Interpreting Corpus (CECIC). Their finding suggested that interpreted English uses more frequently passive structures than both original spoken English and translated English. They pointed out that "passive construction is far less frequent than active construction in the Chinese language, since its use is generally linked to something undesirable or unfortunate" (p.633), therefore an increased use of passive construction in interpreted English indicates "a remarkable tendency towards normalization" (ibid.).

of “constrained language” (H. Kruger & Van Rooy, 2016a), given its bi-mediation status (being translated and non-native) which adds further complexity to the bilingual cognitive processing. That said, it also provides counterevidence to the “complexity principle” argued by Rohdenburg (1996, p. 149), which states that “more explicit grammatical alternatives tend to be preferred in cognitively more complex environments”, and confirmed by Kruger and Van Rooy (2016a), who report an overuse of ‘that’ relative clause on object position (TOBJ), ‘that’ relative clause on subject position (TSUB), and demonstratives (DEMO) in two constrained language varieties (L2 writing and translated texts) as “a kind of cognitive ‘crutch’ to facilitate processing for the text producer as much as for the text receiver” (p.46). Their findings, however, are based on linguistic co-occurrence patterns rather than individual patterns of linguistic features, which are currently beyond the scope of this chapter focusing on a unidimensional analysis. Chapter five offers a detailed discussion.

Besides the mixed picture of the simplified and less elaborated trend of interpreted language, the Mann-Whitney U test in Table 4.1 and 4.2 also reported distinctive features of SI regarding the degree of explicitness. A number of linguistic features associated with increased explicitness were significantly overused (or underused) in SI, such as an overuse of causative adverbial subordinators (CAUS, U=8771.500, p=.031), conditional adverbial subordinators (COND, U=5267.500, p=.000), conjuncts (CONJ, U=6743.500, p=.000), other adverbial subordinators (OSUB, U=8081.000, p=.002), ‘that’ verb complements (THVC, U=7856.000, p=.001), coordinating conjunction (CC, U=7739.500, p=.000), and preposition or subordinating conjunction (IN, U=7120.000, p=.000), and an underuse of subordinator-that deletion (THATD, U=7020.000, p=.000). The use of subordinators and conjunctions often makes explicit the logical relations within the discourse (K. B. Hu & Tao, 2009, 2013; Van Rooy et al., 2010), while the use of full form ‘that’ verb complementizer has long been regarded as a strong indicator for increased syntactic explicitness (K. B. Hu & Tao, 2013; Kajzer-Wietrzny, 2012; H. Kruger, 2018; Olohan & Baker, 2000), though the reasons behind have yet to be fully disentangled.

There are also SI-specific distribution patterns in relation to NS, but these patterns cannot be readily summarized from the perspective of translation universals. For instance, the statistical tests reported an overuse of hedges (HDG, U=8457.000, p=.001), necessity modals (NEMD, U=6666.000, p=.000), possibility modals (POMD, U=8098.000, p=.002), second person pronouns (SPP2, U=5583.000, p=.000), and an underuse of amplifiers (AMP, U=2841.500, p=.000), discourse particles (DPAR, U=7167.500, p=.000), first person pronouns (FPP1, U=4940.500, p=.000), third person pronouns (TPP3, U=6395.500, p=.000), private verbs (PRIV, U=5791.000, p=.000), among others, in interpretations versus non-interpretations. It seems that interpreted language exhibit more features of uncertainty (e.g., hedges, possibility modals) and personal involvement (e.g., second person pronouns) than unmediated spoken language, while the latter is featured by linguistic features cueing stancetaking (e.g., private verbs) and discourse coherence (e.g., discourse particles). Kruger and Van Rooy (2016a) report significantly lower frequency of possibility modality in both translated and L2 (written) English, in relation to non-translated native English, suggesting "a general eschewal of the overt making of stance in favor of a more objective, informational style" in the two constrained language varieties (p.38). While the present author concurs to the idea that interpreters tend to avoid stancetaking expressions compared with native speakers, as will be testified in Chapter five, the overuse of possibility modals may suggest the interpreters' hedging strategies. Moreover, in terms of the usage patterns of personal pronouns in SI versus NS, it is also likely that source language shines through, since in the LegCo proceedings it is not uncommon that Legislative Members often address the primary speaker (such as the Chief Executive) directly using second person pronoun ("you"), while such way of addressing is often discouraged in the UK Parliamentary setting.

Statistical tests between SI and WT revealed that 58 out of 79 linguistic features showed statistically significant differences, with 27 overused and the remaining 31 underused in SI, as shown in Table 4.3 and 4.4 (see Appendix 5 for a full overview). In general terms, this also suggests that linguistic variation patterns of SI differ substantially from WT, which may point to intermodal (spoken vs. written) differences, as been tested in previous

studies (e.g., Bernardini et al., 2016; Ferraresi et al., 2018; Shlesinger, 2008; Shlesinger & Ordan, 2012).

Table 4.3 27 linguistic features overused in SI compared with WT

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
ANDC	2422.000	20567.000	-13.102	.000
CAUS	9239.000	27384.000	-5.522	.000
COND	10410.500	28555.500	-4.182	.000
DEMP	4673.500	22818.500	-10.588	.000
DPAR	2302.000	20447.000	-13.942	.000
EMPH	10145.000	28290.000	-4.478	.000
EX	4829.500	22974.500	-10.419	.000
FPP1	6417.500	24562.500	-8.640	.000
HDG	9235.500	27380.500	-8.100	.000
PIT	11286.500	29431.500	-3.203	.001
PRED	11182.000	29327.000	-3.320	.001
PRMD	10626.500	28771.500	-3.940	.000
SPP2	5589.500	23734.500	-9.579	.000
TO	10131.500	28276.500	-4.493	.000
TPP3	12047.500	30192.500	-2.353	.019
VPRT	4997.000	23142.000	-10.226	.000
XX0	8454.000	26599.000	-6.366	.000
BEMA	7870.500	26015.500	-7.017	.000
CONT	44.000	18189.000	-16.813	.000
PROD	6875.500	29403.500	-3.234	.001
PUBV	11258.500	29403.500	-8.216	.000
THATD	8042.500	26187.500	-6.835	.000
WHCL	801.500	27946.500	-5.005	.000
RP	10899.000	29044.000	-3.639	.000
WP	12217.000	30362.000	-2.165	.030
TOP10	11683.000	29638.000	-2.688	.007
SW	3670.000	21815.000	-11.707	.000

Table 4.4 31 linguistic features underused in SI compared with WT

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
AWL	4137.500	15312.500	-11.186	.000
TTR	10887.500	22062.500	-3.655	.000
CONC	10686.500	21861.500	-4.483	.000
CONJ	3410.000	14585.000	-12.002	.000
DT	9990.500	21165.500	-4.650	.000
DWNT	8677.500	19852.500	-6.134	.000
JJ	7258.500	18433.500	-7.700	.000
NN	7798.000	18973.000	-7.098	.000
NOMZ	9064.000	20239.000	-5.684	.000
PHC	8590.500	19765.500	-6.213	.000

PIN	6304.000	17479.000	-8.766	.000
SYNE	11973.500	23148.500	-2.446	.014
TIME	7846.500	19021.500	-7.045	.000
TOBJ	11733.500	22908.500	-2.730	.006
BYPA	12290.500	23465.500	-2.103	.036
PASTP	12091.000	23266.000	-3.145	.002
PEAS	8826.500	20001.500	-5.950	.000
PIRE	12117.500	23292.500	-2.758	.006
PRESP	6649.000	17824.000	-8.416	.000
SERE	5909.000	17084.000	-9.662	.000
SPAU	5464.500	16639.500	-9.705	.000
STPR	10820.500	28965.500	-3.845	.000
WHOBJ	11509.500	22684.500	-4.011	.000
WHSUB	12204.000	23379.000	-2.189	.029
WZPAST	5965.500	17140.500	-9.160	.000
WZPRES	11876.000	23051.000	-2.547	.011
POS	11907.500	23082.500	-2.511	.012
STTR	6674.500	17849.500	-8.304	.000
ASL	3037.000	14212.000	-12.382	.000
LD	6824.000	17999.000	-8.186	.000
LW	4211.500	155386.500	-11.103	.000

Compared with the above interlingual comparable comparison between SI and NS, the intermodal comparison between interpreting and translation reveals more clear-cut patterns as would be expected between speech and writing. The Mann-Whitney U test in Table 4.3 and 4.4 reported some linguistic features indexing the more simplified nature of SI with respect to informativeness, repetitiveness, and lexical sophistication. Compared to WT, SI is characterized by an overuse of top 10 vocabulary coverage (TOP10, U=11683.000, p=.007), shorter words (SW, U=3670.000, p=.000), and an underuse of standardized type-token ratio (STTR, U=6674.500, p=.000), average sentence length (ASL, U=4211.500, p=.000), lexical density (LD, U=6824.000, p=.000), longer words (LW, U=4211.500, p=.000), adjectives (JJ, U=7258.500, p=.000), total other nouns (NN, U=7798.000, p=.000), nominalizations (NOMZ, U=9064.000, p=.000), phrasal coordination (PHC, U=8590.500, p=.000), and by-passives (BYPA, U=12290.500, p=.036). In other words, the statistically distinctive features in SI seem to support the lexical simplification hypothesis from an intermodal perspective. Meanwhile, the U test also reported features suggesting syntactically less sophisticated and elaborated nature of SI, featured by an underuse of “dispreferred structures” (Biber, 1988) and

different types of clauses, such as pied-piping relative clauses (PIRE, U=12117.500, p=.006), sentence relatives (SERE, U=5909.000, p=.000), split auxiliaries (SPAU, U=5464.500, p=.000), stranded preposition (STPR, U=10820.500, p=.000), *WH* relative clauses on object position (WHOBJ, U=11509.50, p=.000), *WH* relative clauses on subject position (WHSUB, U=12204.000, p=.029), past participial *WHIZ* deletion relatives (WZPAST, U=5965.500, p=.000), present participial *WHIZ* deletion relatives (WZPRES, U=11876.000, p=.011), and others. Based on these variation patterns, it is encouraging to say that intermodally speaking, interpreting is both lexically and syntactically more simplified than translation.

In addition to the aforementioned linguistic patterns, the U test also reported some overused features in SI suggesting strongly the more informal and involved nature of interpreting as a form of spoken language (e.g., Chafe, 1982; Chafe & Tannen, 1987), such as demonstrative pronouns (DEMP, U=4673.500, p=.000), first person pronouns (FPP1, U=6417.500, p=.000), second person pronouns (SPP2, U=5589.500, p=.000), third person pronouns (TPP3, U=12047.500, p=.000), analytic negation (XX0, U=8454.000, p=.000), contractions (CONT, U=44.000, p=.000), pro-verb do (PROD, U=6875.500, p=.000), and *WH*-pronouns (WP, U=12217.000, p=.030). The highly frequent use of pronouns, “a marker of orality in target language” (Shlesinger & Ordan, 2012, p. 48), indicates strongly the contextualized nature of SI as spoken language, despite the fact that simultaneous interpreters are often physically isolated from the source speakers. By contrast, translated texts are decontextualized or autonomous, and they often reply on elaborated syntactic structures to make explicit the information to be conveyed (Chafe & Tannen, 1987; H. Kruger & Van Rooy, 2016a, 2016b; Quirk et al., 1985). This also has some implications for the lower level of explicitness in interpreting with respect to translation.

To sum up, the variation patterns of 79 linguistic features of the three language varieties reveal some contradicting, albeit correlated, lexical patterns discussed in previous studies. While SI is characterized by an overuse or underuse of linguistic features indicating more

simplified, explicit, and perhaps more conservative language use compared to unmediated spoken language, it is also featured by linguistic features associated with a lower degree of syntactic elaboration (or explicitness), and higher information density. A more refined multidimensional analysis may help reveal if and how these distinctive features may actually co-occur to realize a shared function of increased or decreased explicitness. Intermodally speaking, SI is characterized by more simplified, informal, while syntactically less explicit language use compared to WT. However, the author cautions that all interpretations of the shallow statistics generated by statistical tests are only exploratory rather than conclusive. Without knowing the actual co-occurrence patterns of these identified features (either overused or underused in SI), it is hard to decide if their occurrence is due to mere chance. Chapter five aims to shed light on this. In the following section, the author zooms in on two popular 'universal' features of mediated language, that is, lexical simplification and explicitation (or increased explicitness), based on several often-used linguistic indicators from previous studies to observe in detail linguistic variations both across and within (the subgenres of) the three language varieties. The overarching goal is to testify if the most acknowledged 'universal' features of native translation and interpreting are also applicable to, or even more prominent in L2 interpreting, particularly with respect to comparable comparison (SI vs. NS), and if and how different subgenres may influence these patterns.

4.2 Linguistic variation across and within three language varieties

4.2.1 Exploring lexical simplification from comparable and intermodal perspectives

4.2.1.1 Data analysis

Statistical tests carried out in section 4.1 have provided an initial glimpse over the widely accepted hypothesis that "translated texts are more simplified than non-translated target originals", except for the distribution patterns of linguistic features related to information density (such as nominal features, passives, etc.). To make a more direct comparison with previous studies from both comparable and intermodal perspectives, the author focuses

on and visualizes the variation patterns of four popular linguistic parameters, i.e., standardized type-token ratio (STTR) and top 10 vocabulary coverage (Top 10) as indicators for lexical variety and repetitiveness, lexical density (LD) as an indicator for informativeness, and average sentence length (ASL) as an indicator for syntactic sophistication, both across and within the three language varieties. Before the visualization, the Kruskal-Wallis H test was performed to see general variation patterns across the three varieties. The Kruskal-Wallis test extends the Mann-Whitney U test by allowing comparison between more than two groups, meaning that it is possible to compare across three varieties to get a broader picture about the variation patterns of linguistic features. The test results, however, are only meant to be a reference, as two out of the three groups (i.e., NS and WT) are not directly comparable.

Table 4.5 Kruskal-Wallis H test of simplification features across NS, SI and WT

Linguistic parameters	Language variety	N	Mean Rank	Chi-square	df	Asymp.Sig
STTR	NS	138	285.45	94.254	2	.000
	SI	149	148.26			
	WT	190	276.42			
TOP10	NS	138	230.45	8.711	2	.013
	SI	149	266.34			
	WT	190	223.77			
LD	NS	138	165.41	121.610	2	.000
	SI	149	200.13			
	WT	190	322.93			
ASL	NS	138	239.50	189.773	2	.000
	SI	149	122.34			
	WT	190	330.13			

The Kruskal-Wallis H test in Table 4.5 shows that there are statistically significant differences in terms of the use of STTR (X^2=94.254, p=.000), Top10 (X^2=8.711, p=.013), LD (X^2=121.610, p=.000), and ASL (X^2=189.773, p=.000) across the three varieties, with a mean rank STTR 285.45 for NS, 148.26 for SI, and 276.42 for WT, a mean rank Top10 230.45 for NS, 266.34 for SI, and 223.77 for WT, a mean rank LD 165.41 for NS, 200.13 for SI, and 322.93 for WT, and a mean rank ASL 239.50 for NS, 122.34 for SI, and 330.13 for WT. Based solely on the mean rank, three out of the four indicators (i.e., STTR, Top 10, and ASL) mark an overall more simplified nature of SI compared with NS, except for lexical density (LD) associated with informativeness. Intermodally speaking, all four

indicators feature a trend towards simplification in interpreted texts. However, since such observations were based solely on the mean rank instead of linguistic variation of the whole text segments, detailed analysis illustrating the range of variations is still needed, as plotted in the following Figures (see Figure 4.1 to 4.10).

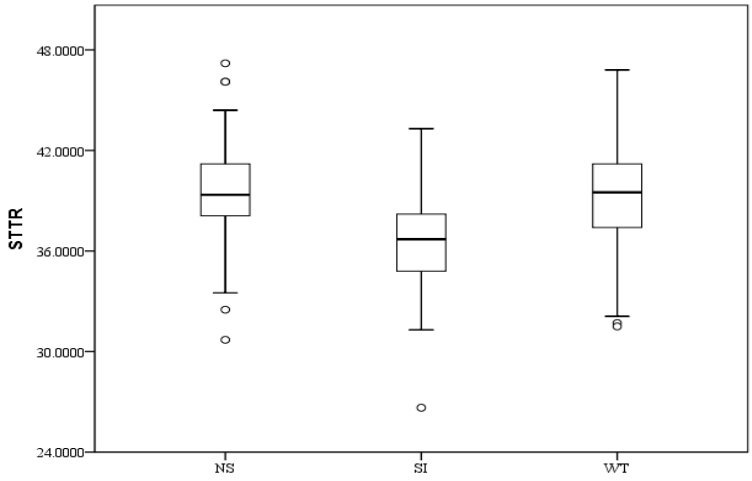

Figure 4.1 Variation of standardized type-token ratio (in percentage) across NS, SI, and WT

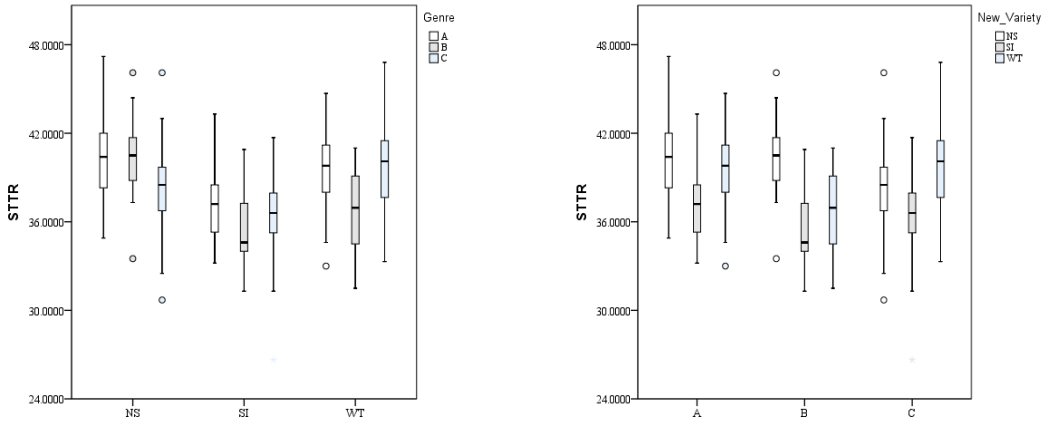

Figure 4.2 Variation of standardized type-token ratio (in percentage) within NS, SI, and WT

(Note*: Left Figure, internal variation among NS, SI, and WT; Right Figure, subgenre comparison across NS, SI, and WT; A = genre "Questions to the Prime Minister/Chief Executive; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = unmediated native speech; SI = simultaneous interpreting into B; WT = written translation into B)

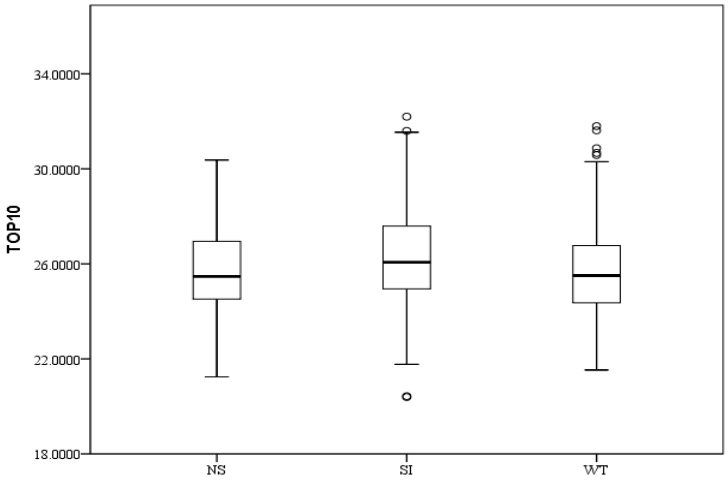

Figure 4.3 Variation of top 10 vocabulary coverage (in percentage) across NS, SI, and WT

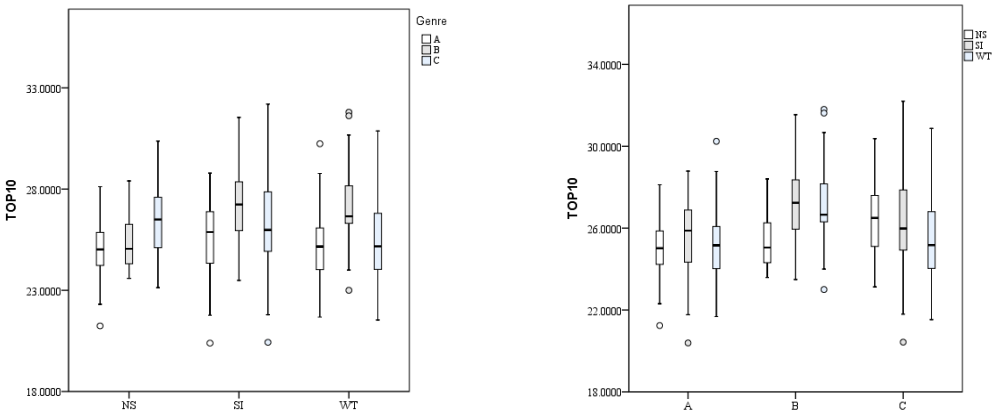

Figure 4.4 Variation of top 10 vocabulary coverage (in percentage) within NS, SI, and WT

(Note*: Left Figure, internal variation among NS, SI, and WT; Right Figure, subgenre comparison across NS, SI, and WT; A = genre "Questions to the Prime Minister/Chief Executive; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = unmediated native speech; SI = simultaneous interpreting into B; WT = written translation into B)

Figure 4.1 to 4.4 present the variation of two linguistic features (i.e., STTR and Top10) as indicators for lexical variety and repetitiveness both across and within non-interpretations, interpretations, and translations. Conforming to the initial analysis done in section 4.1, SI is overall much less lexically varied and more repetitive from both comparable and intermodal perspectives. This pattern is particularly prominent in terms

of STTR, as the majority of the interpreted texts rely on only 35 to 38 types out of 100 tokens (STTR between 35% to 38%), with two outliers characterized by standardized type-token ratio even lower than 30 (and 25). In terms of the consistency pattern of lexical repetitiveness across subgenres, Figure 4.2 and 4.4 (right figures) report mixed findings. For STTR, the mean rank suggests consistently more simplified trend of interpreted texts in relation to the corresponding subgenres in non-interpreted and translated texts. By contrast, the consistency trend is much more blurred regarding linguistic variation of top10 vocabulary coverage, suggesting possible genre-sensitive lexical patterns. An interesting pattern to note is that the variation patterns within SI and WT subgenres follow a similar trend (see the left Figures in Figure 4.2 and 4.4), which may well indicate source language interference. To verify this assumption, a parallel analysis aligning source texts with both interpreted and translated texts needs to be done, which is currently beyond the scope of this study.

In terms of informativeness (operationalized as lexical density) of the three varieties (see Figure 4.5 and 4.6), SI is found to be more lexically dense than native speech in terms of its mean rank (U=8741.000, p=0.28), a finding conforming to Kajzer-Wietrzny (2012, 2015), Ferraresi et al. (2018), and partly Dayter (2018), while less so than written translation (U=6824.000, p=0.000). Despite this finding, in terms of the similarity between interpretations and non-interpretations, and interpretations and translations with regard to lexical density, based on the range of linguistic variation of LD across the three varieties (see Figure 4.5), SI resembles NS more than WT, indicating a stronger influence of modality (being spoken or written) over ontology (being mediated or unmediated), lending some support to the more 'spoken' nature of interpreting (Shlesinger & Ordan, 2012). Such an influence can also be testified in terms of the consistency pattern across subgenre comparisons (see right figure in Figure 4.6), where SI subgenres show an overall consistent pattern of being less lexically dense than corresponding WT subgenres, while the evidence of SI being consistently more informative is not equally strong viewed from a comparable perspective, particularly with respect to genre B ("*Questions to the Secretaries/Ministers*") comparison, as the mean rank of SI_B suggests that interpreted

language is less lexically dense than NS_B, implicating a possible genre influence over the general variation patterns in this regard.

Figure 4.6 (left figure) confirms genre variation in terms of lexical density across the three English varieties. Overall, genre C ("*Debates*") is characterized by lower lexical density than Q&A genres, indicating its less informative nature. This can be attributed to the different functions served by Debates and Q&As, i.e., to persuade and to offer (new or old) information.

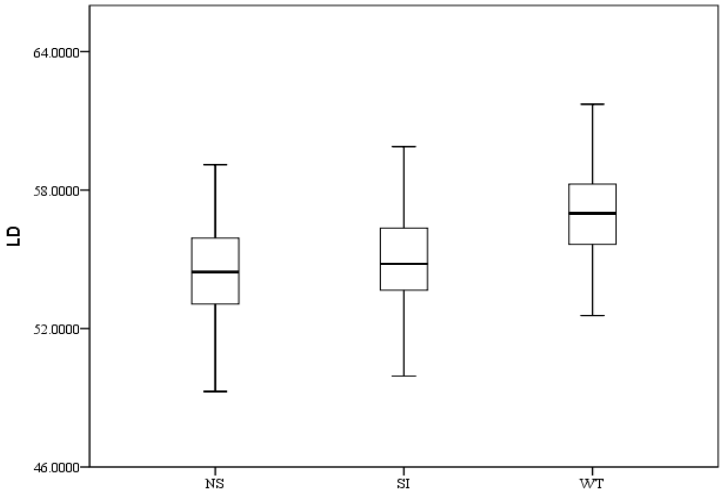

Figure 4.5 Variation of lexical density (in percentage) across NS, SI, and WT

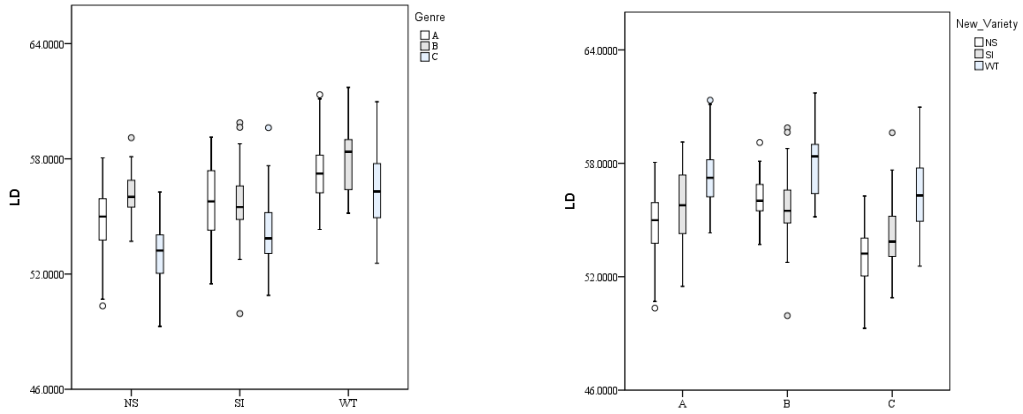

Figure 4.6 Variation of lexical density coverage within NS, SI, and WT

(Note*: Left Figure, internal variation among NS, SI, and WT; Right Figure, subgenre comparison across NS, SI, and WT; A = genre "Questions to the Prime Minister/Chief Executive; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = unmediated native speech; SI

= simultaneous interpreting into B; WT = written translation into B)

The last indicator for simplification is average sentence length (ASL). For both spoken and written components of the LegCo+ corpus, i.e., SI, NS, and WT, average sentence length is calculated by dividing the total number of running words in each text segment by the number of sentences in that text segment, as done in Bernardini et al. (2016). For SI and NS, punctuation markers were added during the transcription, based on the intonation of the speakers/interpreters as well as syntactic information, which makes the calculation of average sentence length much easier. Figure 4.7 and 4.8 plot the variation patterns of average sentence length both across and within the three English varieties. SI shows consistent patterns of being syntactically less sophisticated than both NS and WT, except for subgenre comparison between NS_B and SI_B (genre B "*Questions to the Ministers/Secretaries*") which suggests an overall opposite trend (see right figure in Figure 4.8). In terms of internal variation patterns of the different subsets in SI and WT (see left figure in Figure 4.8), once again a similar variation trend has been identified, with subgenre "*Questions to the Secretaries*" (i.e., SI_B and WT_B) resorting to noticeably longer sentence length than both "*CE Questions and Answers*", and "*Debates*". This finding contradicts Li and Wang (2012) who report longer sentence length of simultaneously interpreted discourse from Cantonese, and the average sentence length they report is 22.55 words. In this study, the average sentence length of the interpreted texts stands between 16 to 17 words, except for "*Questions to the Secretaries*" (SI_B), the average sentence length of which reaches around 22 words. Source language interference may have contributed to this variation trend, same with the variation trends of STTR and top 10 within SI and WT.

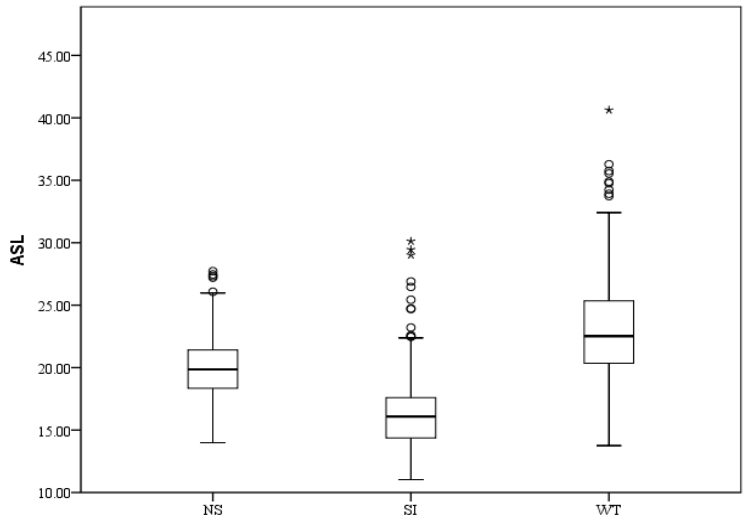

Figure 4.7 Variation of average sentence length (per 100 tokens) across NS, SI, and WT

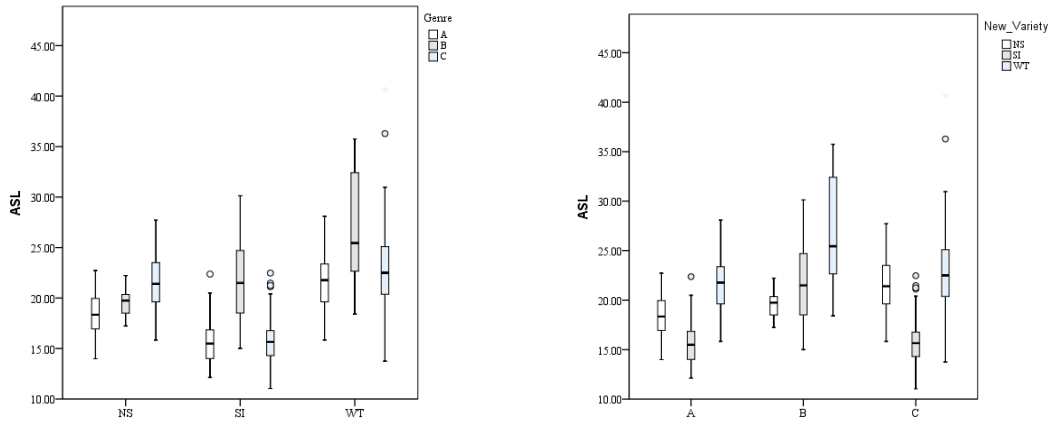

Figure 4.8 Variation of average sentence length (per 100 tokens) within NS, SI, and WT

(Note*: Left Figure, internal variation among NS, SI, and WT; Right Figure, subgenre comparison across NS, SI, and WT; A = genre "Questions to the Prime Minister/Chief Executive; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = unmediated native speech; SI = simultaneous interpreting into B; WT = written translation into B)

4.2.1.2 Discussion of results

To sum up the results reported in section 4.2.1.1, the analysis shows an overall coherent pattern of lexical simplification of interpreted English with respect to native spoken English and translated English from the same source, albeit diverging consistency patterns when subgenre variations are taken into consideration.

Starting from a monolingual comparable analysis, SI resorts to more simplified language use, characterized by lower lexical variety (STTR), more repetitive language use (top10), and lower syntactic sophistication (average sentence length), compared with native speech. However, a contradictory trend is identified in terms of lexical density, in which case SI shows a more informative nature than NS. Kajzer-Wietrzny (2012, 2015) report similar contradictory findings: while interpreted texts were found to rely more on the use of high frequency words, as well as list heads, they were also found to be more lexically dense than original spoken texts. Three possible reasons are suggested by Kajzer-Wietrzny (2012, 2015) regarding higher lexical density in interpreted language, including the interpreters' avoidance of redundancy, the interpreters' use of the condensation strategy due to "time constraint" (Shlesinger, 1995), and a possible explicitating shift from referential to lexical cohesion as suggested by Shlesinger (1995) and Gumul (2006). The present author could argue the same possible reasons for the increased informativeness of SI. Section 4.1, for example, reports an overuse of nominal features in SI in relation to NS, which may suggest that simultaneous interpreters explicitate referential cohesive ties to lexical ones. Another possible reason is argued by Shlesinger (1989) that higher lexical density may indicate higher degree of planning. Although the present author has no direct access to the mode of delivery (read, impromptu, or mixed) of simultaneous interpreters, based on SI transcriptions, as well as the paralinguistic features annotated in the transcription (such as constant sound of page flipping), she finds overall better preparation of simultaneous interpreters in translating "*CE Q&As*" and "*Questions to the Secretaries*" subgenres than translating the subgenre "*Debates*" sessions. The variation patterns of lexical density in Figure 4.6 confirm the author's assumption. Plausible as these explanations may sound, it should be borne in mind that without knowing the actual co-occurrence patterns of correlated linguistic features (such as positive correlation between nominal features and lexical density, or negative correlation between cohesive markers and lexical density), it would be too arbitrary to jump into conclusions.

When subgenre comparisons from comparable and intermodal perspectives are examined,

this more simplified nature of SI is not always consistent and clear-cut, especially in terms of variation patterns of top 10 and lexical density. Previous studies drawing on parliamentary data seldom consider (sub-)genre variation within the parliamentary setting, which may mask the nuances of different variation patterns among different parliamentary sessions. In this study, a genre difference, albeit not as equally strong as genre/register variation studies on translational language, is observed, evidenced in particular in "*Questions to the Secretaries*" in mediated texts regarding variation of average sentence length. This may have implications for future studies on the consistency of lexical patterns of interpreted language across genre/register comparison. In addition, this finding may also suggest that SI into B (or L2 interpreting) does not necessarily show more pronounced simplification patterns compared with native interpreting, since the identified patterns are subject to (sub-)genre influence and thus not consistent enough, as shown in Figures 4.4 and 4.6.

Moving on to an intermodal point of view, interpreted texts show consistent patterns of being more simplified than translated texts of the same source, a finding in line with both Bernardini et al. (2016) and Ferraresi et al. (2018) based on their EPTIC data. Such consistent variation patterns implicate a strong influence of modality, studies on which (Chafe, 1982; Chafe & Tannen, 1987) often show a more simplified and fragmented nature of spoken language in relation to written language. However, it may also highlight the differences of intrinsic constraints experienced by (simultaneous) interpreters and translators. As argued by Kruger and Van Rooy (2016a, p. 27), "[l]anguage production in translation is cognitively constrained by the fact that it involves bilingual language activation and is circumscribed by a previously produced text. Translation is also characterised by normative constraints that determine target-language and -culture acceptability". Language production in (simultaneous) interpreting is even more constrained, especially in terms of cognitive constraints, situational constraints, and linguistic constraints (Lanstyak & Heltai, 2012). Cognitively speaking, interpreters need to cope with multi-tasking (i.e., listening, comprehension, memory, production and coordination) (Gile, 1995/2009) with limited processing capacity, and in certain situations

they may resort to simplification as a coping strategy when their attentional resources are in shortage. Situational constraints mean that the working conditions of (simultaneous) interpreters differ from those of translators, such as the setting, access to primary speakers, access to prepared speeches or Powerpoint of main speakers, background noise, etc.. In particular, interpreters are paced by original speakers, and their language production is realized "on the spot" (Pöchhacker, 2016), which may have contributed to a more simplified language use in interpreted texts. As far as linguistic constraints are concerned, interpreters have more limited linguistic resources than translators, so there may be cases when interpreters have to leave out certain information, use less varied lexis/vocabulary in their output, thus leading to an overall more simplified language use.

Based on these observations, several reflections can be made concerning this unidimensional analysis of linguistic patterns of interpreted language.

To start with, the current research, based on a less investigated and genetically distinct language pair Cantonese/English, confirms, in general, the lexical simplification trend of interpreted English from both comparable and intermodal perspectives. This may suggest, both echoing and extending Shlesinger and Ordan's (2012) observation, that interpreting is an extreme case of spoken language as well as translational language.

Closely related to the first reflection is that, working direction (A-to-B, or B-to-A) in this specific case may have negligent influence on the simplification patterns of linguistic features in the interpreting output. Previous expectation was that the simplified pattern of SI into a B language (L2) would be more prominent and consistent with respect to that of native interpreting, due to the much harsher constraints experienced by simultaneous interpreters (Donovan, 2005; Seleskovitch, 1987). The mixed consistency patterns across subgenre comparisons provided counterevidence to this expectation. One possible reason is that since the (professional) interpreters under investigation work unanimously into one single direction (A-to-B) in the LegCo setting, the influence of directionality may have been factored out compared to those who constantly shift between A-to-B and B-to-A working directions as practiced in the European market. That being said, a more well-

designed research replicating previous studies (e.g., Ferraresi et al., 2018; Russo et al., 2006; Sandrelli & Bendazzoli, 2005), with effect sizes of the linguistic features under discussion reported, might be a better way to see the potential interaction between directionality and simplification.

Another reflection, as reported in section 4.1.1.1, is that this generally more simplified trend of interpreting is also (sub-)genre-sensitive with respect to certain linguistic features, due to specifically internal (sub-)genre variations within the same/similar legislative setting. A more fine-grained analysis taking into consideration various text/speech types needs to be done before reaching any conclusion.

A fourth reflection is that modality or mediation mode, as argued by Shlesinger (2008), may indeed have a larger influence than mere mediation status (mediated or unmediated), given the more consistently simplified nature of interpreted English compared to translated English, as well as the greater similarity between interpreted English and original spoken English in terms of lexical density and average sentence length. Still, the specific constraints intrinsic to SI, especially the cognitively more challenging nature of L2 interpreting, cannot be ruled out.

Last but not least, source language influence, though without direct reference to source speeches (ST), may have played a role on the similar variation patterns within SI and WT (see Figure 4.2, 4.4, 4.6 and 4.8). Upon further reflection, such similarity might be related to the production conditions as well as the very nature of the three subgenres in ST. Take the variation pattern of average sentence length (see Figure 4.8) as an example. Among the three subgenres in both SI and WT, '*Questions to the Secretaries*' stand out from the other two, featured by much longer average sentence length. Such a noticeable difference can be traced back to ST which is composed of both written/scripted Q&As characterized by a higher level of literateness (e.g., longer sentences, more formal language use), and spoken (often unscripted) Q&As. Viewed from another perspective, the heavy influence of the linguistic patterns of source speeches may provide counterevidence to the equalizing effect suggested by Shlesinger (1989, pp. 170–171), which states that

"interpretation diminishes the orality of markedly oral texts and the literateness of markedly literate ones" and that "the range of the oral-literate continuum is reduced in simultaneous interpreting" However, to verify this point, direct comparison between ST and SI in terms of the oral-literate usage patterns of linguistic features needs to be done, which is beyond the scope of the current study.

4.2.2 Exploring explicitness from comparable and intermodal perspectives

4.2.2.1 Data analysis

In contrast with the limited number of linguistic indicators for lexical simplification (i.e., list heads, STTR, high frequency words, lexical density, and average sentence length), linguistic indicators for the explicitation pattern in interpreting are much more diversified, manifested in particular by a number of explicitating shifts reported in Gumul's studies (2006, 2007, 2008, 2017, 2020). The more fine-grained explicitating shifts aside, many studies (e.g., Dayter, 2018; K. B. Hu & Tao, 2009; Kajzer-Wietrzny, 2012, 2015; Morselli, 2018; Shlesinger, 1995; Shlesinger & Ordan, 2012) focus on the parameters used often in translation studies associated with informational and logical explicitness, such as the use of optional 'that' verb complementizer, cohesive ties, linking adverbials and apposition markers, as well as part-of-speeches (PoS) distributions. In this section, the author follows these traditions by presenting both internal and external variations of eight entwined linguistic features, including conjuncts (CONJ), causative adverbial subordinators (CAUS), concessive adverbial subordinators (CONC), conditional adverbial subordinators (COND), other adverbial subordinators (OSUB), 'that' adjective complement (THAC), 'that' verb complement (THVC), and subordinator-that deletion (THATD), tagged by the Multidimensional Analysis Tagger (MAT) introduced in Chapter three, in NS, SI and WT. Out of the eight linguistic features, the four types of adverbial subordinators were categorized and investigated as one: total adverbial subordinators (ASUB). Although the three 'that' complement usage, i.e., THAC, THVC, and THATD, can be summarized as optional/zero 'that' complement clauses, the author decided to

examine them separately so as to make more direct comparison with previous findings.

Before the visualization, as done in section 4.2.1, a Kruskal-Wallis H test was performed to get a quick glimpse of the general differences across NS, SI and WT in respect of these features, as shown in Table 4.6.

Table 4.6 Kruskal-Wallis H Test of explicitation features across NS, SI and WT

Linguistic parameters	Language variety	N	Mean Rank	Chi-square	df	Asymp.Sig
CONJ	NS	138	126.87	252.471	2	.000
	SI	149	190.63			
	WT	190	358.37			
ASUB	NS	138	178.17	52.956	2	.000
	SI	149	296.65			
	WT	190	237.97			
THAC	NS	138	341.33	115.250	2	.000
	SI	149	193.76			
	WT	190	200.16			
THVC	NS	138	203.62	13.486	2	.001
	SI	149	260.41			
	WT	190	247.91			
THATD	NS	138	331.13	131.806	2	.000
	SI	149	258.14			
	WT	190	157.08			

The Kruskal-Wallis H test in Table 4.6 showed that there are statistically significant difference in terms of the use of CONJ (X^2=252.471, p=.000), ASUB (X^2=52.956, p=.000), THAC (X^2=115.250, p=.000), THVC (X^2=13.486, p=.001) and THATD (X^2=131.806, p=.000) across the three, with a mean rank CONJ 126.87 for NS, 190.63 for SI, and 358.37 for WT, a mean rank of ASUB 178.17 for NS, 296.65 for SI, and 237.97 for WT, a mean rank THAC 341.33 for NS, 193.76 for SI, and 200.16 for WT, a mean rank THVC 203.62 for NS, 260.41 for SI, and 247.91 for WT, and a mean rank THATD 331.13 for NS, 258.14 for SI, and 157.08 for WT. From a comparable perspective, SI shows overall a consistent pattern of being more explicit than NS, except for 'that' adjective complement, which was distinctively overused in NS. From an intermodal perspective, however, only three out of the five features, i.e., CONJ (U=3410.000, p=.000), ASUB (U=10494.000, p=.000), and THATD (U=8042.500, p=.000), showed statistically significant difference between SI and WT, with CONJ being noticeably overused in WT while ASUB and THATD overused in SI. This suggests mixed variation patterns of

explicitness trend of interpreting compared to translation. While the overuse of conjuncts and underuse of subordinator-that deletion in translation indicate that translation is more explicit than interpreting, the overuse of total adverbial subordinators in interpreting suggests the opposite. A more in-depth analysis may help account for such contradictory trend. The following figures (Figures 10 to 19) present respectively the variation patterns of the five features both across and within NS, SI, and WT.

Conjuncts tagged in this research include both linking adverbials (i.e., consequently, in consequence, as a consequence, as a result, hence, therefore, thus) and part of the apposition markers (i.e., namely, in other words, that is, that is to say) as examined in Kajzer-Wietrzny (2012), and many others (e.g., furthermore, likewise, in addition, in conclusion, alternatively). It may thus offer a much broader picture of the explicitness pattern of the interpreted texts from both comparable and intermodal angles. Figure 4.9 shows that, consistent with the statistical tests, overall interpreters use more conjuncts than native English speakers, leading to more explicit interpreting outputs. In comparison, interpreters use far less conjuncts than translators, since the number of conjuncts in translated texts quadrupled that in interpreted texts. Comparison between SI and NS contradicts partially the findings reported in Kajzer-Wietrzny (2012) and Morselli (2018), in which no distinctive patterns of interpreted texts in linking adverbials have been found based on the comparable analysis (interpreted vs. non-interpreted texts). In terms of apposition markers, however, both have observed more frequent use in interpreted vs. non-interpreted texts (Kajzer-Wietrzny, 2012; Morselli, 2018), leading to contradictory patterns of explicitness in interpreting with respect to these two indicators. In the current research, the author does not make any distinction between linking adverbials and apposition markers, thus it is not possible to compare directly with the aforementioned studies. Nevertheless, Figure 4.9 and Figure 4.10 demonstrate that, generally SI is more explicit than NS, even across subgenre comparisons in the two varieties (except for NS_C vs. SI_C). An opposite trend is observed from an intermodal perspective, where SI shows consistently much less explicit nature than WT across all subgenre comparisons as visualized in Figure 4.10. Two further observations can be made: first, as indicated by the

range of variation (measured by the upper and lower quartiles) in each subset in the three language varieties, original spoken texts show the least variation, followed by interpreted texts, while translated texts have the largest internal variation, indicating possible contradiction to the "levelling-out" hypothesis (Baker, 1996; Laviosa, 2002). Second, in terms of the similarity between SI and NS, and between SI and WT, SI resembles NS more in the variation patterns of conjuncts, suggesting greater influence exerted by modality or modes of mediation than the mere status of mediation.

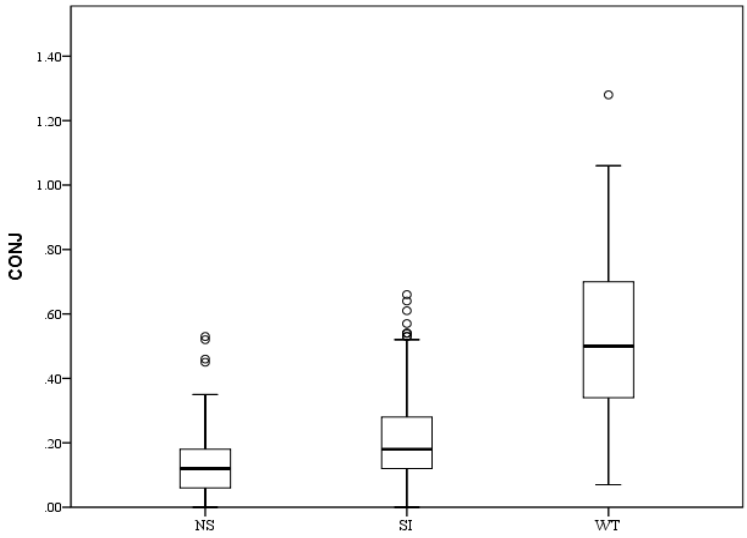

Figure 4.9 Variation of conjuncts (per 100 tokens) across NS, SI, and WT

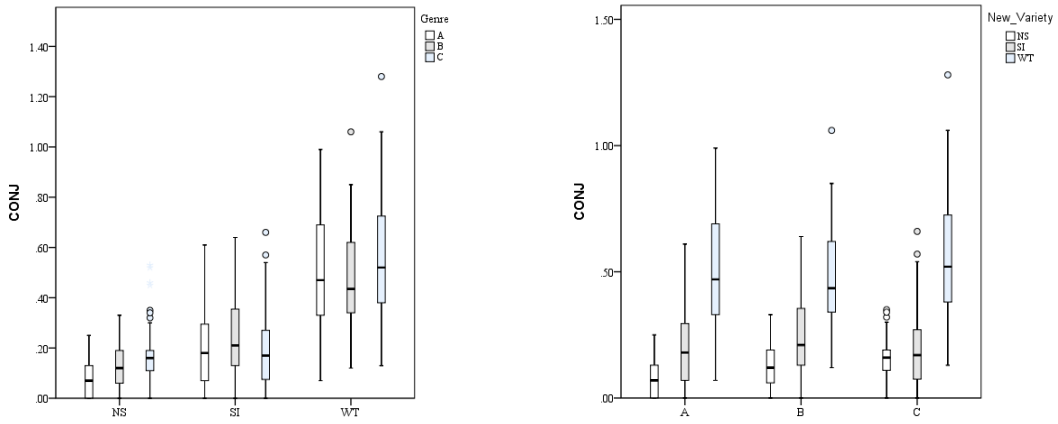

Figure 4.10 Variation of conjuncts (per 100 tokens) within NS, SI, and WT

(Note*: Left Figure, internal variation among NS, SI, and WT; Right Figure, subgenre comparison across NS, SI, and WT; A = genre "Questions to the Prime Minister/Chief Executive; B = genre

"Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = unmediated native speech; SI = simultaneous interpreting into B; WT = written translation into B)

Figure 4.11 and 4.12 plot the variation of total adverbial subordinators (ASUB) across and within the three varieties as an indicator for logical explicitness. In terms of comparable analysis, SI again shows more explicit pattern than NS, and this pattern is relatively consistent across subgenres, although it is not as equally strong in NS_A vs. SI_A considering the range of variation (see Figure 4.12). When it comes to intermodal comparison between translation and interpreting, contrary to the variation pattern of conjuncts (CONJ) discussed above, SI is also characterized by a higher degree of logical explicitness than translation, and this pattern is fairly consistent across the three subgenres. Taken together, interpreters use more frequently total adverbial subordinators to explicitate the dependency relations between the main clause and the dependent clause than both native speakers and translators. The wider range of variation, indicated by upper and lower quartiles, within SI texts also suggests such usage varies from text to text, which may be attributed to a number of reasons, such as preparedness of the speech, and source speech delivery. In addition, Figure 4.12 presents, once again, similar variation patterns within the three subsets in SI and WT, manifesting a possible source language influence.

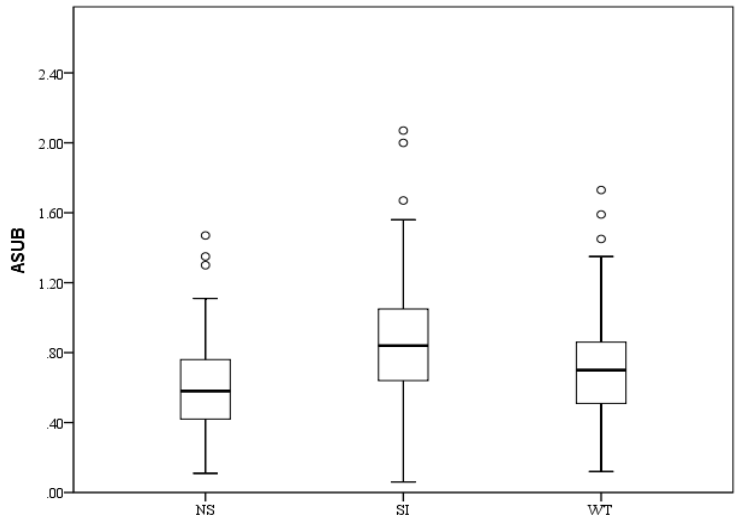

Figure 4.11 Variation of total adverbial subordinators (per 100 tokens) across NS, SI,

and WT

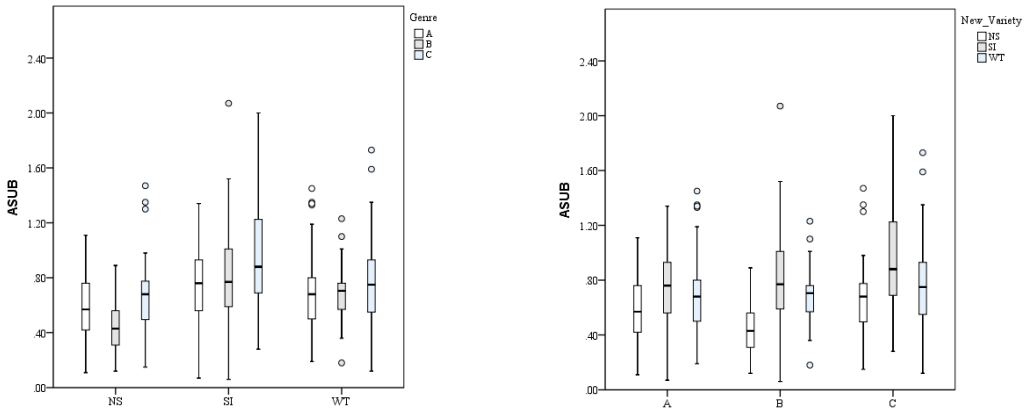

Figure 4.12 Variation of total adverbial subordinators (per 100 tokens) within NS, SI, and WT

(Note*: Left Figure, internal variation among NS, SI, and WT; Right Figure, subgenre comparison across NS, SI, and WT; A = genre "Questions to the Prime Minister/Chief Executive; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = unmediated native speech; SI = simultaneous interpreting into B; WT = written translation into B)

The retention/omission of optional 'that' complement is a most popular indicator for explicitation in both translation and interpreting. Previous studies (Kajzer-Wietrzny, 2012; H. Kruger & Van Rooy, 2012; Olohan & Baker, 2000) focus on the use of optional 'that' or 'that' omission after a limited number of reporting verbs, such as *say*, *tell*, *think*, and *believe.* Based on these very few parameters, mediated texts are found to spell out 'that' form more often than unmediated texts, indicating a preference for increased explicitness. The variation patterns of optional 'that' clauses investigated in this study, however, cover a much broader range of 'that' complement clauses, including 'that' adjective complement (THAC[31]), 'that' verb complement (THVC[32]), and subordinator-that deletion (THATD). It is believed that such a choice can offer a broader picture regarding the explicitation patterns across the three English varieties, and avoid the skewedness of the use of optional 'that' after certain reporting verbs. Figure 4.13 presents the variation patterns of 'that'

[31] Examples: (1) "Isn't is **clear *that*** the failure of western security strategy in the middle-east and elsewhere is the main driver of this migration crisis …" (NS_A3_04)
(2) I'm **pleased *that*** we have secured the continuation of qualifications in community languages. (NS_B4_05)
[32] Examples: (1) "I don't **think *that*** it is something which is passive." (SI_A4_07)
"It **seems *that*** I am talking about a subdivided unit in Sham Shui Po." (WT_C4_10)

complement clauses, including ‘that’ adjective clauses, ‘that’ verb clauses, and omission of optional ‘that’, across NS, SI, and WT. Overall, native English speakers prefer the use of ‘that’ complement clauses more than simultaneous interpreters. Translators, by contrast, rely least often on ‘that’ complement clauses. While this piece of information does not tell us straightforward the possible explicitation patterns, it does suggest an overall difference in terms of the preference for post-predicate ‘that’ clauses, which is closely related to the expression of personal stance (Biber et al., 1999). This pragmatic function (i.e., stancetaking) of ‘that’ clauses aside, Biber et al. (1999) also report register difference, in which ‘that’ complement clause structures are most commonly used in conversation, followed by fiction and news, while least common in academic prose. That is, texts with more oral features tend to rely more on post-predicate that-clauses than those with more literate features, which also seems to be the case in Figure 4.13.

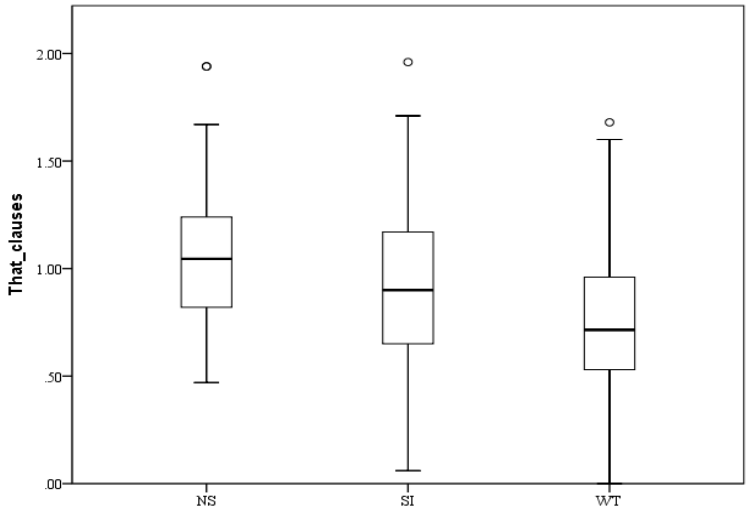

Figure 4.13 Variation of ‘that’ complement clauses (per 100 tokens) across NS, SI, and WT

To examine closely the patterns of syntactic explicitness, the author displayed the variation patterns of the three ‘that’ clause structures separately. Figure 4.14 and 4.15 demonstrate the variation of ‘that’ adjective clauses, a feature under-explored in previous studies, both across and within NS, SI, and WT. Contrary to expectation, SI is featured by statistically less frequent ‘that’ adjective clauses compared to native speech

(U=4081.500, p=.000), while there is no statistically significant difference between SI and WT in this regard (U=13613.500, p=.512). At first sight, this finding seems unexpected, since previous studies on optional 'that' verb complements provide strong evidence for the more explicit nature of mediated texts (Kajzer-Wietrzny, 2012; H. Kruger & Van Rooy, 2012; Olohan & Baker, 2000). Biber et al. (1999, p. 671) offer a possible explanation, stating that "[t]hat adjectives that control a *that* complement clause all convey stance", such as the speaker's/writer's degree of certainty, affective psychological states, or evaluation of situations, etc.. Examined from this point of view, this variation seems plausible as it may suggest that both interpreters and translators try to avoid stancetaking during translation. Figure 4.15 reveals further interesting patterns that all the subgenres in SI and WT are very homogeneous in this respect, while the use of 'that' adjective complements in NS obviously reveals (sub-)genre variations, with subgenre NS_B ("*Oral questions to the Ministers*") more marked in the use of 'that' adjective complements than NS_A and NS_C. Given the homogeneity of linguistic variation in translation and interpreting, and across SI and WT subgenres, the underuse of 'that' adjective complements may well be a translation-specific (i.e., mediation-specific) feature.

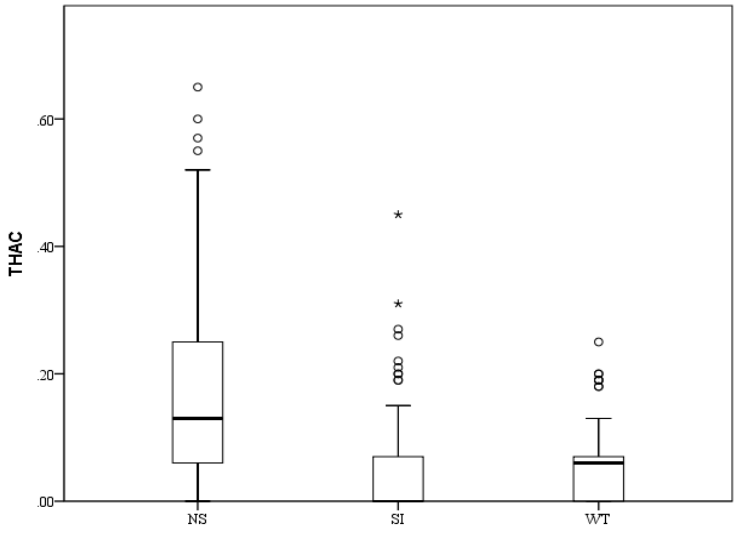

Figure 4.14 Variation of 'that' adjective complement (per 100 tokens) across NS, SI, and WT

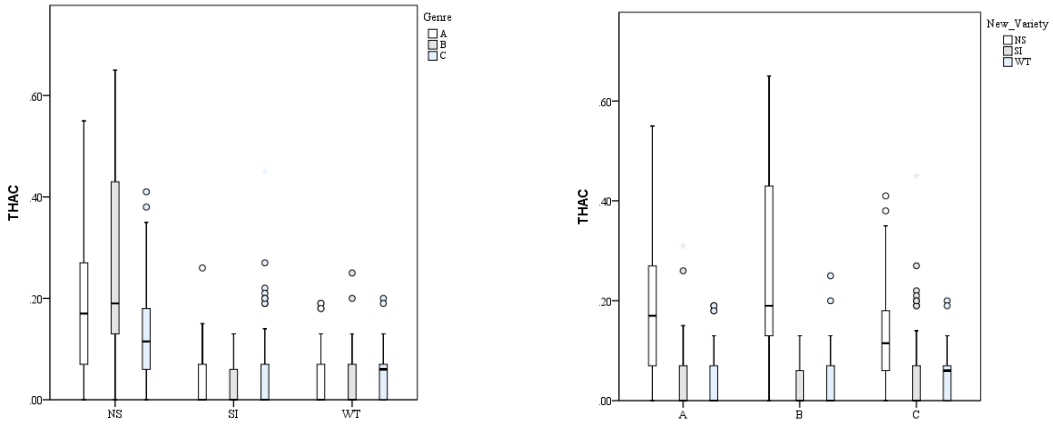

Figure 4.15 Variation of 'that' adjective complement (per 100 tokens) within NS, SI, and WT

(Note*: Left Figure, internal variation among NS, SI, and WT; Right Figure, subgenre comparison across NS, SI, and WT; A = genre "Questions to the Prime Minister/Chief Executive; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = unmediated native speech; SI = simultaneous interpreting into B; WT = written translation into B)

The use of 'that' verb complement (THVC) and subordinator-that deletion (THATD) or optional 'that' omission will be considered together, as shown in Figure 4.16, following previous studies (Kajzer-Wietrzny, 2012; Morselli, 2018). Several interesting patterns are observed. For a start, native speakers tend to use optional or zero 'that' verb complements equally compared with interpreters and translators, who show a noticeable preference for verbalizing optional 'that', as been observed also in previous studies (Kajzer-Wietrzny, 2018; H. Kruger & Van Rooy, 2012; Olohan & Baker, 2000). This may have some implication for the shared feature of translation and interpreting as mediated language, evidenced by the similar variation patterns in optional/zero 'that' usage (see Figure 16). In terms of monolingual comparable comparison, SI shows a coherent pattern of being more explicit than NS, highlighted by an overuse of optional 'that' while an underuse of 'that' omission. Intermodally speaking, SI seems to be less explicit than WT given its overuse of 'that' omission, while it does not differ distinctively from translation in optional 'that' verbalization. These initial observations are partly in line with Kajzer-Wietrzny (2012), in which she reports an explicitating trend of both interpreted and translated texts in all subcorpora of TIC, irrespective of source languages, and that trend is even more pronounced in translated texts.

In terms of the consistency of the observed variation patterns, the trend is not very clear-cut (see the right figure in Figure 17). While Figure 4.16 shows that interpreters tend to spell out optional 'that' more often than native speakers, this has only been partially confirmed across subgenre comparisons, as an opposite trend has been identified in SI_B vs. NS_B. Figures 4.17 and 4.18 (left figures) also demonstrate subgenre variations in this respect, where SI_B stands out from the others (i.e., SI_A, and SI_C), showing overall a less explicit pattern. When the use of optional and zero 'that' verb complements (i.e., THVC and THATD) is viewed in a complementary manner, meaning an overuse of 'that' verb complement should ideally predict an underuse of subordinator-that deletion, the variation patterns of SI in Figures 4.17 and 4.18 are quite blurred. Various factors might be at play, which will be discussed in section 4.2.2.2.

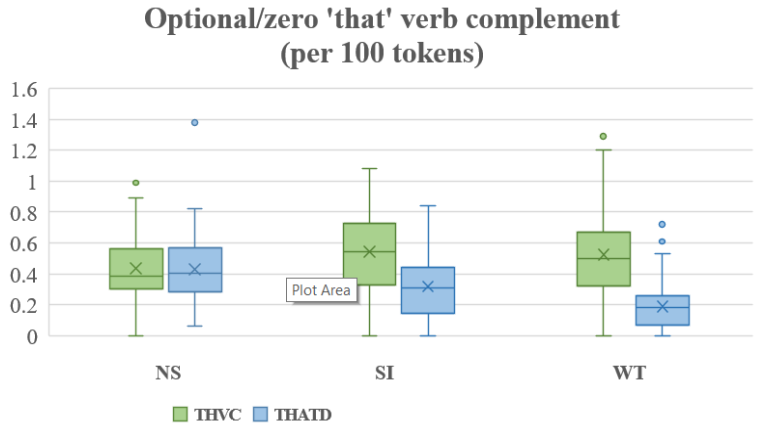

Figure 4.16 Variation of optional/zero 'that' verb complement across NS, SI, and WT

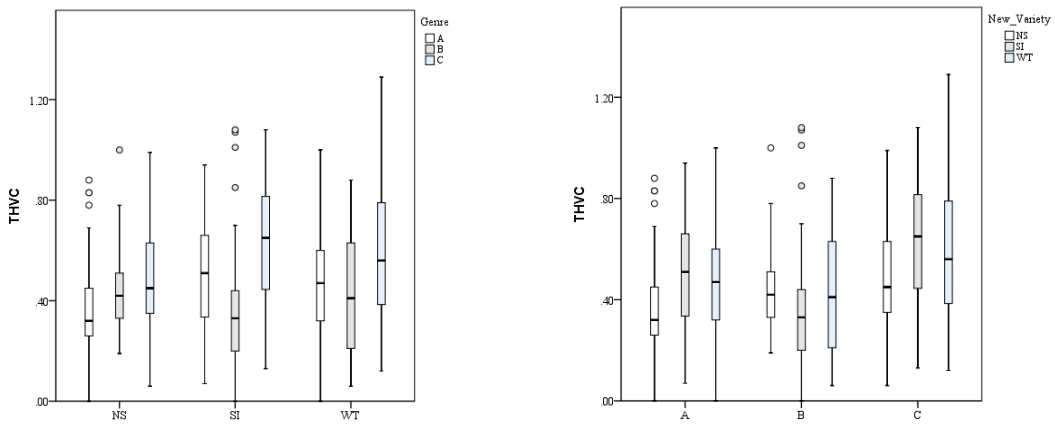

Figure 4.17 Variation of 'that' verb complement (per 100 tokens) within NS, SI, and WT

(Note*: Left Figure, internal variation among NS, SI, and WT; Right Figure, subgenre comparison across NS, SI, and WT; A = genre "Questions to the Prime Minister/Chief Executive; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = unmediated native speech; SI = simultaneous interpreting into B; WT = written translation into B)

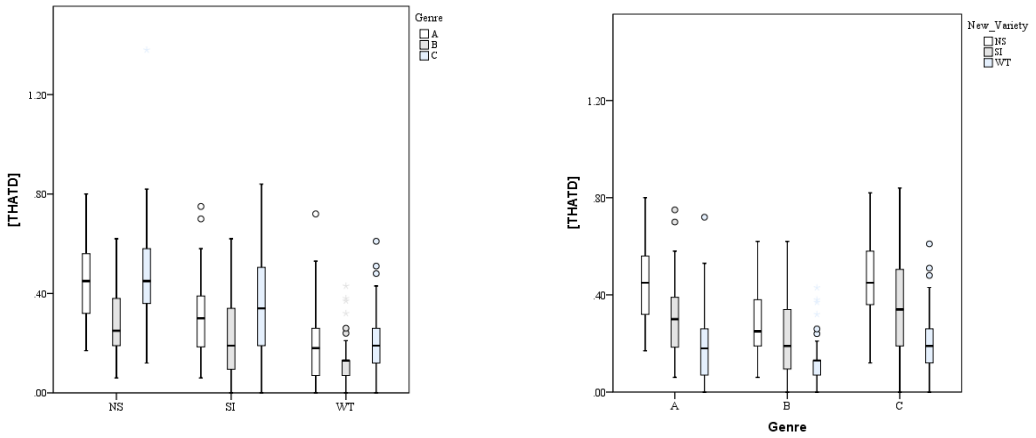

Figure 4.18 Variation of subordinator-that deletion (per 100 tokens) within NS, SI, and WT

(Note*: Left Figure, internal variation among NS, SI, and WT; Right Figure, subgenre comparison across NS, SI, and WT; A = genre "Questions to the Prime Minister/Chief Executive; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = unmediated native speech; SI = simultaneous interpreting into B; WT = written translation into B)

4.2.2.2 Discussion of results

To summarize the results, in terms of the variation patterns of increased explicitness operationalized by eight linguistic features grouping into five (i.e., CONJ, ASUB, THAC, THVC, THATD), SI show mixed patterns viewed from comparable and intermodal perspectives, respectively.

As far as comparable comparison is concerned, confirming many of the previous studies (K. B. Hu & Tao, 2009, 2013; Kajzer-Wietrzny, 2012), SI shows overall increased explicitness than NS, except for 'that' adjective complement, which is found to be significantly underused in both translation and interpreting. One possible reason is the interpreters' avoidance of stancetaking, as 'that' adjective complement is believed to be associated with the expression of personal stance (Biber et al., 1999). The higher degree of explicitness in terms of the remaining four indicators (i.e., conjuncts, total adverbial

subordinators, 'that' verb complement, subordinator-that deletion) in SI can be attributed to a number of reasons, such as higher information density, formality, higher cognitive complexity, source language interference, preference for specific matrix verbs (such as *think*, *say*, *tell*), time constraint, modality, and risk-averse/disambiguation concerns (Biber et al., 1999; Jaeger, 2010; Kajzer-Wietrzny, 2012, 2018; H. Kruger, 2018, 2019; H. Kruger & De Sutter, 2018; Olohan & Baker, 2000; Pym, 2005; Tagliamonte & Smith, 2005). Kajzer-Wietrzny (2018), for example, argues that information density is 'a strong predictor of *that*-mentioning' (p.101) based on the TIC corpora; while Kruger (2019), in an effort to disentangle the possible factors for optional/zero 'that' complementizer, finds "strong support the pragmatic risk-avoidance account of translational explicitation than for cognitive-complexity (or processing strain) account", since translators tend to spell out full form 'that' "even in contexts of low complexity, cognitive demand and communicative risk […] and even in registers where zero is the norm" (p.23). To account for the correlation between the use of optional 'that' and the possible contributors, a well-designed multifactorial analysis as done by Kruger (2018, 2019), Kruger and De Sutter (2018), and De Sutter and Vermeire (2020), etc., is needed.

In terms of subgenre variations of increased explicitness in SI, one subgenre – SI_B ("*Questions to the Secretaries*") – stands out, especially in relation to optional/zero 'that' complementizer. Contrary to the general explicitation pattern, SI_B shows decreased explicitness compared with corresponding NS_B, manifested in its overall less frequent use of 'that' verb complement and more frequent use of subordinator-that deletion. One possible reason may be attributed to the specific production conditions of this subgenre, or mode of delivery of its source speeches (ST), since it includes both scripted (or written-to-be-read) and unscripted Q&As. Besides, as no direct information is available about the preparedness of simultaneous interpreters, the author can only surmise that simultaneous interpreters may have swapped between fully prepared translation of the scripted Q&As, and the unscripted and unprepared translation of the impromptu Q&As of source speakers, which may eventually lead to mixed variation patterns of optional 'that' compared with the other two subgenres.

The intermodal comparison, by contrast, reveals no conclusive patterns, as two out of the five indicators (i.e., conjuncts and subordinator-that deletion) suggest a more explicit nature of translations versus interpretations, one (i.e., total adverbial subordinators) suggests the opposite trend, while the remaining two (i.e., 'that' adjective complement and 'that' verb complement) fails to reveal any statistical significance between the two mediated language varieties. This finding contradicts Hu and Tao (2009, 2013) who report a syntactically more explicit nature of interpreted English from Chinese in relation to native spoken English and translated English of a similar genre. Their studies, however, focus on a different mediation mode, i.e., consecutive interpreting, which may have played a role in the variation patterns of explicitation features. These mixed and seemingly contradictory patterns suggest the complexity of language of mediation (such as translation and interpreting), given especially the various cognitive, social, and cultural constraints (Baker, 1999). For example, in terms of an increased level of explicitness in translation versus interpreting, the ontological differences between the two as pieces of written and spoken language may have played an important role. The lack of social context of translated texts as a form of writing, as well as its higher degree of formality (Biber, 1999; H. Kruger & Van Rooy, 2016a; Quirk et al., 1985), often requires translators to make explicit the information to be conveyed to avoid potential ambiguation. This does not necessarily deny the possibility of interpreted texts to be explicit on other levels, such as the use of connective ties. Due to the highly cognitive-taxing nature of simultaneous interpreting, interpreters may sometimes resort to certain connective ties, such as adverbial subordinators, as padding strategies to buy time and also to compensate for inevitable accuracies (see Defrancq et al., 2015; Gumul, 2017; Tang, 2018). Another possible reason for the more explicit nature of interpreted language in total adverbial subordinators may be related to the "into B" working direction, as testified in Gumul (2017) that explicitation becomes more salient in retour interpreting than in native interpreting due to the more demanding processing capacity management in retour. However, the present author cautions that Gumul's (2017) explicitating shifts are based on parallel comparison between source and interpreted texts, which are essentially S-universals (Chesterman, 2004, 2017) rather than T-universals as focused here.

The above discussions on the increased level of explicitness in SI have several implications. First and foremost, a translation-specific effect may exist irrespective of different modes of mediation (i.e., translation and interpreting). Based on the results reported in section 4.2.2.1, the author tentatively proposes that an underuse of 'that' adjective complement and an overuse of full form 'that' verb complement may well be translation-specific (i.e., mediation-specific) features: while the former shows both translators' and interpreters' avoidance of stancetaking, the latter may indicate their risk-avoidance consideration to 'play safe' and avoid ambiguation. Secondly, the inclusion/exclusion of certain linguistic features (e.g., 'that' adjective complement) may have a direct influence on the lexical patterns to be investigated, which points out the question about the selection of linguistic features. For example, if previous studies had included the variation pattern of 'that' adjective complement as an indicator for increased/decreased explicitness in mediated and unmediated texts, the results may be quite different from those focusing on optional 'that' verb complement. Thirdly, when efforts are made to disentangle the possible contributors of a certain lexical pattern, correlation of linguistic features should be taken into consideration. Take the use of optional/zero 'that' complementizer as an example. Previous analysis shows that preference for full form 'that' or 'that' omission is closely correlated with several linguistic features, such as passives, infinitives, personal pronouns, and features indexing information density (Elsness, 1984; McDavid, 1964; Rohdenburg, 1996; Tagliamonte & Smith, 2005; Thompson & Mulac, 1991). A text laden with passives, infinitives and of higher information density tend to spell out 'that' to avoid ambiguity, while a text characterized by personal pronouns may well opt for 'that' omission. Section 4.1 reports that SI is featured by an overuse of passives, nominal features, higher lexical density, and infinitives in relation to NS, which may have contributed to the preference of full form 'that' complementizer in interpreted texts. This thus seems to highlight, once again, the necessity of a multivariate approach to disentangle possible contributors of lexical patterns, as well as to examine co-occurrence patterns of linguistic features under investigation.

4.3 Summary

The unidimensional analysis described in section 4.1 and 4.2 offers an initial glimpse over the general linguistic patterns of SI (into B) in relation to NS and WT. The interlingual comparable analysis between interpretations and non-interpretations lends some support to the widely attested 'translation universals' that mediated language is more simplified, explicit, and normalized/conventional than unmediated language, except for 'that' adjective complementizer and linguistic features indexing information density. Intermodally speaking, SI is characterized by more prominent and consistent patterns of lexical simplification, but the explicitness patterns are less straightforward, especially when the dynamic interaction between explicitation and cognitive load in simultaneous interpreting is taken into consideration (Gumul, 2020). To disentangle various contributors, a fine-grained multifactorial analysis as done by Kruger and De Sutter (2018), and Kruger (2019) needs to be done, which will be the direction for future follow-up studies.

The reported findings of this unidimensional analysis need to be considered against the background of the limitations of this method. To begin with, the selection of linguistic features to be examined has a direct influence on the outcome of the general tendencies of translation and interpreting, in which case the reported patterns of mediated texts being more simplified, explicit, and normalized/conventional might be due to mere chance. For instance, the inclusion of optional 'that' adjective provides strong counterevidence for the explicitness hypothesis of the optional 'that' usage. That said, the preference for optional 'that' verb complementizer in mediated language has also been confirmed in the Cantonese-English language pair under study, adding further evidence to the possibility that the tendency to verbalize full form 'that' verb complementizer may be an intrinsic feature of mediated language, irrespective of source languages, as well as other forms of "constrained language" (Kajzer-Wietrzny, 2018; H. Kruger & Van Rooy, 2016a, 2016b).

A second matter closely related to the first one is the negligence of other potential dimensions revealed by the selected linguistic features, as pointed out also by Evert and

Neumann (2017), Dayter (2018), and Kruger (2019). While voicing out their criticism against the unidimensional/univariate analysis, Evert and Neumann (2017) draw attention to the fact that one linguistic feature can be attributed to different, and probably correlated, lexical patterns. For instance, the overuse of nominal features identified in section 4.1 in SI compared to NS not only indicates higher degree of informativeness of the interpreted texts, it may also be associated with referential explicitness as reported in Shlesinger (1995) and Gumul (2006) (see also Hansen-Schirra et al., 2007; Steiner, 2008), as well as an increased level of formality. Kruger (2019) also points out that "[i]ncreased explicitness may […] be a collateral effect of a conservative preference for a more formal style motivated by risk avoidance", which "suggests that explicitation may be one dimension of another feature of translated language, usually described in terms such as conventionalization, normalization, standardization and conservatism" (ibid., p.23). A unidimensional analysis thus fails to capture the underlying dimensions on which the linguistic features might load, resulting in seemingly contradictory findings hard to interpret.

A third consideration goes to the various constraints "probabilistically" (H. Kruger, 2018) conditioning the different variation patterns of the three language varieties, which are hard to disentangle resorting to simple univariate techniques, such as Mann-Whitney test. More sophisticated multivariate techniques unveiling the correlation as well as interaction among different factors, as done in Kruger (2018, 2019), Kruger and De Sutter (2018), can help address this issue.

All that said, the unidimensional analysis carried out in this Chapter further facilitates our understanding of the complex nature of mediated language. Interpreting, whether native or retour, has its own linguistic patterns distinct from both non-mediated spoken language and mediated language of a different mode (such as translation), due to various conditioning factors. To get a detailed picture of how this "multifaceted and multidimensional" (De Sutter & Lefer, 2020, p. 18) nature of SI is manifested, a multidimensional analysis adapting Biber's (1988) MD approach on register variation has

been carried out in the following chapter.

Chapter 5 Co-occurrence linguistic patterns of L2 interpretese: A multidimensional analysis

The previous Chapter has offered us a general picture about the linguistic variation patterns both across and within NS, SI, and WT from a unidimensional perspective. Overall, L2 *interpretese* shares great similarity with L1 *interpretese* as borne out in previous studies from both comparable and intermodal perspectives. Correlated patterns have also been identified, such as the potential correlation between an increased level of explicitness and more simplified language use. There are also other patterns reported in Chapter four which lack ready interpretation, indicating the insufficiency of a unidimensional analysis. In this chapter, the author tries to reveal the multifaceted nature of simultaneous interpreting utilizing the multidimensional (MD) approach pioneered by Biber (1986, 1988) on register variation. Details of the MD analysis, including the statistical analysis, the functional interpretation of the identified factors, and typical features specific to L2 interpreting, i.e., L2 *interpretese*, along different dimensions, will be outlined in the following sections.

5.1 Exploration of the co-occurrence patterns: A multidimensional analysis

The major statistical technique for a multidimensional analysis, as clarified in Chapter three, is exploratory factor analysis, which aims to reduce a large number of variables (i.e., 79 linguistic features in this study) to several underlying constructs, i.e., factors or dimensions in functional terms, based on the assumption that "linguistic features co-occur to realize certain functions". The KMO and Bartlett's tests carried out in Chapter three have confirmed the suitability of the research data for factor analysis. In the following sections, the results of the major steps taken for factor analysis are presented.

The principal axis factoring method extracts 20 factors all together, with 51.554% of the total variance explained (see Appendix 6). That is, 20 co-occurring patterns out of 79 linguistic features account for more than half of the variation across NS, SI, and WT.

However, it is not realistic to keep all 20 factors. For one thing, the more the number of factors to be kept, the more likely that "they are not theoretically well-defined" (Biber, 1988, p. 88). For another, as seen from Appendix 6, factor 1 accounts for the largest share of the total variance, with a cumulative 19.067%, while factor 2 and 3 account for 6.121% and 4.510% respectively. In comparison, factor 6 accounts for around 2% of the total variance, while the remaining factors account for only 1.5% or less than 1% of the total variance. Therefore, a decision was made to determine the number of factors to be kept based on both parallel analysis as explained in Chapter three, and the scree plot below (see Figure 5.1). The parallel analysis determines that eight factors should be kept, since the initial eigenvalues of the first eight factors extracted based on the current research data are higher than the mean eigenvalues of the randomly generated correlation matrix. The scree plot also shows a flattening line of the contribution of eigenvalues after the first eight factors. It is therefore decided to keep eight factors for current comparison.

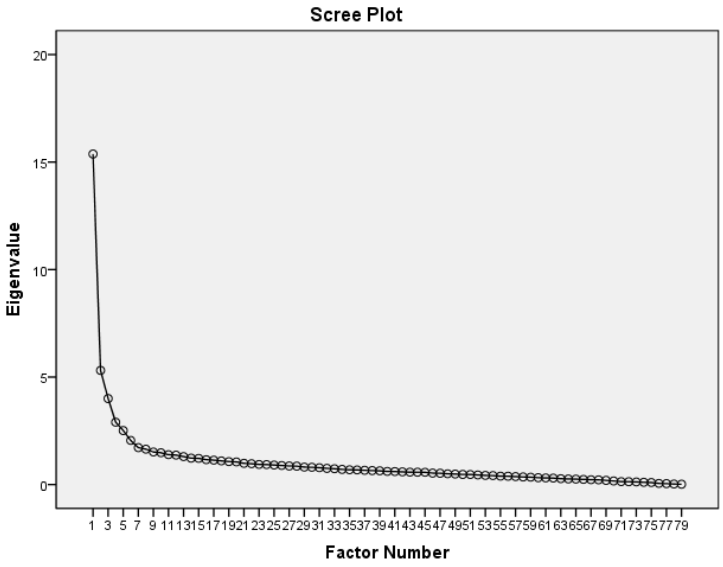

Figure 5.1 Scree plot of the eigenvalues of 79 linguistic features

Since the number of factors to be extracted had been fixed, the author run a second time factor extraction, aiming to reveal the total variance explained by the eight factors. Table 5.1 shows that the first eight factors account for about 40% of the total variance across the three English varieties (i.e., NS, SI and WT), with factor 1 accounting for the largest

share (about 17%), followed by factor 2 (about 6%). A factorial structure based on the rotated factor extraction generated by Varimax has been provided (Table 5.2; see Appendix 7 for a full version), each factor associated with a number of linguistic features featuring larger-than-0.30 factor loading.

Table 5.1 Total variance explained by the first eight factors

Total Variance Explained

Factor	Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	14.966	18.945	18.945	13.301	16.836	16.836
2	4.721	5.976	24.921	4.593	5.814	22.650
3	3.445	4.361	29.281	3.129	3.961	26.611
4	2.442	3.091	32.372	2.766	3.501	30.112
5	1.857	2.350	34.723	2.123	2.688	32.799
6	1.441	1.824	36.546	1.956	2.476	35.275
7	1.134	1.435	37.981	1.799	2.277	37.552
8	1.054	1.335	39.316	1.394	1.764	39.316

Extraction Method: Principal Axis Factoring.

Table 5.2 The rotated factorial structure

8 factors extracted	(number of) linguistic features with a factor loading larger than 0.30
Factor 1 (37)	present tense (.792), contractions (.787), shorter words (.731), be as main verb (.697), discourse particles (.684), demonstrative pronoun (.649), first person pronoun (.645), second person pronoun (.607), pro-verb do (.571), independent clause coordination (.543), analytic negation (.537), subordinator that deletion (.513), pronoun it (.500), emphatics (.479) existential there (.429), causative adverbial subordinators (.426), wh-pronouns (.417), hedges (.408), predicative adjectives (.389), wh-clauses (.383), public verbs (.379), demonstratives (.324), total adverbs (.304), average word length (-.897), longer words (-.831), average sentence length (-.706), total prepositional phrases (-.703), total other nouns (-.623), attributive adjectives (-.608), past participial WHIZ deletion relatives (-.581), phrasal coordination (-.571), nominalization (-.539), present participial clauses (-.529), conjuncts (-.456), lexical density (-.434), determiner 'the' (-.402), split auxiliaries (-.317)
Factor 2 (15)	that relative clauses on object position (.647), amplifiers (.613), that relative clauses on subject position (.574), first person pronoun (.524), that adjective complements (.486), demonstratives (.439), private verbs (.431), sentence relatives (.431), standardized type-token ratio (.375), wh-pronouns (.355), place adverbials (.323), conditional adverbial subordinators (-.437), analytic negation (-.436), second person pronouns (-.322), existential 'there' (-.310)
Factor 3 (12)	split auxiliaries (.590), conjunct (.489), total adverbs (.485), downtoners (.462), time adverbials (.419), concessive adverbial subordinators (.359), perfect aspect (.353), synthetic negation (.339), that verb complements (.309), contractions (-.408), top 10 coverage (-.321), independent clause coordination (-.315)
Factor 4 (9)	third person pronoun (.533), wh-pronoun (.477), wh relative clauses on subject position (.377), be as main verb (.372), nominalization (-.466), gerunds (-.408), lexical density (-.319), preposition or subordinating conjunction (-.313), longer words (-.302)

Factor 5 (4)	lexical density (.460), top 10 coverage (-.776), determiner 'the' (-.669), shorter words (-.485)
Factor 6 (5)	suasive verbs (.588), public verbs (.434), predictive modals (.341), that verb complements (.341), standardized type-token ratio (-.403)
Factor 7 (4)	coordinating conjunction (.352), total other nouns (-.507), possessive endings (-.361), lexical density (-.340)
Factor 8 (3)	pied-piping relative clauses (.326), average sentence length (.317), particles (-.371)

Table 5.2 presents the rotated factorial structure of the 79 linguistic features along eight dimensions. One of the advantages of the rotated factor extraction is that, "each linguistic feature tends to load on only one factor, and each factor is characterized by those relatively few features that are most representative of the underlying construct" (Biber, 1988, p. 104). Conforming to the results reported in Figure 5.1 and Table 5.1, factor 1 captures the largest number of co-occurring linguistic features (i.e., 37 out of 79), carrying different weights of loadings, followed by factor 2 and factor 3, while factor 8 has only 3 linguistic features loaded on it with relatively small weight of loadings (absolute value less than .40). To make sense Table 5.2, interpretation of the eight factors as textual dimensions has been carried out as reported in section 5.1.1.

5.1.1 Interpretation of the factors as textual dimensions

Interpretation of the factors as textual dimensions is based on the assumption that "a cluster of features co-occur in text because they are serving some common function in those texts" (Biber, 1988, p. 91). Such an interpretation is crucially important since it provides "the foundation for determining the function(s) underlying a set of features" (ibid., p.101). However, as emphasized also by Biber (1988, p. 92), "while the co-occurrence patterns are derived quantitatively through factor analysis, interpretation of the dimension underlying a factor is tentative and requires confirmation, similar to any other interpretative results." Several things should be borne in mind when it comes to the interpretation of the rotated factorial structure. First of all, the interpretation is only tentative and requires further confirmation. Second, interpretation of the factors as textual dimensions is based on the "assessment of the communicative functions most widely shared by the features" (ibid., p.87), and these communicative functions are interpreted

based on previous literature on linguistic variation between spoken and written registers. Third, since the linguistic features grouped on each factor have different weights of factor loadings, "greater attention is given to those features with the largest loading" during the interpretation (ibid.). By saying "the largest loading", Biber (1988) means the absolute value of the factor loading, be it positive or negative. In addition, since positive and negative loadings show groups of features that are distributed in text "in a complementary pattern", while interpreting factors as dimensions, "both the negative and positive cluster of features must be taken into consideration" (ibid., p.88). In the following parts, interpretations of the eight factors are done one by one.

5.1.1.1 Interpretation of factor 1

As shown in Table 5.2, factor 1 is the most powerful factor representing the linguistic variation across native speech, simultaneous interpreting, and written translation. Out of 37 co-occurring linguistic features, 23 have an absolute value of factor loading equal to or larger than 0.50, suggesting those features are highly representative of the underlying construct. A close examination of the 37 linguistic features reveals that most of them overlap with the identified features (with both positive and negative loadings) in the first Dimension ('Involved vs. Informational Production') categorized in Biber (1988). For example, along the positive side of the dimension, factor 1 is characterized by linguistic features associated with more affective, involved, fragmented, and informal language use produced under high time constraints. The use of contractions, for instance, is most representative for informal usage, especially in spoken language. Shorter words (less than three letters), according to Zipf (1949), often carry a general meaning rather than an informational focus. The use of present tense, demonstrative pronouns, first and second pronouns, and pronoun *it* indicates 'on-line' production as well as a high degree of involvedness. Subordinator-that deletion (or 'that' omission), apart from being a possible indicator for decreased explicitness in mediated language, is also closely related to an informal style of the texts, as argued by Quirk et al. (1985) and Biber (1995).

The negative loadings reflect high information density and "a careful integration of information in a text" (Biber, 1988, p. 104) that are not time-constrained, such as average word length, longer words (more than seven letters), average sentence length, prepositional phrases, nouns, adjectives, phrasal coordination, etc., as reported also in Biber (1986, 1988) and X. Hu et al. (2016). Given the great similarity between the current research and Biber (1988) with respect to factor 1, it is decided that Dimension 1 is labelled as 'Involved versus Informational Production'.

5.1.1.2 Interpretation of factor 2

Factor 2 shows a clearly distinct pattern compared with Biber's (1988) Dimension 2 – 'Narrative versus Non-narrative Concerns'. However, 15 linguistic features (11 positive loadings and 4 negative loadings) loaded on this construct overlap partially with those features categorized in Biber's (1988) Dimension 6 – 'On-line Information Elaboration', such as subordination features represented by 'that' relative clauses on object position and 'that' clauses as adjective complements, and demonstratives associated with "informal, unplanned types of discourse" (p.113). The nature of 'On-line Information Elaboration', as rightfully explained by Van Rooy et al. (2010), is that it "captures exactly such a tension between informational density, which results from preparation, and on-line production strain, which results whenever the prepared speech isn't read verbatim but represented from notes and thus subjected to reformulation under time pressure" (p.343). Therefore, different from the 'Informational Production' in Dimension 1 which emphasizes planned and integrated informational production, the co-occurring features in factor 2 highlight more real-time constraints, which result in "a fragmented presentation of information accomplished by tacking on additional dependent clauses, rather than an integrated presentation that packs information into fewer constructions containing more high-content words and phrases" (p.113). Another important function shared by these grouping features is the expression of personal stance, which has also been reported in Biber (1988). For instance, relative clauses, which are the non-essential parts of a sentence, provide more explicit and elaborate information concerning the referents in a text. First-person

pronouns, “markers of ego-involvement in a text”, are often correlated with cognition verbs (such as private verbs) to indicate “discussion of mental processes” (Biber, Appendix II, p.225). The co-occurrence of first-person pronouns and private verbs, with these relativization features indicates strongly that they may be used to express the speaker’s/writer’s value judgement. This may be further confirmed by the use of sentence relatives, which, according to Biber et al. (1999), “are most commonly used to convey an attitude or value judgement about a proposition” (p.867). Amplifiers are also often used to express strong feelings or views of the speaker/writer, and the full form ‘that’ adjective complements, as analyzed in Chapter four, “can be used for elaboration of information relative to the personal stance of the speaker” (ibid., p.114). In explaining the correlation between on-line information elaboration and stancetaking expression, Biber (1988) argues that this “indicates that those discourse tasks which involve the explicit marking of an individual’s stance are frequently also tasks that demand informational production under real-time constraints” (p.160). The negative features with relatively high loadings (conditional adverbial subordinators, and analytic negation), by contrast, indicate a more objectified point of view. To fully capture the communicative functions of these grouping features, the author tentatively names Dimension 2 as ‘On-line Information Elaboration with Stancetaking Concerns’.

5.1.1.3 Interpretation of factor 3

Factor 3 captures 12 linguistic features, with one feature carrying a factor loading closer to 0.60 (split auxiliaries, 0.590) while four larger than 0.40 (conjuncts, total adverbs, downtoners, and time adverbials). Based on Nini (2014), split auxiliaries are identified every time an auxiliary is followed by one or two adverbs and a verb form, in which case the co-occurrence of split auxiliaries and adverbs is not surprising. The use of split auxiliaries has not received enough attention from traditional grammarians (see also Biber, 1988), but according to Biber (1988), they are “more common in certain written genres than in typical conversation” (Appendix II, p.244). In a previous study, he (1986) finds that split auxiliaries often co-occur with linguistic features with strong informational

focus such as passives, nominalizations, and prepositions. The use of conjuncts, as discussed in Chapter four, is closely related to an increased explicitness of logical relations, and is found to be more frequent in written texts than spoken texts (Altenberg, 1986; Biber, 1986; Ochs, 1979). The co-occurrence patterns of total adverbs, downtoners and concessive adverbial subordinators have also been captured in Biber's (1988) Dimension 7, which he tentatively names as 'Academic Qualification or Hedging', since these features are used to "qualify the extent to which an assertion is 'known' in academic discourse" (p.114). However, due to the smaller factor loadings of these three features in Dimension 7, Biber discards this dimension in the end. In this study, the factor loadings for these features are relatively high, which suggest they have higher weight in accounting for the underlying construct, along with split auxiliaries, conjuncts, and time adverbials. When these features are taken together, they all point to more elaborated and precise description.

The negative features, although being rather weak representation of this dimension, indicated by their small factor loadings, index reduced and simplified language use, such as contractions and top 10 vocabulary coverage. Therefore, Dimension 3 is tentatively labelled as 'Precise versus Simplified Description'.

5.1.1.4 Interpretation of factor 4

Nine linguistic features group together in factor 4, four having positive weights and five negative weights. The four positive loadings seem to have a narrative focus, represented by features such as third person pronoun, WH-pronoun, and WH relative clauses on subject position. Third person pronouns "mark relatively inexact reference to persons outside of the immediate interaction" (Biber, 1988, Appendix II, p.225). They are often reported to co-occur with past tense and perfect aspect "as a marker of narrative, reported (versus immediate) styles" (ibid.). The use of WH-pronouns also implies a reported style, in that something or somebody under discussion is not in the immediate context. As far as WH relative clauses on subject position is concerned, Ochs (1979) argues that relatives

are often used for more exact and explicit references in planned discourse (see also Biber, 1988). All three linguistic features with high positive loadings point to referents that may not be present in the immediate context. The negative features of factor 4 are all nominal features demonstrating high information density, but only two out of the five features (i.e., nominalization and gerunds) have relatively high factor loadings while the other three (i.e., lexical density, preposition, and longer words) have factor loadings only slightly higher than the threshold 0.30. Biber (1988) explains that, when interpreting factors as dimensions, greater attention should be given to features with the largest loading, and in this case, nominalization and gerunds, which "have been particularly been taken as markers of conceptual abstractness" (Biber, 1988 Appendix 11, p. 225). Dimension 4 is thus interpreted as 'Narrative versus Abstract focus'.

Before looking into the remaining four factors (factor 5, 6, 7, and 8), one thing worth mentioning is that, Biber (1988, p. 88) argues that "[i]n general, five salient loadings are required for a meaning interpretation of the construct underlying a factor". In that case, the eight-factor solution determined in this study can be excluded as "over-factoring" since three out of the eight factors (i.e., factor 5, 7, and 8) have less than five salient loadings. However, the author decided to tolerate at least four salient loadings in this research, as following Biber's suggestion would mean that factor 5, 7, and 8 all being excluded, which may mask the nuances of the differences across the three language varieties. Based on the new criteria, a seven-factor solution is thus retained.

5.1.1.5 Interpretation of factor 5

Factor 5 is characterized by one linguistic feature with positive weight while three with negative weights. Lexical density, as discussed in Chapter three and four, calculates the proportion of lexical words to total running words (or sometimes, to grammatical words), to indicate the informativeness of the text. While the other three features, i.e., top 10, determiner 'the', and shorter words (less than 3 letters), all point to a functional concern. So factor 5, or Dimension 5, is named as 'Lexical versus Functional Concern'.

5.1.1.6 Interpretation of factor 6

Factor 6 reveals a much-discussed lexical pattern of translation, that is, optional 'that' after reporting verbs, although reporting verbs carry more factor weight than 'that' usage. It shows clearly that the use of 'that' verb complements co-occurs frequently with suasive and public verbs. According to Biber (1988), "suasive verbs imply intentions to bring about some change in the future (e.g., *command*, *stipulate*)", while public verbs "involve actions that can be observed publicly; they are primarily speech act verbs, such as *say* and *explain*, and they are commonly used to introduce indirect statements" (Appendix II, p.242). The other co-occurring feature, i.e., predicative modals, also indicates things that will happen in the future. It seems that speakers/writers rely more often on these co-occurring features to persuade the audience/readers by predicting the possible outcomes in the future. The marked underuse of one salient negative weight, that is, STTR, offers a possible explanation that these persuasive moves are made under real time constraints. Dimension 6 is thus tentatively labelled as 'On-line Persuasion'.

5.1.1.7 Interpretation of factor 7

Factor 7 has four linguistic features loaded on it, including coordinating conjunctions, total other nouns, possessive endings, and lexical density, and their factor loadings are moderately high (>=0.34). The positive feature of coordinating conjunctions identifies conjunctions such as *and*, *but*, *so*, *either*, *or*, which serve to coordinate or conjoin two sentences, phrases, or words, cueing more coordinated language use. The other features with negative weights, i.e., total other nouns, possessive endings, and lexical density, index more compact information that highlights a possessive relationship, as exampled in "people's income" (WT_A2_01), "patients' conditions" (WT_B3_01), "a person's talent" (SI_A2_07), et cetera. Dimension 7 is thus interpreted as 'Coordinating versus Possessive Functions'.

5.1.1.8 Summary of the textual dimensions

Based on Table 5.2, seven factors are found to have strong factorial structures carrying more than four linguistic features. Factor 1 captures the largest variance, followed by factor 2 and 3. Based on the different co-occurring linguistic features sharing certain communicative functions, the author has given interpretative labels to each of the factors as dimensions.

Dimension 1 is labelled as 'Involved versus Informational Production', following the original label in Biber (1988). The positive end of this pole targets at an interactive, involved, informal, and reduced language production, while the negative end marks an informational, integrated, and formal language production. Biber (1988) argues that this dimension is "the most fundamental parameter of variation" in both texts of English and texts of other languages (Biber, 1988, p. 115). In a later study, he (2010) renames this dimension as "Clausal versus Phrasal", arguing that this might be a universal dimension underlying other register comparisons as well.

Dimension 2, named as 'On-line Information Elaboration with Stancetaking Concerns', distinguishes texts that are produced under strict time constraints aiming for stancetaking elaboration from those that are more reserved in terms of stance expressions.

Dimension 3, labelled as 'Precise versus Simplified Description', captures the difference between texts that are more formal, explicit, and precise in terms of language usage, versus texts that rely more on reduced and simplified language description.

Dimension 4, interpreted as 'Narrative versus Abstract Focus', distinguishes texts with different focuses: the features on the positive end of this dimension are associated with a narrative, or reported focus, in that the subjects under discussion are often not in the immediate context; while the features on the opposite end point to more abstract description highlighted by nominalizations and gerunds.

Dimension 5 is labelled 'Lexical versus Functional Concerns', and this dimension is rather straightforward as suggested by the name. Dimension 6, 'On-line Persuasion', captures optional 'that' verb complements associated with suasive and public verbs. The last dimension, 'Coordinating versus Possessive Functions, distinguishes coordinated texts from texts of compact information highlighting possessive relations.

5.1.2 Textual relations in NS, SI and WT

5.1.2.1 Factor scores and textual relations

The primary goal of this research is to examine the variation of linguistic patterns across native speech, simultaneous interpreting, and written translation, with a specific aim to isolate linguistic patterns specific to simultaneous interpreting (into B). One way to achieve this goal is to compute factor scores for each dimension. Once the factor score or dimension score has been computed for each text (or text segment), "the similarities and differences among genres (the textual 'relations') can be analyzed with respect to these scores to support or refute hypothesized interpretations" (Biber, 1988, p. 92).

Computation of factor scores is done by summing the number of the standardized score[33] of the linguistic features having salient loadings larger than 0.35 on that factor (Biber, 1988). When a linguistic feature has been loaded on more than one factor, there are two ways to deal with this. One is advocated by Biber (1988), following Gorsuch's (1983) suggestion, to compute only the one with the largest weight "to assure the experimental independence of the factor scores" (p.93). The other way is taken by X. Hu et al. (2016) to take into consideration every linguistic feature regardless of their smaller or larger weights, since they believe it is "unwise to prejudge the issue of whether or not a feature might be important to more than one dimension" (p.20). The present author, however,

[33] The standardization procedure has been expounded in detail in Biber (1988, pp. 93–94). To calculate the standardized score of a linguistic feature, information should be known as regards the mean value, standard deviation, as well as the raw frequencies of that linguistic feature. For instance, if a text had a frequency of 113 past tense verbs, its mean value being 40.1 and standard deviation 30.4, the standardized score would be calculated using the following formula: the standardized score = (raw frequency – mean value) / standard deviation, which is (113-40.1)/30.4=2.4. The factor score would be calculated by summing all standardized scores of the linguistic features having salient loadings (larger than 0.35) on that factor.

does not concur with X. Hu et al's (2016) argument since the importance of one linguistic feature along the different dimensions can already be determined by its factor loadings, as explicitated by Biber (1988). However, she does believe that it is problematic to exclude linguistic features that have relatively lower weights, since when Biber (1988) interprets the factors as textual dimensions from a functional perspective, he does not exclude those features with lower weights when they are loaded on more than one dimension. Therefore, it seems unwise to exclude them in the calculation of factor scores. Following this line of thought, the author decided to adopt X. Hu et al.'s (2016) approach for factor score calculation. That is, when a linguistic feature has loaded on more than one dimension, as long as it has a salient loading larger than 0.35 (the original threshold for factor interpretation is 0.30), it will be computed for factor score calculation.

Based on the standardization procedure, the factor scores for all 477 text segments in each dimension have been calculated to investigate the linguistic variations across the three language varieties, as shown in Table 5.3.

Table 5.3 Descriptive dimension statistics for the three language varieties

Dimension 1: 'Involved versus Informational Production'
Dimension 2: 'On-line Information Elaboration with Stancetaking Concerns"
Dimension 3: 'Precise versus Simplified Description'
Dimension 4: 'Narrative versus Abstract Focus'
Dimension 5: 'Lexical versus Functional Concerns'
Dimension 6: 'On-line Persuasion'
Dimension7: 'Coordinating versus Possessive Functions'

Dimension	Mean	Minimum Value	Maximum Value	Range	Standard Deviation
…….. NS…….					
Dimension 1	1.32	-7.94	13.61	21.55	4.27
Dimension 2	6.85	-3.77	17.20	20.97	3.45
Dimension 3	-1.18	-6.58	7.83	14.41	2.47
Dimension 4	1.29	-3.85	6.87	10.72	2.05
Dimension 5	-0.93	-4.93	3.72	8.65	1.51
Dimension 6	0.38	-2.69	4.26	6.95	1.29
Dimension 7	-0.80	-4.18	2.51	6.69	1.39
……SI……..					
Dimension 1	2.55	-10.91	20.37	31.28	6.39
Dimension 2	-2.33	-13.31	4.80	18.11	3.69
Dimension 3	-1.53	-7.67	6.31	13.98	2.92

Dimension 4	-0.28	-4.91	7.64	12.55	2.07
Dimension 5	0.51	-6.74	6.72	13.46	2.11
Dimension 6	-0.56	-4.88	4.73	9.61	1.53
Dimension 7	-0.20	-3.85	6.36	10.21	1.85
……..WT……..					
Dimension 1	-3.23	-12.75	8.63	21.38	4.00
Dimension 2	-2.83	-14.05	6.12	20.17	3.77
Dimension 3	2.15	-7.32	11.42	18.74	3.76
Dimension 4	-0.59	-5.79	8.17	13.96	2.04
Dimension 5	0.26	-5.60	6.75	12.35	1.97
Dimension 6	0.16	-5.27	8.78	14.06	1.84
Dimension 7	0.73	-3.43	6.43	9.86	1.77

Table 5.3 presents the factor scores for each dimension of the three language varieties, including the mean score, minimum and maximum value, the range (that is the score difference between the maximum and the minimum value), as well as standard deviation. For example, the mean score of Dimension 1 (1.32) for NS indicates that native speech is more marked in the use of linguistic features that have positive loadings, such as present tense, contractions, shorter words, discourse particles, pronouns, etc. that are associated with unplanned, informal and involved language production, while unmarked in the use of negative features associated with integrated and informational language production. The maximum score of 13.61 indicates strongly the involved and informal nature of certain text segment within NS, while the minimum value of -7.94 points to the opposite formal and informational end of the continuum. The standard deviation shows whether the factor score of each text segment within one language variety groups around the mean score. As far as the standard deviation (4.27) for Dimension 1 of NS is concerned, it shows that the factor scores of the text segments within NS are scattered around the mean score, indicating great variation within NS subgenres.

Viewed from a holistic perspective, Table 5.3 offers more information than can be explicitated here. For a start, there are great internal variations within the three language varieties (NS, SI, and WT) along all seven dimensions, with Dimension 1 – 'Involved versus Informational Production' revealing the largest variation. This is particularly noticeable for SI in which the range of variation amounts to 31.28, ranging from -10.91 showcasing a strong informational focus to 20.37 cueing strongly involved and informal

production. While the range of Dimension 1 for SI is determined by two extreme text segments (the minimum and the maximum value), the high standard deviation (6.39) adds further evidence to the great variation within SI subgenres. Second, mediated texts (SI and WT) demonstrate overall higher internal variation than unmediated texts (NS) along the seven dimensions. Such heterogeneity of the distributional patterns of linguistic features in mediated texts may offer counterevidence to the "levelling-out" (Baker, 1996) or "convergence" hypothesis which claim "relatively higher level of homogeneity of translated texts" (Laviosa, 2002, p. 72). As far as SI and NS data are concerned, it does not lend support to this hypothesis along all seven dimensions, as interpreted texts are more varied (i.e., higher standardized deviation) than non-interpreted originals in each dimension. In addition to the two observations, comparison between SI and WT in terms of the maximum and minimum value along the seven dimensions also seems to suggest that under some extreme situations, interpreted texts may not differ from translated texts as has been expected, which may have implications for the shared constraints of translation and interpreting as forms of constrained/mediated language. These initial observations suggest the complexity of language production, in that for each language variety (mediated or not), there seems to be no absolute uniform patterns among different subsets or subgenres within that language variety, as reported also in Biber (1988) with respect to the linguistic variation between (various kinds of) spoken and written registers.

Table 5.4 presents overall F and correlation values for each dimension across NS, SI, and WT utilizing General Linear Model procedures following Biber (1988). The aim for such tests is to see whether the differences of dimension scores of the three language varieties are statistically significant, and how strong the predictive power of each dimension is in distinguishing them. The F value, based on Biber (1988, p. 126), "is a test of statistical significance, indicating whether a dimension can distinguish among genres to a significant extent". The p (probability) value indicates the probability that F value is significant. R*R values show the predicative power of each dimension as they "directly indicate the percentage of variance in the dimension scores that can be predicated by knowing the genre distinctions" (ibid.). Based on the different values presented in Table

5.4, all seven dimensions can distinguish the three varieties to a statistically significant extent. However, the predictive powers of these dimensions vary substantially, with only one dimension having a R*R value larger than 50%, while two having R*R values larger than 20%. In Biber's (1988) study, four out of the six dimensions featuring larger-than-50% R*R values, while the lowest value is about 17%, which "is still large enough to be noteworthy" (p.127). Based on the different R*R values, it seems that Dimension 2 ('On-line Information Elaboration with Stancetaking Concerns') is the most powerful predictor among the three varieties, meaning that there is a strong correlation between (sub-)genre distinctions and the values of dimension scores. This is followed by Dimension 3 ('Precise versus Simplified Description'), and Dimension 1 ('Involved versus Informational Production'). While the remaining dimensions reveal statistically significant differences across the three varieties, they are not strong predictors since only a small portion of the variation in values for these dimension scores can be accounted for by knowing the language variety categories (or subgenre categories within) of texts.

Table 5.4 F and correlation scores for the seven textual dimensions

Dimension 1: 'Involved versus Informational Production'
Dimension 2: 'On-line Information Elaboration with Stancetaking Concerns'
Dimension 3: 'Precise versus Simplified Description'
Dimension 4: 'Narrative versus Abstract Focus'
Dimension 5: 'Lexical versus Functional Concerns'
Dimension 6: 'On-line Persuasion'
Dimension7: 'Coordinating versus Possessive Functions'

Dimension	F value	Probability (p)	R*R
1	65.298	.000	21.6%
2	328.923	.000	58.1%
3	70.854	.000	23%
4	32.435	.000	12%
5	23.766	.000	9.1%
6	14.126	.000	5.6%
7	34.012	.000	12.6%

The much less strong correlation (indicated by R*R values) between genre distinctions (i.e., three language varieties, and their subgenres) and the values of dimension scores in comparison with Biber (1988) is not surprising, since his study covers a large number of

situational settings that serve for different purposes, while the current research focuses on texts produced under similar legislative settings but are of different mediation status. Interpretation of this table cannot be done without considering the actual positioning of the three varieties along the seven dimensions, which turns out to be more revealing evidenced in the following sections.

5.1.2.2 Relations along dimensions

Section 5.1.2.1 presents the full descriptive statistics of the factor scores of the three language varieties, but it is not straightforward to see how the three language varieties are similar to or differ from each other along the seven textual dimensions. One effective way to determine such similarity or variation is to visualize the mean factor or dimension scores for each language variety, as presented in the following figures (Figures 5.2 to 5.8), based on the statistics presented in Table 5.3. The same method can also be used to determine genre variations across the nine subsets/subgenres (see section 5.1.3). Three key matters need to be considered while interpreting the following figures, following Biber's advice (1988, p. 129), which include: (1) the similarities and differences among the language varieties as well as the subgenres with respect to their mean factor/dimension scores; (2) the co-occurring features underlying the dimension under discussion, including both features with positive weight and those with negative weight; and (3) the underlying functions these features serve for, as interpreted in section 5.1.1.

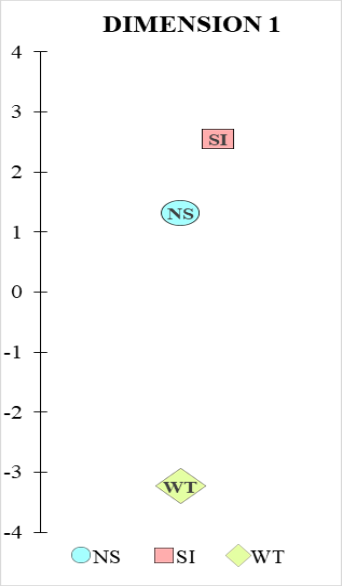

Figure 5.2 Mean scores of Dimension 1 for NS, SI, and WT

Dimension 1: 'Involved versus Informational Production'

(Note*: NS = native speech; SI = simultaneous interpreting into B; WT = written translation into B)

Figure 5.2 plots the mean scores of Dimension 1, 'Involved versus Informational Production', for the three English varieties. A high positive score indicates involved and informal language use produced 'on-line', characterized by co-occurring linguistic features such as present tense, contractions, shorter words, discourse particles, demonstrative pronouns, first and second pronouns, subordinator-that deletion, emphatics, etc.; while a high negative score points to carefully integrated and informational language production, marked by frequent co-occurrences of average word length, longer words, prepositional phrases, total other nouns, adjectives, phrasal coordination, nominalization, and others. Figure 5.2 shows that, both NS and SI as forms of (un/mediated) spoken discourse are marked in 'involved production', while WT, as a form of mediated written discourse, situates at the 'informational production' end of the continuum. In terms of the 'involvedness', SI is more marked than NS, evidenced by its higher mean dimension score. The reasons for the more marked involvedness of interpreted texts might be complex, as will be discussed later. Nonetheless, it seems that overall Dimension 1 tends to be a potential candidate for isolating interpreted language from both unmediated spoken language and translated language.

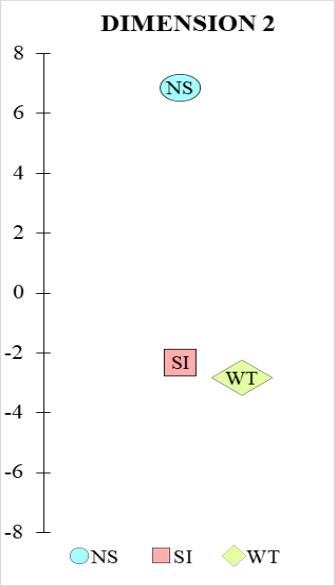

Figure 5.3 Mean scores of Dimension 2 for NS, SI, and WT

Dimension 2: 'On-line Information Elaboration with Stancetaking Concerns'

(Note*: NS = native speech; SI = simultaneous interpreting into B; WT = written translation into B)

Figure 5.3 plots the mean scores of Dimension 2, 'On-line Information Elaboration with Stancetaking Concerns', for NS, SI, and WT. A high positive score indicates that the texts are informational, but produced under real time constraints, while a negative score reveals a "conditioned" description. The informational focus, as explained in section 5.1.1, emphasizes fragmented presentation of information that is added 'on-line', manifested in the use of subordination features associated often with stancetaking concerns, such as 'that' relative clauses on object position (e.g., the dog that bit me), 'that' relative clauses on subject position (e.g., the dog that I saw), 'that' adjective complements (e.g., I'm glad that you like it), and sentence relatives, along with other linguistic features highlighting immediate referents, such as first person pronouns, demonstratives, and private verbs. The opposite end of the continuum groups four features while only two have relatively larger weight, i.e., conditional adverbial subordinators and analytic negation. Figure 5.3 shows that native speech is extremely marked for its on-line expression of personal stances given its very high mean dimension score, interpreting and translation, by contrast, are unmarked for this dimension. As a matter of fact, SI does not differ much from WT in this dimension, based on their mean dimension scores. One possible reason might be

the translators' and interpreters' risk-management strategy to remain neutral and refrain from stancetaking (Pym, 2005). Another possible contributor might be the bi-mediation status of translation and interpreting (Translated, and L2/non-native) compared with the unmediated status of native speech. R. Xiao 's study (2009) on the linguistic variation of different English varieties reveals that the 'on-line information elaboration' dimension can readily distinguish native from non-native varieties of English, as the latter are unmarked in the use of syntactically elaborated features due to socio-cultural and language acquisition issues. In this study, a similar trend has also been identified, although the reasons behind are much more complex as translation and interpreting are not only non-native, but most importantly, mediated. Nonetheless, D2 captures very noticeable differences between SI and NS from a comparable perspective, but texts of different mediation modes are not very distinguishable.

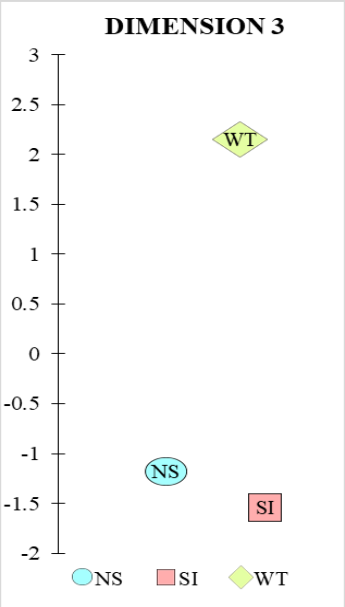

Figure 5.4 Mean scores of Dimension 3 for NS, SI, and WT

Dimension 3: 'Precise versus Simplified Description'

(Note*: NS = native speech; SI = simultaneous interpreting into B; WT = written translation into B)

Figure 5.4 plots the mean scores for Dimension 3, 'Precise versus Simplified Description', for the three varieties. A high positive score means that the texts under discussion are characterized by frequent use of co-occurring linguistic features indexing more precise

and explicit description, such as split auxiliaries, conjuncts, total adverbs, downtoners, time adverbials, etc., while a high negative score marks simplified language, featured by contractions, top 10 vocabulary coverage, and independent clause coordination. Different from Dimension 2, interpreted texts in this dimension resemble non-interpreted target originals in that they are both characterized by simplified language usage and unmarked in precise and elaborate description. By contrast, there is noticeable distinction between interpreting and translation as different modes of mediation. This finding is not surprising, and it has offered further evidence to the 'oral' features of SI and NS as spoken discourse, and the 'literate' features of WT as written discourse. Viewed from an intermodal perspective, Figure 5.4 also lends some support to the claim that "interpreters simplify more than translators" (Bernardini et al., 2016), indicating a strong influence of modality as argued by by Shlesinger (2008) and testified by Bernardini et al. (2016) and Ferraresi et al. (2018).

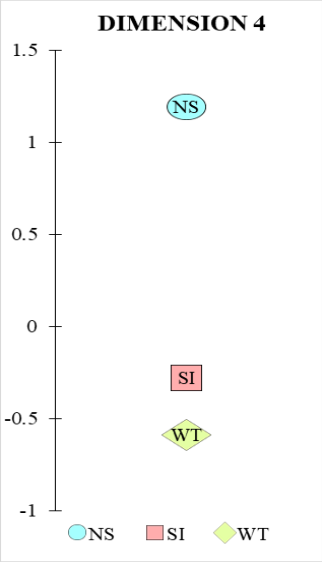

Figure 5.5 Mean scores of Dimension 4 for NS, SI, and WT

Dimension 4: 'Narrative versus Abstract Focus'

Dimension 4, 'Narrative versus Abstract Focus', in Figure 5.5 captures ontological differences between interpretations and non-interpretations, but the difference between interpretations and translations are not very distinguishable. In particular, NS has moderately high positive scores, meaning that it is marked in frequent use of third-person

pronouns, *WH*-pronouns, and *WH*-relative clauses, while infrequent use of nominalizations, gerunds, and lexical density, in relation to SI. This finding confirms partially the previous unidimensional analysis carried out in Chapter four in which non-mediated spoken language is found to be less lexically dense than interpreted language, especially in terms of the distributional patterns of nominal features. In terms of SI (and WT), although they have shown statistically significant difference from NS with respect to the distributional patterns of co-occurring features in D2, given their intermediate dimension scores (approximately ranging from -0.5 to 0.5 based on Biber's description), they are not very marked in either narrative or abstract focus. That is, compared with the more marked narrative focus of NS, SI (and WT) are more or less equally characterized by co-occurring features associated with the two focuses, although WT is slightly more prominent in the use of nominal features, as indicated by its mean dimension score.

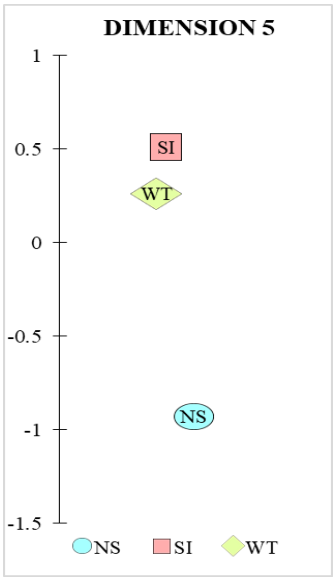

Figure 5.6 Mean scores of Dimension 5 for NS, SI, and WT

Dimension 5: 'Lexical versus Functional Concerns'

(Note*: NS = native speech; SI = simultaneous interpreting into B; WT = written translation into B)

Figure 5.6 lends further support to the slightly more 'lexical' focus of interpretations versus non-interpretations. However, such a lexical focus is not very marked due to the relatively intermediate dimension score of SI, meaning that interpreted texts are not noticeably marked with respect to either lexical or functional concerns. The same can also

be said for translations. In comparison, NS has a moderately high dimension score indexing marked functional focus while unmarked lexical focus, which are featured by relatively more frequent co-occurrence patterns of top10 vocabulary coverage, determiner 'the', and shorter words (less than 3 letters), while less frequent use of lexical density. Dimension 5, in general, distinguishes SI from NS, but not from WT, and their overall differences are not as equally noticeable as in other dimensions considering the narrower range of dimension score differences.

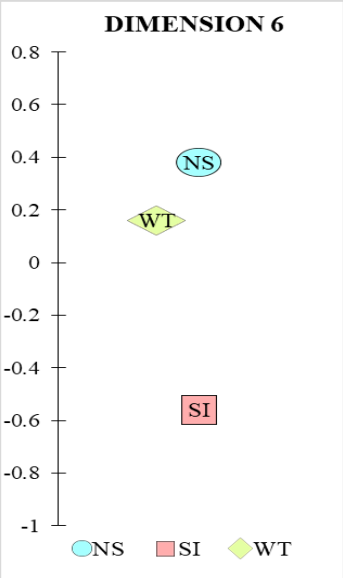

Figure 5.7 Mean scores of Dimension 6 for NS, SI, and WT

Dimension 6: 'On-line Persuasion'

(Note*: NS = native speech; SI = simultaneous interpreting into B; WT = written translation into B)

Figure 5.7 isolates SI from both NS and WT, but their overall variation patterns are generally undistinguished due to the relatively narrower range of the mean dimension scores. As far as WT and NS are concerned, there is no clear characterization as to their persuasive or non-persuasive preference, while SI, in comparison, is very slightly unmarked in terms of the persuasive focus, featured by less frequent use of suasive verbs, public verbs, 'that' verb complement, and predictive modals, while slightly more frequent use of standardized type-token ratio. Upon first reflection, these co-occurrence patterns in SI versus NS and WT seem to offer counterevidence to the linguistic patterns identified through a unidimensional analysis (Chapter four), which reports statistically significant

underuse of STTR and overuse of 'that' verb complement in SI from a comparable perspective, and statistically significant underuse of STTR and 'that' verb complement in SI from an intermodal perspective. The author defends that interpretation of the variation patterns needs to be done against the shared functions these grouping features serve for, rather than examine separately individual linguistic features. In addition, emphasis should be given to those features with higher factor loadings during interpretation, which, in this case, are suasive and public verbs associated with persuasive concerns expressed under real time constraints. Given the slightly unmarked nature of SI in this dimension, it may well be the result of the interpreters' risk-avoidance consideration to remain neutral.

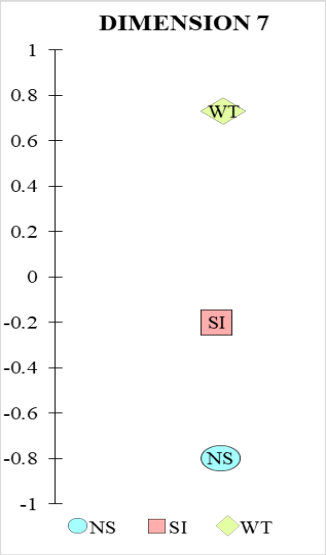

Figure 5.8 Mean score of Dimension 7 for NS, SI, and WT

Dimension 7: 'Coordinating versus Possessive Functions'

(Note*: NS = native speech; SI = simultaneous interpreting into B; WT = written translation into B)

Figure 5.8 also presents relatively intermediate positioning of all three language varieties along this dimension, with WT being slightly more marked in the use of coordinating conjunctions, and unmarked in the co-occurrence of total other nouns, possessive endings, and lexical density for possessive relations, while NS situating towards the opposite end. SI, in comparison, does not show any marked usage of the two sides. That is, SI resorts to a balanced use of both coordinating conjunctions, and the co-occurring nouns, possessive endings, and lexical density. The slightly more frequent use of coordinating

conjunctions, such as *and*, *so*, *but*, in WT offers counterevidence to Chafe and Danielewicz's study (1986, p. 103) which reports overall more frequent use of coordinating conjunctions in spoken rather than written language so as to avoid "the more elaborate interclausal relations". Previous Mann-Whitney U test carried out in Chapter four, however, fails to reveal any statistical significance between SI and WT in coordinating conjunctions. Interpretation of the results, as emphasized here again, needs to be done based on the complementary co-occurring features along this dimension. While coordinating conjunctions are often associated with spoken language, the use of possessive forms instead of prepositional ones (such as *of*) also mark a strong informal focus, as found to be the case in NS.

Based on the above analysis, it seems that there is no single, absolute difference among the three English varieties, despite ontological or modality differences. More often than not, SI resembles either NS or WT to varying extents, as shown in Figures 5.2 to 5.8. Along certain dimensions, the three varieties are relatively undistinguished given their relatively intermediate mean dimension scores, further consolidated by the relatively low R*R values presented in Table 5.4. While the R*R values demonstrate that the first three dimensions have the largest predictive powers among the three varieties, evidenced also by the aforementioned plots and analysis, they fail to show which dimension is the most noticeable discriminator from both comparable (SI vs. NS) and intermodal (SI vs. WT) perspectives. To make up for this deficiency, additional plots visualizing the mean factor/dimension score differences along the seven dimensions are presented below (see Figures 5.9 to 5.11). The underlying assumption is that, "[t]he larger the absolute value of the factor score difference, the greater the difference between the two groups" (X. Hu et al., 2016, p. 24). In X. Hu et al. (2016), the largest factor score difference identified in Dimension 2, which they labelled as 'Translational versus Non-translational', reports a factor score difference of 0.73, and it is considered as the most noticeable discriminator for distinguishing translations from non-translations.

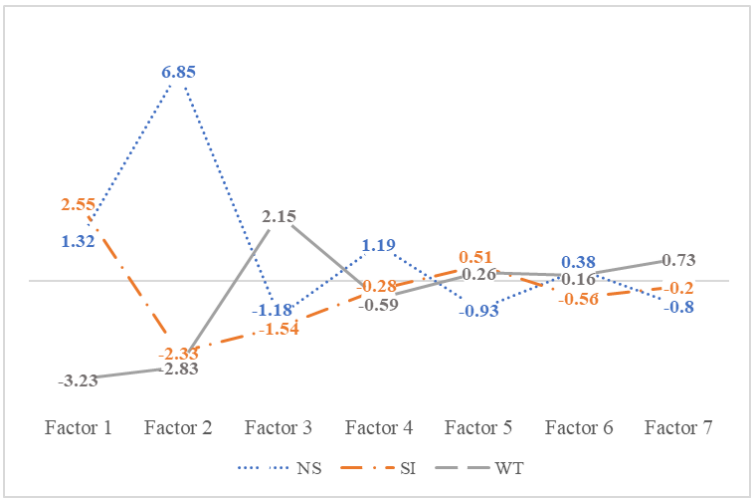

Figure 5.9 Mean factor scores for NS, SI, and WT

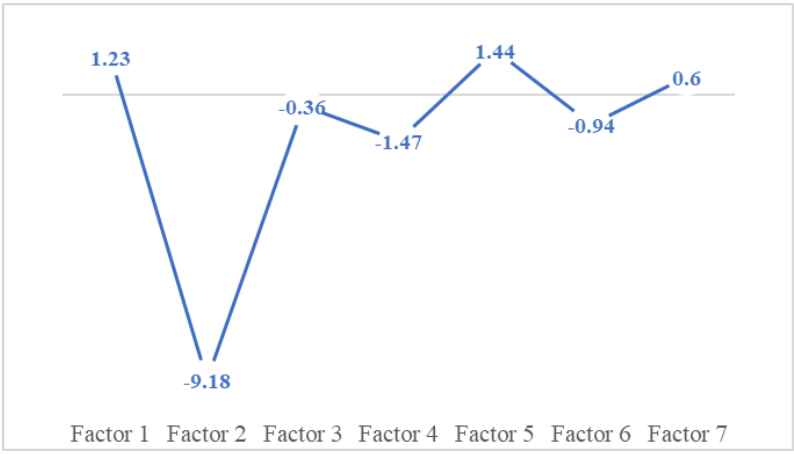

Figure 5.10 Mean factor score differences between SI and NS

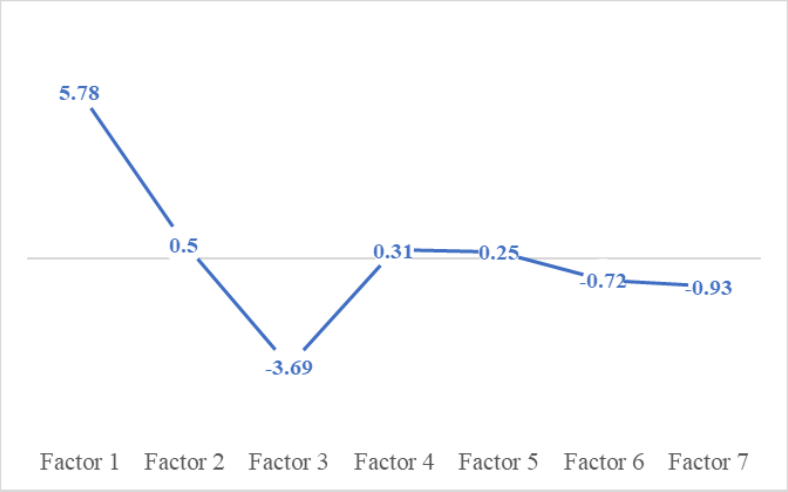

Figure 5.11 Mean factor score differences between SI and WT

Figure 5.9 presents the mean factor/dimension scores for the three language varieties

while Figure 5.10 and 5.11 examine their mean factor/dimension score differences from comparable and intermodal perspectives, respectively. Figure 5.9 shows that, overall, interpretations, translations, and non-interpretations situate differently along the seven dimensions, but there is no absolute difference among the three varieties along many dimensions (such as Dimension 5, 6 and 7). Viewed from either a comparable or an intermodal angle (see Figure 5.10 and 5.11), the absolute values of their factor score differences also vary substantially.

To be more specific, factor 1, or Dimension 1, 'Involved versus Informational Production', captures the largest difference (5.78) between SI and WT, followed by Dimension 3 (3.69), Dimension 7 (0.93), and Dimension 6 (0.72), while their differences along the other dimensions are negligible. In terms of ontological differences between SI and NS, Dimension 2, 'On-line Information Elaboration with Stancetaking Concerns', exhibit the most dramatic variation, with a mean factor score difference reaching 9.18. By contrast, the other dimensions, i.e., Dimension 4, 5, 1, and 6, all reveal variations (1.47, 1.44, 1.23, and 0.94 respectively) to a much lesser degree. In general terms, there seems to be less disparities between SI and WT than between SI and NS along many dimensions, as suggested by the mean factor/dimension score differences. This finding contradicts previous unidimensional analysis which reveals greater resemblance between SI and NS as (un/mediated) forms of spoken discourse, with respect to linguistic variation of certain linguistic indicators such as conjuncts (CONJ) and total adverbial subordinators (ASUB) reported in Chapter four. This seemingly unexpected finding about the greater resemblance between translation and interpreting highlights the different perspective offered by a multidimensional analysis that a unidimensional analysis fails to uncover. Meanwhile, it may also have implications for the shared nature of mediated (and also constrained) language varieties. Before jumping into a conclusion, a more fine-grained analysis in terms of consistency patterns with respect to the more (un-)marked features of SI along the seven dimensions has been carried out in section 5.1.3 to investigate possible genre influence as well as genre variations.

5.1.3 Consistency patterns across genre comparisons

It was noted in the previous section that there are great internal variations within the three language varieties, evidenced by the range of dimension scores as well as standard deviations (see Table 5.3). To find out if the distinctive co-occurrence patterns of SI along the seven dimensions are consistent across SI subgenres compared with the corresponding NS and WT subgenres, the author decided to explore the textual relations (or variations) among the nine subsets/subgenres across NS, SI, and WT. As clarified in Chapter three, in the preliminary phase for corpus construction, the author tried to make as comparable as possible the three English components in the LegCo+ corpus, so three pairs of comparable subgenres were included in NS, SI, and WT, including "*Questions to the Prime Minister/Chief Executive*" (A), "*Questions to the Ministers/Secretaries*" (B), and "*Debates*" (C).

Before the exploration, the author emphasizes that a genre influence is identified when cross-subgenre comparison between SI and NS, or between SI and WT reveals different (un-)marked co-occurrence patterns as reported in the above section. Otherwise, it can be said that there is no subgenre influence on the (un-)marked co-occurrence patterns of SI from comparable and/or intermodal perspectives, and these patterns are consistent along the seven dimensions. A distinction is also made here regarding genre variation and genre influence. As the names have suggested, genre variation refers to variation among genres in terms of the distribution/variation patterns of linguistic features, while genre influence deals with the influence of different genres over the consistency of the linguistic patterns specific to SI in general. While it is often expected that there would be genre or subgenre variations within each language variety in terms of the variation patterns of linguistic features, genre influence is not necessarily expected, when the overall patterns are always consistent across subgenre comparisons.

Table 5.5 presents the descriptive dimension statistics for the corresponding subgenres across NS, SI, and WT. Similar to Table 5.3, this Table also include the mean, minimum

and maximum scores, range, and standard deviation of each dimension score for each subgenre/subset. Based on the data presented, it is possible to cross-examine the variation patterns both across corresponding subgenres and within each language variety.

Table 5.5 Descriptive dimension statistics for the corresponding subgenres across NS, SI, and WT

Dimension 1: 'Involved versus Informational Production'
Dimension 2: 'On-line Information Elaboration with Stancetaking Concerns"'
Dimension 3: 'Precise versus Simplified Description'
Dimension 4: 'Narrative versus Abstract Focus'
Dimension 5: 'Lexical versus Functional Concerns'
Dimension 6: 'On-line Persuasion'
Dimension7: 'Coordinating versus Possessive Functions'

Dimension	Mean	Minimum Value	Maximum Value	Range	Standard Deviation
……NS_A…..…..……					
Dimension 1	0.70	-6.07	9.89	15.97	3.30
Dimension 2	6.50	-1.07	12.05	13.12	2.75
Dimension 3	-1.98	-6.58	2.45	9.03	2.00
Dimension 4	0.90	-2.53	6.55	9.08	1.94
Dimension 5	-1.37	-4.94	0.95	5.89	1.25
Dimension 6	0.63	-2.19	2.97	5.16	1.21
Dimension 7	-0.35	-2.47	2.50	4.97	1.12
……SI_A……..					
Dimension 1	3.79	-10.91	16.28	27.20	6.19
Dimension 2	-2.14	-8.59	4.77	13.37	2.83
Dimension 3	-1.51	-5.65	6.31	11.96	2.49
Dimension 4	-1.02	-4.17	1.76	5.93	1.33
Dimension 5	0.41	-3.57	4.17	7.73	1.80
Dimension 6	-0.72	-3.89	2.11	6.00	1.10
Dimension 7	-0.11	-3.29	4.63	7.92	1.54
……WT_A…….					
Dimension 1	-3.36	-12.75	8.63	21.38	4.07
Dimension 2	-2.34	-12.09	3.93	16.03	3.42
Dimension 3	2.43	-4.27	9.30	13.57	2.78
Dimension 4	-0.75	-4.33	3.49	7.82	1.74
Dimension 5	0.33	-3.11	2.95	6.06	1.51
Dimension 6	0.27	-3.63	3.05	6.68	1.34
Dimension 7	0.92	-1.84	4.62	6.46	1.55
……NS_B…….					
Dimension 1	-0.35	-7.94	7.66	15.60	3.52
Dimension 2	6.31	0.52	12.09	11.56	3.27
Dimension 3	-2.25	-6.06	0.85	6.91	1.73
Dimension 4	0.86	-1.57	4.17	5.74	1.58
Dimension 5	-0.78	-3.05	3.32	6.37	1.51

Dimension 6	0.23	-1.90	2.31	4.21	1.03
Dimension 7	-0.29	-2.81	2.51	5.31	1.36
Dimension 1	-0.35	-7.94	7.66	15.60	3.52
……SI_B…….					
Dimension 1	-0.58	-10.89	19.01	29.90	5.88
Dimension 2	-5.71	-13.31	2.38	15.70	3.54
Dimension 3	-3.41	-7.03	2.36	9.39	2.64
Dimension 4	-0.74	-4.91	2.72	7.63	2.29
Dimension 5	1.67	-2.81	5.90	8.71	2.00
Dimension 6	-1.61	-4.88	1.74	6.62	1.64
Dimension 7	0.90	-2.02	6.36	8.38	2.22
Dimension 1	-0.58	-10.89	19.01	29.90	5.88
……..WT_B……..					
Dimension 1	-2.82	-9.97	3.77	13.56	3.81
Dimension 2	-6.01	-14.05	3.41	17.45	4.05
Dimension 3	-0.72	-7.32	7.47	14.80	3.94
Dimension 4	-0.78	-5.79	3.33	9.12	2.25
Dimension 5	1.48	-0.88	6.75	7.63	1.89
Dimension 6	-1.21	-5.27	2.12	7.40	1.61
Dimension 7	1.11	-3.36	6.43	9.79	2.30
……NS_C..…….					
Dimension 1	2.63	-7.27	13.61	20.88	4.91
Dimension 2	7.39	-3.77	17.20	20.97	3.99
Dimension 3	-0.01	-5.43	7.83	13.26	2.64
Dimension 4	1.57	-3.85	6.87	10.72	2.29
Dimension 5	-0.65	-3.97	3.72	7.69	1.64
Dimension 6	0.25	-2.69	4.26	6.95	1.45
Dimension 7	-1.41	-4.18	2.21	6.39	1.39
……SI_C……..					
Dimension 1	2.86	-9.86	20.37	30.23	6.40
Dimension 2	-1.19	-11.03	4.80	15.83	3.57
Dimension 3	-0.84	-7.67	5.70	13.37	3.02
Dimension 4	0.42	-4.28	7.64	11.92	2.20
Dimension 5	0.15	-6.74	6.72	13.46	2.23
Dimension 6	-0.05	-3.16	4.73	7.89	1.55
Dimension 7	-0.68	-3.85	3.70	7.55	1.73
……WT_C…….					
Dimension 1	-3.26	-10.71	7.21	17.92	4.05
Dimension 2	-2.14	-13.37	6.12	19.49	3.40
Dimension 3	2.87	-5.70	11.42	17.12	3.89
Dimension 4	-0.43	-5.63	8.17	13.81	2.15
Dimension 5	-0.18	-5.60	6.44	12.04	2.12
Dimension 6	0.52	-3.02	8.78	11.80	2.01
Dimension 7	0.48	-3.43	4.92	8.35	1.70

(*Note: A = subgenre “Questions to the Prime Minister/Chief Executive”; B = subgenre “Questions to the Secretaries/Ministers”; C = “Debates”; NS = native speech; SI = simultaneous interpreting (into B); WT = written translation (into B))

Several observations can be made based on Table 5.5. First of all, SI subgenres (i.e., SI_A, SI_B, and SI_C) seem to be more varied than corresponding NS subgenres (NS_A, NS_B, and NS_C) along most of the dimensions, given its higher standard deviations as well as the larger range between maximum and minimum values. In other words, mediated spoken texts are found to be less homogeneous than unmediated ones along different dimensions, offering counterevidence to the "levelling out" or "convergence" hypothesis mentioned before (Baker, 1996; Laviosa, 2002). Such (sub-)genre variation within SI is particularly noticeable along Dimension 1, where the highest standard deviation (SD) score reaches 6.40 (SI_C), while the highest SD scores for NS and WT are 4.91 (NS_C) and 4.07 (WT_A) respectively. Second, based on the mean dimension scores for the corresponding subgenres across NS, SI, and WT (e.g., NS_A, SI_A, and WT_A), a subgenre influence may have been identified along some of the dimensions. More straightforward visualizations have been provided (see Figures 5.2 to 5.8) to illustrate in detail how subgenres may have influenced the overall variation patterns of SI compared to NS and WT along different dimensions. Last but not least, the extreme cases (i.e., minimum and maximum values) in each of these subgenres may offer further evidence to the fact that there are no single absolute differences between languages of different mediation status (SI vs. NS), or different modes of mediation (SI vs. WT). For example, while SI in general is found to be more marked in the use of positive linguistic features associated with involved, unplanned, and informal language use along Dimension 1, the minimum values of -10.91 in SI_A, -10.89 in SI_B, and -9.86 in SI_C reveal the opposite, highlighting a strong informational focus.

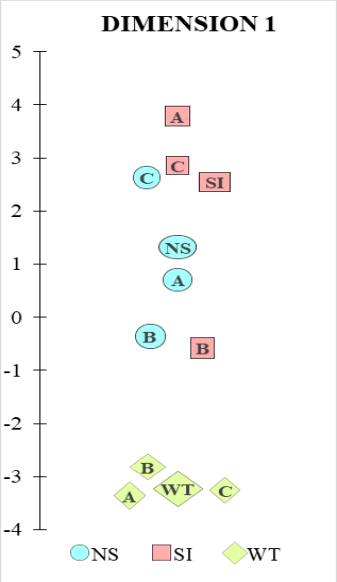

Figure 5.12 Mean scores of Dimension 1 for each of the subgenres in NS, SI, and WT Dimension 1: 'Involved versus Informational Production'

(*Note: A = genre "Questions to the Prime Minister/Chief Executive"; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = native speech; SI = simultaneous interpreting (into B); WT = written translation (into B))

Figure 5.12 plots the mean scores of Dimension 1 for each of the subgenres in the three language varieties. The main purpose, as is the same for the following figures, is to see if the identified co-occurrence patterns of linguistic features specific to SI reported in the above section can be consistently confirmed when subgenre comparisons are carried out. Figure 5.2 reports that Dimension 1 sets SI apart from both NS and WT (as also shown here), in that SI viewed as a whole is more marked by its involved and informal language production, characterized by more frequent use of grouping features such as present tense, contractions, shorter words, discourse particles, demonstrative/first-person/second-person pronouns, independent clause coordination, while unmarked by the use of co-occurring features highlighting informational, integrated, and well-planned language use. Figure 5.12 shows that such (un-)marked use of linguistic co-occurrence is not consistent from a comparable perspective, indicating a strong genre influence over the generally more marked nature of SI as being more involved and informal than NS, while the identified patterns are very consistent in terms of intermodal subgenre comparisons.

Specifically, among the three genres A, B, and C, only SI_A shows more marked use of involved features than corresponding NS_A, confirming the general patterns of SI viewed as a whole, while the other two corresponding subgenres (i.e., SI_B vs. NS_B, and SI_C vs. NS_C) shows no distinguishable difference. As a matter of fact, SI_B ("*Questions to Secretaries*") has a rather intermediate dimension score, indicating that this subgenre is unmarked in both the use of co-occurring features associated with involved and informal production, and those with informational and integrated language production. It may well be the case that the more marked nature of SI as a whole (Figure 5.2) compared to NS is genre-sensitive. The intermodal comparison between SI and WT subgenres, however, shows a consistent pattern of the more marked nature of SI as being more involved and less informational, offering strong evidence to the influence of modality, that is, mode of mediation on the surface manifestations.

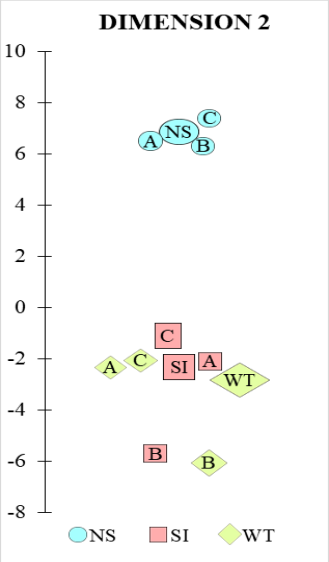

Figure 5.13 Mean scores of Dimension 2 for each of the subgenres in NS, SI, and WT Dimension 2: 'On-line Information Elaboration with Stancetaking Concerns'

(*Note: A = genre "Questions to Prime Minister/Chief Executive"; B = genre "Questions to Secretaries/Ministers"; C = genre "Debates"; NS = native speech; SI = simultaneous interpreting (into B); WT = written translation (into B))

Figure 5.13 conforms to the general patterns reported in Figure 5.3 in which both SI and WT are found to be unmarked in the use of linguistic features associated with on-line information elaboration with stancetaking concerns, while marked in the use of co-

occurring features indicating 'conditioned' description. These patterns are very consistent, as each of the SI subgenres are consistently unmarked in Dimension 2 compared to the corresponding subgenres in NS, while there is no noticeable difference between SI and WT subgenres. Therefore, no strong subgenre influence is reported in Dimension 2.

However, subgenre variations do exist in the two mediated language varieties, where SI_B and WT_B ("*Questions to the Secretaries*") stand out from the other two subgenres, characterized by strongly unmarked use of stancetaking features under strict time constraints while very frequent use of conditional adverbial subordinators and analytic negations. The marked nature of SI_B also contributes to the more marked dimension score difference in relation to NS_B, a finding indicating the influence of genre variation over the general patterns of SI in Dimension 2.

One more observation concerns subgenre comparison between SI and WT. Figure 5.13 shows that there is almost no noticeable difference between the corresponding subgenres, indicating great homogeneity of texts of different mediation modes. It is tentatively argued that this unmarked nature in D2, i.e., infrequent use of features associated with on-line information elaboration with stancetaking concerns, may well be a shared dimension of mediated texts.

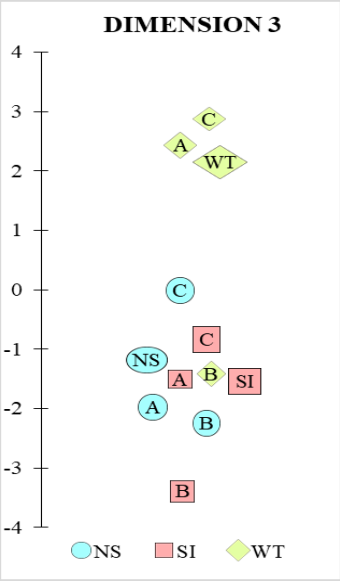

Figure 5.14 Mean scores for each of the subgenres in NS, SI, and WT

Dimension 3: ‘Precise versus Simplified Description’

(*Note: A = genre “Questions to Prime Minister/Chief Executive”; B = genre “Questions to Secretaries/Ministers”; C = genre “Debates”; NS = native speech; SI = simultaneous interpreting (into B); WT = written translation (into B))

Figure 5.14 lends support to the relatively clear-cut picture presented in Figure 5.4, in which SI is found to be noticeably distinguishable from WT, whereas it shares great similarity with NS along this dimension continuum. In terms of the comparable comparison, only SI_B versus NS_B shows a noticeable difference, featured by more marked use of repetitive and simplified linguistic features including contractions, top 10 coverage and independent clause coordination, while the comparison between the other two corresponding subgenres (i.e., SI_A vs. NS_A, and SI_C vs. NS_C) has cancelled each other out, which may have led to the similar variation patterns between SI and NS. Intermodal comparison between SI and WT subgenres, however, is very consistent, as each of the SI subgenres show distinguishable differences from their WT counterparts. Subgenre variations, once again, have been observed among mediated texts, in which both SI_B and WT_B demonstrate more marked patterns than the other subgenres. For example, while WT_A and WT_C are characterized by a marked use of linguistic features associated with more precise and elaborated description, such as split auxiliaries, conjuncts, adverbs, downtoners, concessive adverbial subordinators, WT_B, by contrast, is less marked in this aspect while slightly more marked in the use of simplified and repetitive linguistic features. Therefore, both subgenre influence (over the general situating patterns of SI along D3) and subgenre variations have been observed in this dimension.

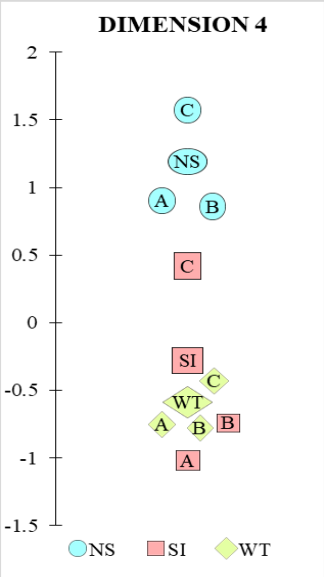

Figure 5.15 Mean factor score for each of the subgenres in NS, SI, and WT
Dimension 4: 'Narrative versus Abstract Concerns'

(*Note: A = genre "Questions to Prime Minister/Chief Executive"; B = genre "Questions to Secretaries/Ministers"; C = genre "Debates"; NS = native speech; SI = simultaneous interpreting (into B); WT = written translation (into B))

Figure 5.15 illustrates distinguishable and consistent differences between SI and NS subgenres, while such differences are not noticeable from an intermodal perspective (except for SI_C vs. WT_C), a finding conforming to Figure 5.5, which reports more marked use of narrative linguistic features by native speakers, while both translators and interpreters resort more or less balanced use of linguistic features associated with narrative and/or abstract concerns given their intermediate dimension scores. In terms of subgenre variations, however, both SI and NS subgenres are less homogeneous than WT subgenres, in that SI_C and NS_C stand out from the remaining others along this dimension. Such subgenre variations have also contributed to the overall positioning of the three varieties as a whole.

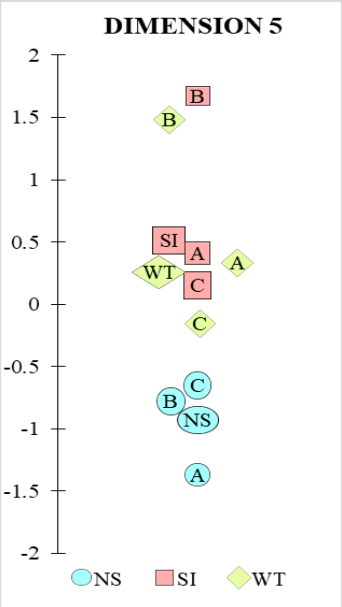

Figure 5.16 Mean scores for each of the subgenres in NS, SI, and WT
Dimension 5: 'Lexical versus Functional Concerns'

(*Note: A = genre "Questions to Prime Minister/Chief Executive"; B = genre "Questions to Secretaries/Ministers"; C = genre "Debates"; NS = native speech; SI = simultaneous interpreting (into B); WT = written translation (into B))

Figure 5.16 has also confirmed the linguistic patterns of SI observed in Figure 5.6, in which SI demonstrates strong differences from NS along Dimension 5, while there is no noticeable difference between SI and WT. While the identified patterns specific to SI are rather consistent in terms of subgenre comparisons, there are noticeable subgenre variations within mediated texts, in which SI_B and WT_B stand out from the others, exhibiting a strong lexical focus, while the remaining subgenres are much more balanced in the use of features associated with either lexical or functional concerns. Subgenre variations are less acute within NS, though NS_A also distinguishes itself from the others by being more marked in terms of its functional focus.

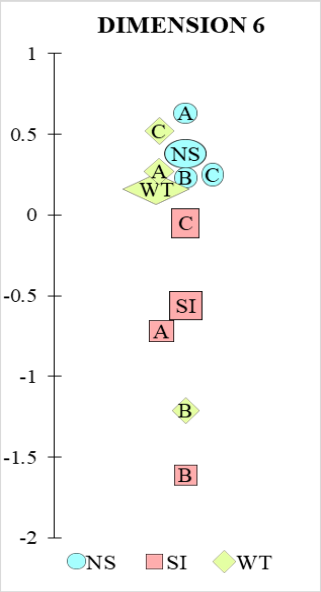

Figure 5.17 Mean scores of Dimension 6 for each of the subgenres in NS, SI, and WT Dimension 6: 'On-line Persuasion'

(*Note: A = genre "Questions to the Prime Minister/Chief Executive"; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = native speech; SI = simultaneous interpreting (into B); WT = written translation (into B))

Dimension 6 in Figure 5.7 isolates SI from both NS and WT, characterized by slightly unmarked use of linguistic features associated with on-line persuasion, including suasive verbs, public verbs, predicative modals, and 'that' verb complements, while slightly marked use of standardized type-token ratio. Figure 5.17, however, shows that this SI-specific pattern is not always consistent. In terms of comparable comparison, while SI_A and SI_B manifest distinguishable differences from NS_A and NS_B with respect to their mean dimension scores, the difference between SI_C and NS_C is not equally noticeable, indicating a possible genre influence along this dimension. In terms of intermodal comparison, while SI subgenres show consistent differences from WT subgenres, such differences are not always equally prominent. In addition, due to the relatively narrow range of the mean dimension scores, most of these subgenres are not very marked in either side of this dimension, despite their dimension score differences. Strong variations within SI subgenres have been observed, among which SI_B is more marked for its infrequent use of on-line persuasion features. Similarly, WT_B also stands out from the other two, situating towards the negative pole of the continuum. By contrast, NS subgenres are very

homogeneous in this dimension.

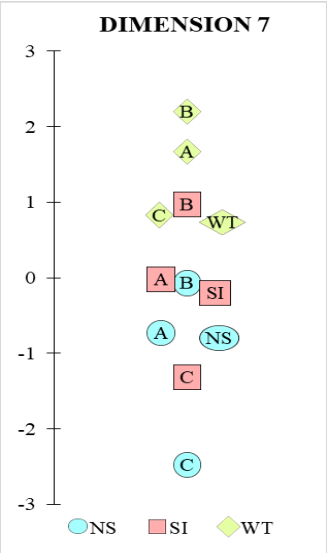

Figure 5.18 Mean scores of Dimension 7 for each of the subgenres in NS, SI, and WT Dimension 7, 'Coordinating versus Possessive Functions'

(*Note: A = genre "Questions to the Prime Minister/Chief Executive"; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = native speech; SI = simultaneous interpreting (into B); WT = written translation (into B))

Figure 5.18 shows that the co-occurrence patterns identified in Figure 5.8 with respect to SI are consistent from both comparable and intermodal perspectives. That is, each of the SI subgenres situates between WT and NS subgenres, as is the case when they are observed as a whole (Figure 5.8). Overall, WT and its subgenres are distinguishable from the others (except for SI_B), characterized by their marked use of coordinating conjunctions while unmarked use of co-occurring nouns, possessive endings, and lexical density. SI and NS, along with their subgenres (except for SI_C and NS_C), are not readily distinguishable, given the relatively intermediate range of dimension scores, which means that they are generally unmarked in the use of either positive features or negative features along this dimension. To be specific, SI_A, with its dimension score around 0, is equally characterized by both positive features and features with negative weights, as is the case for NS_B. SI_B, by contrast, resembles WT_C in that both are marked by more frequent use of coordinating conjunctions, and unmarked by co-occurring features with negative weights, while SI_C shows the opposite trend in relation

to SI_B. None of the three language varieties show high level of homogeneity along this dimension, as their mean dimension scores spread out along the dimension continuum.

Based on the above analysis, the identified co-occurrence patterns of linguistic features in SI (as a whole) compared with NS and WT are not always consistent across subgenres. More often than not, certain SI subgenre exhibits more/less marked language use than the corresponding NS and/or WT subgenre, indicating a possible genre influence. Genre influence (or in other studies, register variation) as a potential factor for translation-specific features has received growing research attention over the past few years, and many studies (see among others, X. Hu et al., 2016; H. Kruger & Van Rooy, 2012; Puurtinen, 2003; Redelinghuys, 2016) have reported that some of these 'universal' features of translation are subject to register or genre variation. That is, while some of the 'universal' features of translation, such as more simplified and explicit language use, are identified within certain genre in relation to a comparable genre in non-translated texts, these features are not (equally) obvious in other genre comparisons. This study also lends some support to this argument. In addition, (sub-)genre variations, as argued also by Biber (1988), also contribute to the overall variation patterns of all three language varieties along the seven dimensions. As far as SI is concerned, it seems that subgenre SI_B ("*Questions to the Secretaries*") stands out along many dimensions, which may suggest the specificity of this subset in mediated texts. To verify this assumption, the author visualizes the mean dimension scores of all three subsets in SI, shown in Figure 5.19.

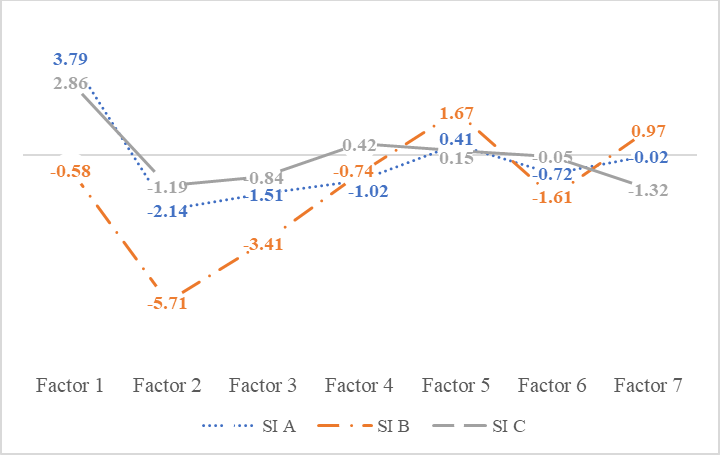

Figure 5.19 Mean factor scores of SI_A, SI_B, and SI_C

(*Note: SI_A = subgenre "Questions to Chief Executive"; SI_B = subgenre "Questions to the Secretaries"; SI_C = subgenre "Debates"; SI = simultaneous interpreting (into B))

Figure 5.19 confirms the author's assumption, as well as part of the results reported in the previous section: subgenre SI_B does stand out from the other two subgenres along most of the seven dimensions, and the differences are most noticeable in Dimension 1 ('Involved versus Informational Production'), Dimension 2 ('On-line Information Elaboration with Stancetaking Concerns') and Dimension 3 ('Precise versus Simplified Description'). For Dimension 1, while both SI_A and SI_C are marked in more involved and informal language production, and unmarked in the use of linguistic features associated with integrated, formal, and informational production, SI_B is characterized by slightly more frequent use of informational linguistic features. SI_B along Dimension 2 also exhibits strongly marked preference for 'conditional' description, and strongly unmarked use of stancetaking features produced under strict time constraints. In terms of Dimension 3, it is, once again, much more marked in terms of simplified language use in relation to the other two subgenres. While the differences along the other dimensions are much less noticeable.

There are two possible reasons for such subgenre variation within SI subsets. One is the possible influence of source speech (ST), especially in terms of its mode of delivery (scripted, unscripted, mixed) or production conditions, while the other is concerned with the preparedness of simultaneous interpreters during translation (with texts or without texts). In terms of source language influence and its production conditions, the subset "*Questions to the Secretaries*", i.e., ST_B, often starts with written Q&As, followed by follow-up spoken Q&As. Based on online video recordings, the written parts of Q&As are often scripted, as source speakers usually read out their prepared questions and answers. By contrast, the follow-up questions and answers are often unscripted and delivered spontaneously. It is highly likely that these specific properties of source speeches have been transferred to the interpreted speeches. If this is the case, the author argues that similar variation patterns can also be observed in the corresponding WT subgenre (i.e., WT_B). Figure 5.20 seems to confirm this hypothesis, as WT_B

distinguishes itself from the others along most of the seven dimensions. However, this does not exclude the possible influence of the preparedness of simultaneous interpreters. For one thing, the comparison between transcriptions of SI and the translated texts (WT) reveals a large proportion of overlapping between the two in the written Q&As part in subset B, "*Questions to the Secretaries*", indicating a strong possibility that the interpreters might interpret with texts. For another, during the transcription phase, sounds of page flipping were heard constantly for the written Q&As but not the others, which may also indicate the likelihood of interpreters working with texts.

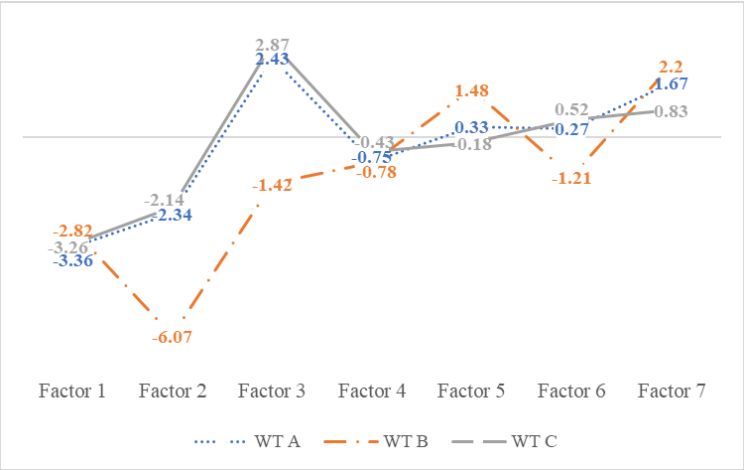

Figure 5.20 Mean factor scores of WT_A, WT_B, and WT_C

(*Note: WT_A = subgenre "Questions to Chief Executive"; WT_B = subgenre "Questions to the Secretaries"; WT_C = subgenre "Debates"; WT = written translation (into B))

5.2 L2 Interpreting as a multi-constrained language variety

5.2.1 L2 Interpretese along the seven dimensions

The above sections have presented separately the textual relations, or overall linguistic patterns of SI, NS, and WT, as well as those patterns of the different subgenres within the three language varieties. To highlight the linguistic features specific to L2 interpreting, i.e., L2 interpretese, in a more systematic manner, the author presents the full picture of the textual relations among all (sub-)genres along the seven dimensions, as illustrated below.

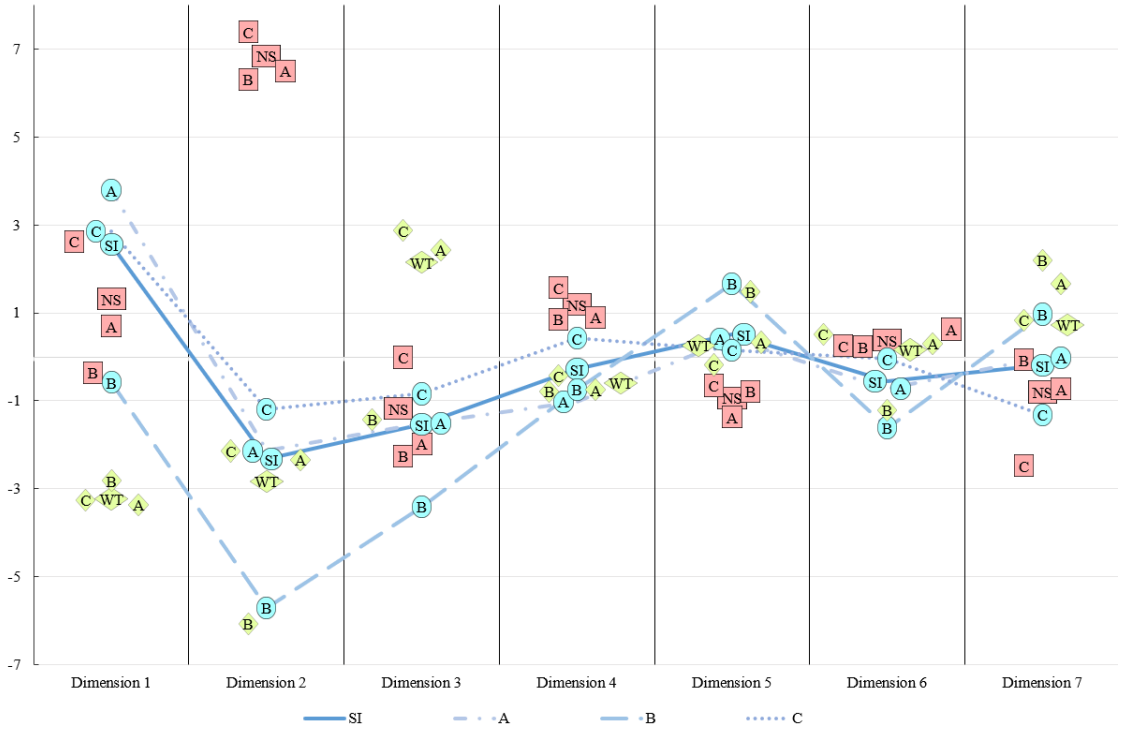

Figure 5.21 Plot of the textual relations among all (sub-)genres, highlighting SI (sub-)genres

(*Note: A = genre "Questions to the Prime Minister/Chief Executive"; B = genre "Questions to the Ministers/Secretaries"; C = genre "Debates"; NS = native speech; SI = simultaneous interpreting (into B); WT = written translation (into B))

Generally speaking, the first three dimensions show more variation patterns across SI, NS, and WT, while variations along the remaining dimensions are not as equally distinguishable, conforming to the statistical tests presented in Table 5.4 (especially with reference to R*R values). Dimension 1 captures the unique co-occurring patterns of SI distinct from both NS and WT viewed globally. However, such distinct patterns are genre-sensitive from a comparable perspective (SI vs. NS), since only the comparison between SI_A and NS_A ("*Questions to the Prime Minister/Chief Executive*") is in line with the general SI-specific patterns, while cross-subgenre comparisons between the other two (SI_B vs. NS_B, and SI_C vs. NS_C) do not reveal much distinguishable differences. In other words, overall L2 interpreting is more marked in the co-occurrence of linguistic features associated with unplanned, involved, and informal language production, while unmarked in the co-occurrence patterns indicating planned, integrated, and informational focus, but such marked usage is subject to genre influence (genre A).

Dimension 2 identifies sharp differences between mediated texts (SI and WT) and unmediated texts (NS), in that native speech is characterized by markedly consistent overuse of linguistic features associated with stancetaking concerns produced under real time constraints, while both interpreting and translation are consistently unmarked in this aspect. Given the homogeneous variation patterns between SI and WT, it is tentatively argued that the marked manifestation in Dimension 2 might be a shared dimension of mediated languages. Dimension 3 reports more noticeable intermodal differences between SI and WT, while such a distinction is relatively moderate between SI and NS. Dimensions 4 and 5 reveal more ontological differences between SI and NS, and the differences are relatively consistent across subgenre comparisons. By contrast, SI and WT also reveal no distinguishable differences, except for one subgenre comparison (such as SI_C vs. WT_C in D4), suggesting, once again, the possible shared nature of mediated languages. Dimension 6, indicated by its very small R*R value (5.6%), report rather mixed patterns especially when subgenre comparisons are taken into consideration; while Dimension 7 shows that SI situates between NS and WT along this continuum, relatively unmarked for the use of either positive or negative features.

To sum up, following the strict criteria for the identification of L2 interpretese, only Dimension 1 seems to be a potential candidate, but needs to be considered against the background of a strong genre influence, while the other dimensions single out either ontological or intermodal differences. When subgenre variations are taken into consideration, these specific patterns of L2 interpreting are not always straightforward, and along many dimensions, SI cannot be readily isolated from NS or WT. Therefore, the L2 *interpretese* identified in the current research are summarized as the following:

Simultaneous interpreting (into B) is characterized by more marked use of a number of co-occurring linguistic features associated with involved and informal language production, operationalized as present tense, contractions, shorter words, be as main verb, discourse particles, demonstrative pronouns, first and second person pronouns, pro-verb do, independent clause coordination, and many others, while unmarked use of co-

occurring features associated with integrated, planned, and formal language production, such as average word length, longer words, average sentence length, total prepositional phrases, total other nouns, adjectives, phrasal coordination, conjuncts, and others, compared with both native speech and written translation (into B) from the same source. However, these co-occurrence patterns are rather genre-sensitive, among which subgenre SI_A "Questions to the Chief Executive" demonstrates strong differences from the corresponding NS_A "Questions to the Prime Minster", conforming to the general SI-specific patterns reported here.

Except from Dimension 1 ('Involved versus Informational Production'), the other dimensions reveal more similarities than differences from both comparable and intermodal perspectives, especially when the relatively intermediate dimension scores (ranging from -0.5 to 0.5) along the dimension continuum are taken into consideration. The comparable analysis reveals that SI (and SI subgenres) can be consistently distinguished from NS (and corresponding NS subgenres) along D2 ('On-line Information Elaboration with Stancetaking Concerns'), D4 ('Narrative versus Abstract Focus'), and D5 ('Lexical versus Functional Concerns'). By contrast, SI and WT along these same dimensions are generally undistinguishable (except for SI_C vs. WT_C along D4), which has strong implications for the shared nature of translation and interpreting as forms of mediated, or even constrained languages (H. Kruger & Van Rooy, 2016a). The intermodal differences are more noticeable along D3 ('Precise versus Simplified Description'), D6 ('On-line Persuasion'), and D7 ('Coordinating versus Possessive Functions'), while SI and NS are not readily distinguishable along the same dimensions, except for D7.

In terms of the specific manifestations of the co-occurring linguistic patterns of SI, overall SI (as a whole) is marked for its involved and informal language use, while unmarked for integrated, planned, and informational language production (D1), unmarked for on-line information elaboration with stancetaking concerns, and marked for conditioned description (D2), moderately marked for more simplified and repetitive language use, and

unmarked for features associated with precision and elaboration (D3), unmarked for co-occurrence of linguistic features related to either narrative or abstract concerns (D4), unmarked for either lexical or functional usage (D5), less marked for on-line persuasion features (D 6), and finally, unmarked for co-occurring features associated with either coordinating or possessive functions (D7).

5.2.2 Implications for 'universal' features of mediated language

The "bottom-up" multidimensional analysis carried out in this chapter reveals sharp differences in relation to the previous "top-down" unidimensional or univariate studies, in that there is no presumption with respect to the 'universal' features of mediated language. Rather, linguistic co-occurrence patterns are automatically identified based on their distribution patterns in all text segments within the three language varieties. Consequently, there is no guarantee that the widely acclaimed 'universal' features of mediated language, such as simplification, explicitation, normalization and levelling out (or convergence), can be equally identified utilizing a different research approach. Nonetheless, based on the identified dimensions along which SI may be distinguished from NS and/or WT, implications can be drawn in terms of the 'universal' features of L2 interpreting as bi-mediated (non-native, and Translated) language.

(1) The generally more simplified pattern of mediated language can be confirmed for L2 interpreting in relation to both unmediated spoken language and L2 translation. One exception, as also been the case for previous studies (see among other, Bendazzoli & Sandrelli, 2005; Kajzer-Wietrzny, 2012; Russo et al., 2006), is that interpreted language is characterized by higher information density compared with non-interpreted language, which may be the result of certain interpreting strategies (such as condensation technique) due to the highly cognitive-challenging nature of simultaneous interpreting. However, this more simplified pattern of L2 interpreting is subject to genre variation as well as genre influence.

(2) The generally more explicit pattern of mediated language cannot be readily identified based on the multidimensional analysis, as no dimension reported in this chapter groups together the often-discussed linguistic indicators for increased explicitness. One dimension, i.e., Dimension 6 ('On-line Persuasion'), seems to be promising, as the co-occurring linguistic features (i.e., suasive verbs, public verbs, predicative modals, and 'that' verb complements) with positive weights overlap with previous univariate studies on optional 'that' after reporting verbs. However, it is argued that the focuses are different. Dimension 6 is more marked in terms of the co-occurrence patterns of reporting verbs (higher factor loadings) with an aim to persuade rather than explicitate. Therefore, the multidimensional analysis cannot provide ready evidence for the increased explicitness pattern.

(3) The generally more normalized/conventional pattern of mediated language might be disproved, based on Pym's (2008b, p. 4) interpretation of Toury's law of growing standardization, which "posits *gross modo* that translations have less internal linguistic variation than non-translations", an interpretation that seems to overlap with the 'levelling out' or 'convergence' hypothesis (Baker, 1993; Laviosa, 2002). Besides, the more marked manifestation of SI in the positive end of Dimension 1, 'Involved and Informal Production', may add further counterevidence to the normalization hypothesis. Nevertheless, the conclusion needs to be made with great caution, since the often-discussed features for normalization, such as fixed expressions or lexical bundles, reformulation markers, etc. (e.g., Kajzer-Wietrzny, 2012; R. Xiao, 2011), have not been investigated in the current research.

(4) The generally levelling-out or converging pattern of mediated language is also disproved along most of the dimensions identified in this study, as SI in general is characterized by greater diversity/internal variations in terms of the co-occurrence patterns along the seven dimensions, compared with NS, the three subgenres of which show overall much more homogeneous variation patterns.

These implications, as well as the results reported here, also offer indirect references to the possible role of working direction (A-to-B, L2) on the linguistic manifestations of simultaneous interpreting. Being a bi-mediated (non-native, and Translated) language variety, L2 interpreting is supposed to show more distinctive patterns than native interpreting, such as consistently more simplified language use, consistently more explicit and perhaps conventional production. Both the unidimensional and the multidimensional analyses carried out in this research fail to show this. Further still, Figure 5.21 shows that SI subgenres are not very distinguishable from NS subgenres along many dimensions, meaning that the differences between non-native, mediated language production and native, unmediated language production are not always clear-cut. This finding seems disappointing, considering especially the much more constrained nature of L2 interpreting. One of the possible reasons, as argued here, might be that as interpreters in the LegCo setting work unanimously into B, instead of shifting constantly between A-to-B and B-to-A working directions, the possible influence of working direction may have been flattened out. Other possible factors, as argued also by Gile (2005, p. 9), such as "[l]anguage-specific and language-pair specific factors, as well as variability in other relevant factors, may offset such calculations to the extent that depending on circumstances, directionality may lose much of its importance."

5.2.3 Possible constraining factors

To account for the mixed and multiple results reported above, it is important to examine the very nature of L2 interpreting as a multi-constrained language variety. By saying "multi-constrained", the author means that L2 interpreting, and as a matter of fact, all language production, is constrained by an interplay of multiple factors, as acknowledged also by Baker (1999), Laviosa (2008), Lanstyak and Heltai (2012), Kruger and Van Rooy (2012), to name just a few. Previous studies often resort to monofactorial perspective, and attribute the 'universal' features to either the constraining influence of source language or source texts (e.g., Dai & Xiao, 2011), the translation-inherent cognitive processing (e.g., Olohan & Baker, 2000), or the influence of register or genre (e.g., H. Kruger & Van Rooy,

2012). The author concurs with Baker's (1999, p. 285) view that "language in general, and the language of translation in particular, reflects constraints which operate in the context of production and reception: these constraints are social, cultural, ideological, and of course also cognitive in nature". That is, there is no single factor that can account for the distinct manifestations of interpreted language compared to non-interpreted or unmediated language, and to translated language from the same source. Recent efforts have been made towards a multifactorial analysis, utilizing more scientific and sophisticated approaches, evidenced in Kruger (2018), Kruger and De Sutter (2018), and De Sutter and Vermeire (2020), to disentangle the possible influence and interactions of the above mentioned factors. The current research, however, does not aim to do the multifactorial analysis, as it can constitute a wholly new, independent, and promising research project. Rather, based on the results reported in this chapter, the author wants to highlight two possible contributors which are believed to have a strong influence on the overall distribution patterns of the co-occurring linguistic features in SI compared with NS, namely, the interpreters' pragmatic risk-avoidance concerns proposed by Pym (2004, 2005, 2008a, 2008b, 2015) and genre influence.

5.2.3.1 Pragmatic risk-avoidance concerns

Translation (including interpreting) is a purposeful activity undertaken in certain social context. The purpose of simultaneous interpreting, as well as the translated documents, in the LegCo setting is to convey to the public, as well as to the outside world, the work and progress that the Hong Kong government is making to address the livelihood issues that are of great concern to the Hong Kong people. Thus, a number of potential risks, that is, "the possibility of not fulfilling the translation's purpose" (Pym, 2004) are involved during the translating process. One immediate risk is interpreters' failed attempts to convey the accurate information from the source speakers, i.e., Members of the Legislative Council representing different political groups, to the public, the people that theses Members represent for.

The general concern here, however, is how the risk-averse behaviors of simultaneous interpreters have contributed to the final manifestation of the linguistic patterns identified in SI compared with NS. Pym (2008b) argues that during the translating process, translators (as well as interpreters) are always involved in the tug of war between reward and risk-taking or risk-avoidance. That is, "translators will tend to take risk X in the presence of reward structure Y" (Pym, 2008b, p. 326); otherwise, they will choose the opposite risk-averse options. The heatedly debated 'universal' features of translated language, in Pym's view, can be approached from the model of risk management, since translation is always full of risks, either high risks or low risks. He explains that the reasons why translators either "tend to standardize language or to channel interference" are because "there are two main ways of reducing or transferring communicative risk" (ibid., p.325). Other 'universal' patterns of translated language can also be explained utilizing risk management, as translators' constant preference for risk averse options "may develop into a deceptively universal behavioral disposition" (p.326) epitomized as various kinds of 'universal' patterns, such as simplification, explicitation, or normalization.

Applied to the current research, it is argued that the dramatic variation between mediated texts (including SI and WT) and unmediated texts (NS) may well indicate translators' and interpreters' awareness of risk management, manifested in particular in the markedly underuse of stancetaking features produced under real time constraints. Previous studies have offered evidence to interpreters' reservation of personal stances during interpreting, although contradictory findings have also been reported, and even if interpreters are found to show any stancetaking behaviours, they are often in line with the institution they represents for (see e.g. Wang & Feng, 2018).

In addition to the dramatic variation between SI and NS along Dimension 2, previous analyses also reveal higher lexical focus (information density) of SI in relation to NS. One possible reason, as discussed before, may be attributed to the interpreters' condensation strategies during translation due to high time constraints. Based on Pym's

risk management model, however, it is argued that the deep-rooted reason for such a condensation technique, and also many other coping strategies, can be attributed to the interpreters' risk-avoidance consideration. In his explanation (Pym, 2004), "[v]arious translation strategies can be used to reduce or maintain levels of risk, and the strategies can consequently be described as having low-risk or high-risk consequences with respect to the problem concerned." The higher information density might be the interpreters' deliberate decision-marking to leave out certain cohesion markers so as to catch up with source speakers and avoid the possible risk of leaving some important information untranslated.

5.2.3.2 Genre influence

Another conditioning factor is genre influence, as attested from previous analyses, which report that along many dimensions, the distinguishable linguistic co-occurrence patterns of SI are genre-sensitive: while certain subgenre comparisons between SI and NS, or between SI and WT lend support to the general distinguishable patterns of SI along certain dimensions, other comparisons fail to reveal any noticeable differences. For instance, while Dimension 1 shows that SI in general is more involved and informal with respect to NS and WT, such a distinguishable pattern might be attributed to one pair of genre comparison, i.e., SI_A vs. NS_A, as comparisons between the remaining corresponding subgenres fail to show any noticeable differences.

Nowadays, more and more academic endeavors have been made to investigate the possible influence of genre, or as used in previous studies, register variation on the 'universal' features of translated language. Sufficient evidence (Kruger & Van Rooy, 2012a; Puurtinen, 2003; Redelinghuys, 2016) has shown that these reported features are not always consistent when different genres or registers are under comparison. Similar studies regarding interpreted language are under-explored given the limited number of "genres" in interpreting settings. More often than not, studies on the features of interpreted language focus only on one certain genre (such as parliamentary discourse as

one broad genre in EPIC) without paying any attention to the possible nuances among subgenres and their influences on the overall distribution patterns of interpreting. Section 5.1.3 offers sufficient evidence to this.

The two main contributors aside, it is acknowledged that the specific bilingual cognitive processing in simultaneous interpreting (and perhaps more specifically, in SI from A to B language direction) also plays an important role in the manifestations of the interpreted language, due to simultaneous interpreters' multitasking on a tightrope (Gile, 1995, 1999). The less clear-cut patterns of explicitation based on the multidimensional analysis, for example, might be attributed to the limited processing capacity of simultaneous interpreters, in that "interpreters have to concentrate on everything that the speaker says, whereas delegates can select the information *they* are interested in" (Gile, 1995, p. 165), leaving the interpreters no extra time for explicitation (see also Pym, 2008a). Meanwhile, the limited processing capacity of simultaneous interpreters may, in turn, lead to more explicitating shifts to mask processing difficulties (e.g., Defrancq et al., 2015; Gumul, 2020).

5.3 Summary

This chapter aims to answer the third major research question proposed in Chapter one, i.e., RQ 3: What are the general co-occurrence patterns of the 79 linguistic features in SI compared with NS and WT? Two sub-questions are also addressed, in relation to the multi-dimensions that SI situates along, as well as the consistency patterns of SI when subgenre comparisons are taken into consideration. To answer these questions, a multidimensional analysis utilizing factor analysis has been carried out. The MD approach identifies, altogether, seven factors, or dimensions interpreted in functional terms, that distinguish SI from NS and/or WT. Technically speaking, only Dimension 1, 'Involved versus Informational Production', captures L2 interpretese as defined in the research, while the remaining dimensions report either similarities or differences between SI and NS, or between SI and WT, among which Dimension 2, 'On-line Information Elaboration with Stancetaking Concerns' reports the most dramatic differences between

SI and NS, while both SI and WT as mediated texts are highly and consistently unmarked in this dimension, suggesting possible shared co-occurring linguistic patterns of mediated texts.

In terms of the similarities/differences of the co-occurrence patterns along the other dimensions (Dimensions 3 to 7), there are no single, absolute differences across the three language varieties, especially when subgenre comparisons are taken into consideration. The seven dimensions identified in this chapter are general underlying parameters of variation described "in relatively global terms" (Biber, 1988, p. 169). They do not, however, represent all the differences defined by the 79 linguistic features.

Although differences concerning the very nature of SI into B are observed in relation to NS and WT, the co-occurring patterns of SI per se along the seven dimensions are not marked enough despite its multi-constrained nature, in that along many dimensions SI (and SI subgenres) situate towards the middle side of the continuum, indicating that it is unmarked for the use of linguistic features with either positive or negative weights. This may have offered counterevidence to Shlesinger and Ordan's (2012) claim that all the features identified for translation (in relation to non-translation) may be all the more salient for interpreting, and that interpreting can be considered as an extreme case of translation.

The multidimensional analysis carried out here is illuminating in that, it shows SI (i.e., L2 interpreting) as a multi-constrained language variety, is multidimensional and multifaceted in nature. Relations between SI and NS, SI and WT, or across the three, can never be considered from one single dimension, or from the distributional patterns of several linguistic features, as it will lead to inaccurate and incomplete description of the nature of interpreted language. The results reported here also have certain implications for the 'universal' features widely hailed in previous scholarship, but they do not show much support to these general tendencies of mediated language, except for one possible feature, i.e., simplification. The contributing factors, however, are not easy to disentangle, and future endeavors need to be done in this regard.

Revealing as this new approach is, Biber (1992, pp.136-137) identifies two potential problems associated with an exploratory approach such as exploratory factor analysis. One goes to the possibility that "the analysis can capitalize on chance co-occurrence patterns", while the other is that "certain results can be difficult to interpret because they have little basis in prior research studies". These two problems are also acknowledged in this research, especially with respect to the second one. Although some of the co-occurring features along one dimension have been identified previously, other features sometimes lack theoretical foundations, such as the positive loadings in Dimension 4, 'Narrative Concern'. Moreover, the labelling for all seven dimensions needs further confirmation, for which Biber (1992b) recommends another theory-based approach named confirmatory factor analysis (CFA), which also points to a new direction for future studies.

These two potential problems aside, an additional shortage, which also happens to be the unique advantage of the MD approach is that, unlike unidimensional analysis, linguistic features utilizing a MD approach are always considered from the viewpoint of co-occurrence patterns, in which case little attention is paid to the distributional patterns of individual linguistic feature along certain dimension. Take the co-occurring positive features on Dimension 1 as an example. SI is found to be much more marked in the positive continuum manifesting an involved and informal focus. However, the unidimensional analysis carried out in Chapter four shows that SI resorts less often to subordinator-that deletion ('that' omission) than NS, but viewed from a multidimensional perspective, such individual differences are concealed by the co-occurrence patterns. A solution might be the combination of the two so as to inform each other the linguistic patterns (either co-occurring or individual) of the linguistic features to be investigated.

Chapter 6 Conclusion

6.1 Summary of main findings

This project starts with the assumption that simultaneous interpreting (with a special focus on SI into a B language, or L2 interpreting), given its multi-constrained nature, must demonstrate certain linguistic patterns that distinguish it from both unmediated, native spoken language and translated language from the same/similar source. To test this hypothesis, a specialized million-size corpus consisting of three English components (i.e., interpreted texts, translated texts, and non-interpreted target originals) and one Cantonese source component (i.e., source texts) is constructed, thanks to the easy accessibility to online proceedings from two legislative settings, i.e., the UK Parliament, and the Legislative Council of Hong Kong. Three comparable genres (in a generic sense), including "*Questions to the Prime Minister/Chief Executive*", "*Questions to the Ministers/Secretaries*", and "*Debates*", are included to investigate possible genre influence on the identified patterns of SI, or L2 interpretese.

The extant research, as reviewed in Chapter two in great detail, relies heavily on the simplistic unidimensional and univariate analysis by carrying out frequency-based comparison of several selected linguistic features between mediated and unmediated language varieties. Based on frequency distribution, hasty conclusions are often drawn about the so-called 'universal' patterns or features of language in mediation. Thought-provoking as these studies are, since they have justified the status of mediated languages (specifically, translation and interpreting) as distinct, but equally important, language varieties in the target culture system and provided new perspectives to examine their very nature, they have also neglected their multifaceted nature shaped and constrained by various conditioning factors. Given this complex nature of language in mediation, it is intriguing to examine the possible similarities and/or differences among texts of different mediation status (e.g., translated, interpreted, unmediated spoken originals, etc.) along (potentially) multiple dimensions. A multidimensional analysis based on Biber's (1988)

variationist approach has been carried out for this purpose.

Before the multidimensional analysis, the author replicates the widely adopted unidimensional method, based on 79 linguistic features carefully selected in this study. The fundamental aim is to see what the unidimensional analysis can inform us about the general linguistic patterns of the three language varieties, and if the much-discussed 'universal' hypotheses can be attested using the LegCo+ data. The following sections provide a brief summary of the main findings based on two distinct research approaches, aiming to address the three major research questions proposed in Chapter one.

6.1.1 Summary of the main findings of the unidimensional analysis

Two main analyses have been carried out in Chapter four. The first analysis concerns with the overall distribution and variation patterns of the 79 linguistic features in NS, SI, and WT (RQ 1); while the second one aims to testify two much-discussed universal hypotheses in translation and interpreting studies, i.e., lexical simplification, and increased explicitness (or explicitation) (RQ 2).

Based on statistical tests on the variation patterns of 79 linguistic features in the three English varieties, several contradicting, albeit correlated, trends have been identified from both comparable (SI vs. NS) and intermodal (SI vs. WT) perspectives. In terms of comparable analysis, while SI is marked in features associated with more simplified, explicit, and conservative language use compared to NS, it is also characterized by an overuse of nominal features indexing high information density. In terms of intermodal comparison, conforming to previous intermodal studies (e.g., Bernardini et al., 2016; Ferraresi et al., 2018), SI shows more simplified and informal patterns than WT, while the explicitness pattern is not straightforward enough. In addition to these patterns under the framework of translation universals, other SI-specific linguistic patterns have also been reported, such as tendencies towards personal involvement, uncertainty expressions, etc..

Moving on to the second analysis on lexical simplification and increased explicitness patterns of SI compared to NS and WT, the author followed the traditional research trajectory by focusing on several selected linguistic features that are found to be most representative of the related patterns, such as STTR, top 10, lexical density and average sentence length for lexical simplification, conjuncts, adverbial subordinators, and optional 'that' usage for increased explicitness. Subgenre comparisons have also been carried out to examine if the identified lexical patterns are consistent across all SI subgenres compared with the corresponding NS and WT subgenres. The comparable analysis on lexical simplification has confirmed the general findings reported in the first analysis, which reports overall more simplified and repetitive patterns of SI, except for lexical density. However, cross-subgenre comparison shows that this simplified pattern of SI is not always consistent, especially with respect to top 10 coverage and lexical density. The intermodal analysis reveals consistent patterns of SI as being more simplified than WT. In terms of increased explicitness, comparable comparison reports overall an increased explicitness in SI, except for 'that' adjective complement which has been markedly underused in interpreting. Intermodal comparison reveals less clear-cut patterns, in that while some of the linguistic indicators under discussion show more explicit patterns in SI, the others show the opposite trend. No conclusive remarks can be made in terms of consistency patterns also. Therefore, the unidimensional analysis can only lend partial support to the findings reported in previous studies.

These general trends aside, two linguistic features are found to be used in similar patterns both across and within (the subgenres of) SI and WT, i.e., 'that' adjective complement and 'that' verb complement. Both mediated language varieties show markedly consistent underuse of 'that' adjective complement, while an overuse of 'that' verb complement (over 'that' omission) compared to unmediated native speech. Due to this marked difference, as well as homogeneity among subgenres, the author tentatively argues that these two linguistic features might be translation-specific (i.e., mediation-specific) features.

6.1.2 Summary of the main findings of the multidimensional analysis

The unidimensional analysis summarized above tries to identify SI-specific linguistic patterns in terms of the distribution of 79 linguistic features, as well as to test two popular translation universal hypotheses. While it reports promising findings, several co-related trends cannot be readily explained. A multidimensional analysis highlighting co-occurrence patterns has been carried out to address this issue.

The multidimensional analysis based on factor analysis identifies altogether eight factors, accounting for about 40% of the total variances among SI, NS, and WT. In the end, the author kept seven factors, as factor 8 groups only three features with relatively small factor loadings, which lacks theoretical foundation for interpretation (Biber, 1988). The seven factors are then interpreted as seven dimensions, based on the assumption that linguistic features co-occur to realize shared communicative functions, after which the SI-specific co-occurrence patterns of linguistic features along the seven dimensions are identified in relation to NS and WT. Analysis of the consistency patterns across subgenre comparison has also been carried out to explore the possible genre influence over the SI-specific patterns along the dimension continuums.

Dimension 1, 'Involved versus Informational Production', captures the linguistic co-occurrence patterns specific to SI compared to NS and WT, in that SI (as a whole) is more marked in features associated with involved, informal, and fragmented language use, such as present tense, contractions, shorter words, *be* as main verb, discourse particles, pronouns, independent clause coordination, etc., features that have positive loadings on Dimension 1, and unmarked in features indicating informational and integrated language production. This seems to have offered sound explanation to part of the findings reported in section 4.1, where both nominal features and features cueing simplification trend and personal involvement have been observed in SI vs. NS. However, (sub-)genre influence needs to be acknowledged, for the SI-specific patterns are not very consistent, in that only subgenre comparison between SI_A and NS_A reveals distinguishable differences

between the two varieties.

Dimension 2, ‘On-line Information Elaboration with Stancetaking Concerns’, reveals the most dramatic variation between SI and NS, while translation and interpreting share great similarity along this dimension. Specifically, SI is marked in terms of an underuse of co-occurring features cueing on-line stancetaking expressions, such as ‘that’ relative clauses on object position, ‘that’ relative clauses on subject position, amplifiers, first person pronouns, ‘that’ adjective complements, demonstratives, private verbs, etc., while an overuse of conditional adverbial subordinators, analytic negation, and second person pronouns indexing conditioned description. Such linguistic patterns are rather consistent among SI subgenres compared to their unmediated counterparts. An interesting pattern between translation and interpreting as two modes of mediation has been observed, which demonstrates great homogeneity between the two in terms of their unmarked manifestation in stancetaking expressions. This finding seems to strongly suggest the shared patterns of mediated languages.

Dimension 3, ‘Precise versus Simplified Description’, reveals sharp intermodal differences, while interpreted language resembles more native spoken language, lending some support to “more spoken than translated” nature of interpreting (Shlesinger & Ordan, 2012). Overall, WT is marked in precise and elaborated description, characterized by split auxiliaries, conjuncts, adverbs, downtoners, and concessive adverbial subordinators, while unmarked in contractions, top 10 coverage, and independent clause coordination. SI, by contrast, shows the opposite trend, which is also slightly marked compared to NS. With respect to the consistency of SI-specific patterns, the intermodal distinction is consistent across subgenre comparisons, while the slightly more marked pattern of simplified description in SI subgenres is levelled out, showing little variation between SI and NS as forms of spoken language along this dimension.

Dimension 4, ‘Narrative versus Abstract Focus’, captures noticeable ontological differences between SI and NS, but not intermodal differences. Given the intermediate dimension score of SI, interpreted language is characterized by relatively balanced use of

grouping features with both narrative and abstract concerns. That is, SI shows no marked preference for either of the complementary co-occurring features, such as third person pronouns, *WH* pronouns, *WH* relative clauses on subject position along the positive side of the continuum, and nominalizations, gerunds, etc. along the negative side. Native spoken language, by contrast, is marked in the use of linguistic features with positive weights, while translation situates at the opposite side of this dimension with relatively intermediate dimension scores. The intermediate positioning of SI along Dimension 4 helps explain one of the findings from the unidimensional analysis in section 4.1., which reports an overuse of both pronouns and nominal features in SI compared with NS. These SI-specific patterns are relatively consistent across subgenre comparisons from a comparable perspective, but the undistinguishable intermodal distinction is challenged in SI_C vs. WT_C ("*Debates*"), indicating possible genre influence and also genre variation.

Dimension 5, 'Lexical versus Functional Concerns', once again demonstrates larger ontological differences than intermodal differences, in that interpreted language in general is slightly marked in lexical features, while slightly unmarked in functional features, a pattern contradictory to that of non-interpreted spoken language. Translated language, similar to interpreted language, is also slightly marked in lexical features vs. functional features, but this pattern is less prominent given its intermediate dimension scores (lower than SI dimension scores). Strong subgenre variations within the three varieties have been observed, where SI_B and WT_B ("*Questions to the Secretaries*") stand out, demonstrating a much stronger lexical concern.

Dimension 6, 'On-line Persuasion', is relatively undistinguishable for the three varieties evidenced by their intermediate dimension scores, although SI is slightly unmarked in on-line persuasion features, compared with NS and WT. However, a closer examination reveals that this SI-specific pattern is not always consistent, due to strong internal variations within SI, where SI_C and SI_B show distinct patterns in relation to the general SI-specific pattern.

Dimension 7, 'Coordinating versus Possessive Functions', reports both ontological and

intermodal distinction across the three varieties globally, with translated language highlighting coordinating function, unmediated spoken language emphasizing possessive function, while interpreted language unmarked in either of the two communicative functions. However, strong internal variations among the three varieties have been observed, which help cancel out the global differences across the three.

To sum up, Dimension 1 seems to be a potential candidate for L2 interpretese, while Dimension 2, capturing the largest ontological variation, might be a shared dimension for translated and interpreted languages. Along the remaining dimensions, there are no absolute difference among translations, interpretations, and non-interpretations, and more often than not, similarities rather than differences are observed.

6.1.3 Implications for the 'universal' features of mediated language

The unidimensional and multidimensional analyses, as recapitulated above, provide both evidence and counterevidence to previous findings regarding the 'universal' patterns or features of language in mediation, particularly with respect to the fourfold TU hypotheses, i.e., simplification, explicitation (in this study, increased explicitness), normalization, and levelling out. Although the current research does not make direct comparisons with previous studies, due to the different research methodologies adopted, certain implications can still be made based on the findings reported in this project.

(1) In terms of lexical simplification, the unidimensional analysis carried out in Chapter four has generally confirmed the overall more simplified nature of interpreted language from both comparable and intermodal perspectives, except for linguistic indicators for information density. The multidimensional analysis also lends partial support to the overall more simplified, involved, and informal nature of interpreting viewed globally. However, genre influence should also be considered when interpreting the results.

(2) In terms of explicitation, or as preferred in this research, increased explicitness, the unidimensional analysis carried out in Chapter four, section 4.2 offers clear evidence to the more explicit nature of interpreted language, compared with non-interpreted spoken originals, while the intermodal comparison reports some inconclusive findings. However, based on the variation patterns of the 79 linguistic features carried out in section 4.1, SI is also found to be characterized by an underuse of linguistic features associated with syntactic elaboration, or syntactic explicitation. Such a finding is partially confirmed in the multidimensional analysis in Dimension 2, 'On-line Information Elaboration with Stancetaking Concerns', where SI is unmarked in the co-occurrence patterns for syntactic elaboration produced on-line. Although the focus (i.e., stancetaking) in Dimension 2 is different, the author argues that it may offer indirect reference to less explicit nature of interpreting. A promising dimension, i.e., Dimension 6 – 'On-line Persuasion', groups together some linguistic features which are often considered as indicators for explicitation, namely, optional 'that' after (certain types of) reporting verbs, the focus here is also different. Therefore, the multidimensional analysis fails to provide straightforward evidence to this universal pattern.

(3) In terms of normalization, the unidimensional analysis done in Chapter four, section 4.1 reports certain distribution patterns of linguistic features, such as an overuse of passive structures in interpreted language, that may have some implications for more normalized language use according to Hu and Tao (2013), but it may also be the result of source language interference, which needs further confirmation based on parallel analysis. The multidimensional analysis may offer some counterevidence to this pattern, manifested by the more marked 'Involved and Informal' nature of SI along Dimension 1. In addition, Pym's interpretation of normalization as less internal variations among translated texts, a view overlapping with the convergence hypothesis (Laviosa, 2002), is also disproved in this study.

(4) In terms of levelling out or convergence patterns, both the unidimensional and the

multidimensional analyses offer strong counterevidence, in that greater internal variations within SI have been reported compared with the corresponding subgenre variations within NS.

Despite the mixed evidence and counterevidence, the author argues that the overarching goal of the current research is not to confirm or refute 'universal' features of translation and interpreting, but rather to uncover and call attention to the multidimensional nature of interpreted language. Thus this research is exploratory rather than confirmatory in nature. The author concurs with Baker's early argument (1999, p. 293) that,

> A detailed description of linguistic features is not an end in itself for a translation scholar: it is merely a means to an end, a first step towards understanding the pressures and constraints under which translators operate and which inevitably leave traces in the language they produce.

Interpreted language as a multi-constrained language variety produced under interwoven constraints is expected to be characterized by linguistic patterns that might differ from unmediated spoken language and translated language. However, these patterns are "[f]ar from being laws that have to be obeyed in order to escape punishment, there are ideas to be pursued, played with, experimented upon, and thereby extended into an open-ended beyond" (Pym, 2008b, p. 315). The search for these specific patterns can cast light on the very nature of interpreting as constrained language production, which in turn can inform us the underlying constraints behind (see also Chesterman, 2004).

6.2 Innovations and significance of the study

Investigation of linguistic features in translated language is not a new endeavor, especially after the wide application of large-scale machine-readable corpora. However, this is not the same case when it comes to interpreted language, as compiling a large-scale interpreting corpus is a much more daunting task (Sandrelli & Bendazzoli, 2005; Shlesinger, 1998). One of the innovations of the current research lies in the construction

of a new large-scale (million-size) parallel, intermodal comparable corpus. Composed of four sub-corpora, including source texts, interpreted texts, translated texts, and non-translated spoken originals, this corpus can be utilized to investigate a number of topics, such as 'universal' features/patterns of interpreting, SI strategies, interpreting norms, and even process-oriented studies on cognitive load, thanks to the paralinguistic annotations (i.e., fillers, repairs, and truncated words) in this corpus.

Apart from the construction of a new translation and interpreting corpus, this research is innovative in that it contributes to the ongoing debates in corpus-based studies on the typical features of translation and interpreting, such as translationese, translation universals, interpretese, and interpreting universals. The emphasis on the L2 aspect of interpreted language is particularly illuminating. Besides, this research highlights the spoken mode of SI by adopting both comparable and intermodal perspectives, instead of categorizing SI as one of the generic translations, which will otherwise offer a limited picture about the very nature of interpreting as being both spoken and mediated language.

In addition, a multivariate instead of the prevailing univariate approach has been adopted, which can be considered as a methodological innovation in corpus-based interpreting studies. The multivariate and multidimensional analysis allows for the identification of co-occurrence patterns of linguistic features in actual language use, based on their shared communicative functions. In this way, it helps uncover the interdependence of linguistic features under discussion, along with the underlying dimensions or constructs that a univariate analysis fails to identify.

This research is believed to be significant both theoretically and practically. Theoretically speaking, by examining L2 interpreting-specific linguistic features, the research, first of all, undoubtedly can help facilitate our understanding of the nature of L2 interpreting. Focusing on interpreting per se, it also helps enrich corpus-based interpreting studies as a major branch of interpreting studies. Secondly, as the corpus under investigation is based on the proceedings of the Hong Kong Legislative Council, this research can help unveil the interpreting norms and constraints in legislative settings. As suggested also by

Laviosa (1989, p.474), the aim of translationese studies is "not merely to unveil the 'third code' per se, but most importantly, to understand the specific constraints, pressures and motivations that influence the act of translating and underlie its unique language". The same applies to interpreting. Thirdly, by examining interpreting-specific linguistic features, the research can also shed light on the cognitive processing of SI under such constrained conditions. For example, Defrancq and Plevoets (2016, 2018) argue that the use of filled pauses 'ehm' may indicate high cognitive load experienced by simultaneous interpreters. Based on the tagged paralinguistic features in the corpus, it is possible to uncover the cognitive constraints experienced by interpreters working from A-to-B language direction.

In practical terms, the corpus under study can be utilized for multiple purposes, such as examining the interpreters' strategies based on parallel sub-corpora, as translation and interpreting are done by professional translators and interpreters at the Hong Kong Legislative Council, which has guaranteed the translation quality. The identified linguistic patterns of L2 interpreting can further shed light on interpreting teaching and training. Student interpreters can have a general knowledge about the typical patterns of interpreting into a B language, while interpreting trainers and teachers can also educate their students about the possible underlying constraints experienced by professional translators and interpreters based on this corpus. Awareness of these can help student interpreters produce more natural-sounding outputs, while at the same time facilitate their understanding regarding practice and theoretical motivations.

De Sutter and Lefer (2020, p. 2) point out three factors that hinder the conceptual and theoretical progress in corpus-based translation studies, including "the strong focus of corpus-based translation scholars on finding linguistic differences rather than similarities, the lack of (advanced) statistical testing and the restricted collaboration with scholar from other fields." The current research strives to cross these barriers, in that more similarities than differences have been observed and emphasized, although the starting point is a search for differences; a powerful multivariate statistical technique has been utilized to

uncover the underlying constructs for the three language varieties; and an interdisciplinary point of view from corpus linguistics and register studies has been adopted as one of the major methods.

6.3 Limitations and future directions

6.3.1 Limitations

This study is among the first ones to apply Biber's (1988) MD approach to the investigation of interpreting-specific linguistic patterns in relation to unmediated spoken language and translated language from the same source. It offers sufficient evidence to the multifaceted nature of simultaneous interpreting (into B), and it demonstrates both differences and similarities across the three varieties from a multidimensional perspective. Despite its innovations and theoretical as well as practical significance, several limitations concerning both the corpus data (i.e., the LegCo+ corpus) and the MD approach adopted are acknowledged here.

In terms of the corpus data, the LegCo+ corpus is not representative of the population of simultaneous interpreting, since it contains only three genres, two out of which are closely related (i.e., both Q&As), and thus the findings reported here cannot be generalized to simultaneous interpreting carried out in other settings. Laviosa (2004, p.13) argues that "[…] a corpus intended to be representative of the population of translated texts will consist of an array of subcorpora presenting differing degrees of relevance but all being regarded as legitimate objects of investigation". The author here defends that, the LegCo+ corpus was not intended to be a fully balanced and representative corpus during its construction phrase, as compiling a(n) (comparable and intermodal) interpreting corpus is already a much more daunting task than constructing any kind of translational and interpreting corpus (Bernardini et al., 2018; Shlesinger, 1998). The inclusion of diversified genres adds further difficulty, since interpretations differ substantially from translations in terms of the diversity of genres with respect to accessibility (e.g., confidentiality issues, interpreters' reluctance to be observed, etc.). This also explains

why the most representative interpreting corpora are almost unanimously based on proceedings in the European Parliament, or the Press Conferences in China. Moreover, as acknowledged by Baroni and Bernardini (2006, p. 264),

> [t]he generality of results should be demonstrated through an accumulation of findings from several experiments, each based on a small homogeneous corpus, rather than through a single experiment with a large varied corpus, where confounding factors would be difficult to control.

Although the size of the LegCo+ corpus is relatively large, the subgenres included are rather homogeneous, since they are taken from legislative settings of a similar nature (i.e., the Hong Kong LegCo and the UK Parliament). Future comparisons can also be made in relation to studies based on corpus data from the European Parliament, which may help shed light on the shared linguistic (co-occurrence) patterns of simultaneous interpreting in legislative/parliamentary settings.

Another limitation about the corpus is its inadequate metadata, particularly metadata about simultaneous interpreters and translators (e.g., age, gender, working experience, working mode, preparedness). Some information, such as gender and preparedness, can be tentatively accessed based on video recordings and transcriptions (with paralinguistic features annotated), as been practiced in Bernardini et al. (2016). It is still acknowledged that such a way lacks objectivity.

In terms of methodological deficiency, the selection of linguistic features to be included in the MD analysis can have a direct influence on the outcome, with respect to the linguistic co-occurrence patterns and the potential dimensions to be extracted. Although the inclusion of 79 linguistic features in the current research have rich theoretical foundations (i.e., register variation and 'translation universals'), there are still many other features discussed previously that have yet to be included, which may eventually have an influence on the linguistic co-occurrence patterns specific to SI as reported here. Besides, as also been acknowledged in Chapter five, interpretations of the dimensions extracted through multidimensional analysis are tentative and need further confirmation. Further

exploitation of confirmatory factor analysis may serve as a complementary way to the current exploratory factor analysis featuring the MD analysis.

6.3.2 Future directions

Future studies may extend the research scope by including other L2 English varieties, such as Hong Kong English, for a direct comparison between interpreted English and L2 English, as well as native English, to identify L2-specific effects under the framework of constrained language from an interdisciplinary perspective. As suggested by Halverson (2003, p. 227), studies in second language acquisition "have provided evidence very similar to that found in translation studies", given the constrained nature of L2/non-native language and interpreted language as bilingual language production. A recent study by Kajzer-Wietrzny (2018) reports shared patterns of non-nativeness between non-native English and interpreted English. It thus seems promising that such an interdisciplinary perspective may help further reveal the multidimensional nature of interpreting.

Another language variety, i.e., non-mediated or native written English, produced in a similar setting, can also be included to identify L2 *translationese*, i.e., linguistic patterns specific to L2 translation (translation carried out from A to B language direction). Further intermodal comparisons between L2 interpreting and L2 translation, with respect to native spoken language and native written language, can also be done to inform the "equalizing effect" proposed by Shlesinger (1989) with respect to their positioning along the oral-literate continuum.

Besides the enrichment of research data, research scope can also be extended in terms of the research methods adopted. Univariate studies replicating previous studies on translationese and interpretese (B-to-A working direction) can be considered to make more direct comparison between L1 and L2 interpreting so as to inform the possible influence of working direction on the linguistic manifestations of interpreted language.

Another line of research can go for a multifactorial analysis utilizing more sophisticated

but revealing techniques to disentangle the possible contributors to the identified linguistic patterns, as done in Kruger and De Sutter (2018). This line of research is particularly promising, thanks to the increasing awareness of interdisciplinary collaboration from translation scholars, as well as the development of more advanced research tools. The promising multifactorial analysis may, in the end, help enlighten the various constraints that translators and interpreters go through, the results of which can be fed into translator and interpreter training in future courses.

Appendices

Appendix 1 Initial version of transcription symbols following Tang (2014)

Features	Symbols	Examples
Short pause (shorter than 2s)	<p>	The teachers in China have reached the number of <p> one point six million
Long pause (longer than 2s)	<p><p>	we've also set up <p><p> new liaison points
Stammer/hesitation	<uh>	the faculty's <uh> training program
Stretched pronunciation	~	for the~ festival
Unusual pronunciation	spelling*	motoblize* (should be mobilize)
False start	spelling*-	The top teachers ex*- extended their congratulations
Intonation	full stop	Yesterday was out Teacher's Day
Intonation	Question mark	What is the key point?
Intonation	Exclamation mark	Thank you for your presence!

Appendix 2 Tests of normality (for NS, SI, and WT respectively)

Tests of Normality (for NS)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AWL	.073	138	.073	.986	138	.195
TTR	.117	138	.000	.894	138	.000
AMP	.081	138	.027	.963	138	.001
ANDC	.069	138	.199	.986	138	.184
CAUS	.122	138	.000	.917	138	.000
CONC	.404	138	.000	.664	138	.000
COND	.113	138	.000	.908	138	.000
CONJ	.134	138	.000	.893	138	.000
DEMO	.109	138	.000	.972	138	.006
DEMP	.086	138	.013	.955	138	.000
DPAR	.108	138	.000	.955	138	.000
DWNT	.110	138	.000	.949	138	.000
EMPH	.058	138	.200*	.979	138	.034
EX	.098	138	.003	.932	138	.000
FPP1	.043	138	.200*	.992	138	.629
GER	.104	138	.001	.943	138	.000
HDG	.445	138	.000	.520	138	.000
INPR	.399	138	.000	.635	138	.000
JJ	.077	138	.044	.979	138	.033
NEMD	.132	138	.000	.860	138	.000
NN	.074	138	.065	.985	138	.129
NOMZ	.097	138	.003	.974	138	.010
OSUB	.131	138	.000	.929	138	.000
PHC	.083	138	.022	.960	138	.001
PIN	.050	138	.200*	.993	138	.706
PIT	.086	138	.014	.963	138	.001
PLACE	.110	138	.000	.949	138	.000
POMD	.069	138	.100	.982	138	.059
PRED	.053	138	.200*	.987	138	.204
PRMD	.085	138	.016	.974	138	.009
RB	.061	138	.200*	.984	138	.117
SPP2	.099	138	.002	.927	138	.000
SYNE	.168	138	.000	.899	138	.000
THAC	.147	138	.000	.900	138	.000
THVC	.111	138	.000	.965	138	.001
TIME	.085	138	.017	.914	138	.000
TO	.055	138	.200*	.992	138	.608

TOBJ	.073	138	.070	.981	138	.054
TPP3	.050	138	.200*	.989	138	.332
TSUB	.119	138	.000	.947	138	.000
VBD	.131	138	.000	.890	138	.000
VPRT	.052	138	.200*	.992	138	.623
XX0	.133	138	.000	.931	138	.000
[BEMA]	.061	138	.200*	.989	138	.325
[BYPA]	.164	138	.000	.894	138	.000
[CONT]	.094	138	.005	.969	138	.003
[PASS]	.042	138	.200*	.987	138	.218
[PASTP]	.433	138	.000	.585	138	.000
[PEAS]	.087	138	.012	.974	138	.010
[PIRE]	.239	138	.000	.792	138	.000
[PRESP]	.162	138	.000	.892	138	.000
[PRIV]	.044	138	.200*	.992	138	.621
[PROD]	.084	138	.019	.933	138	.000
[PUBV]	.090	138	.008	.930	138	.000
[SERE]	.158	138	.000	.930	138	.000
[SMP]	.376	138	.000	.644	138	.000
[SPAU]	.061	138	.200*	.984	138	.117
[SPIN]	.475	138	.000	.531	138	.000
[STPR]	.144	138	.000	.858	138	.000
[SUAV]	.086	138	.014	.954	138	.000
[THATD]	.075	138	.052	.949	138	.000
[WHCL]	.205	138	.000	.883	138	.000
[WHOBJ]	.302	138	.000	.715	138	.000
[WHQU]	.262	138	.000	.771	138	.000
[WHSUB]	.131	138	.000	.923	138	.000
[WZPAST]	.177	138	.000	.868	138	.000
[WZPRES]	.106	138	.001	.956	138	.000
CC	.068	138	.200*	.975	138	.012
DT	.092	138	.006	.972	138	.006
IN	.051	138	.200*	.977	138	.021
POS	.106	138	.001	.924	138	.000
RP	.144	138	.000	.858	138	.000
WP	.068	138	.200*	.982	138	.065
STTR	.056	138	.200*	.990	138	.398
ASL	.072	138	.075	.975	138	.013
TOP10	.071	138	.082	.989	138	.338
LD	.073	138	.072	.987	138	.214
SW	.046	138	.200*	.991	138	.492
LW	.066	138	.200*	.984	138	.119

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tests of Normality (for SI)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AWL	.152	149	.000	.906	149	.000
TTR	.085	149	.010	.984	149	.074
AMP	.095	149	.002	.942	149	.000
ANDC	.056	149	.200*	.976	149	.010
CAUS	.111	149	.000	.936	149	.000
CONC	.457	149	.000	.529	149	.000
COND	.079	149	.024	.963	149	.000
CONJ	.135	149	.000	.924	149	.000
DEMO	.042	149	.200*	.988	149	.232
DEMP	.064	149	.200*	.983	149	.058
DPAR	.114	149	.000	.918	149	.000
DWNT	.153	149	.000	.874	149	.000
EMPH	.090	149	.005	.975	149	.007
EX	.103	149	.001	.962	149	.000
FPP1	.058	149	.200*	.984	149	.082
GER	.151	149	.000	.832	149	.000
HDG	.360	149	.000	.684	149	.000
INPR	.417	149	.000	.603	149	.000
JJ	.119	149	.000	.957	149	.000
NEMD	.119	149	.000	.837	149	.000
NN	.061	149	.200*	.978	149	.019
NOMZ	.132	149	.000	.922	149	.000
OSUB	.102	149	.001	.927	149	.000
PHC	.141	149	.000	.892	149	.000
PIN	.082	149	.015	.959	149	.000
PIT	.115	149	.000	.944	149	.000
PLACE	.163	149	.000	.900	149	.000
POMD	.085	149	.010	.966	149	.001
PRED	.061	149	.200*	.982	149	.052
PRMD	.069	149	.077	.980	149	.028
RB	.061	149	.200*	.992	149	.578
SPP2	.123	149	.000	.916	149	.000
SYNE	.189	149	.000	.759	149	.000
THAC	.315	149	.000	.693	149	.000
THVC	.080	149	.022	.978	149	.019
TIME	.099	149	.001	.917	149	.000
TO	.058	149	.200*	.987	149	.196
TOBJ	.187	149	.000	.890	149	.000
TPP3	.103	149	.001	.944	149	.000

TSUB	.286	149	.000	.812	149	.000
VBD	.085	149	.010	.941	149	.000
VPRT	.058	149	.200*	.987	149	.166
XX0	.036	149	.200*	.988	149	.242
[BEMA]	.054	149	.200*	.989	149	.280
[BYPA]	.200	149	.000	.876	149	.000
[CONT]	.053	149	.200*	.983	149	.067
[PASS]	.076	149	.037	.975	149	.008
[PASTP]	.486	149	.000	.355	149	.000
[PEAS]	.071	149	.065	.953	149	.000
[PIRE]	.433	149	.000	.499	149	.000
[PRESP]	.169	149	.000	.856	149	.000
[PRIV]	.041	149	.200*	.985	149	.104
[PROD]	.145	149	.000	.918	149	.000
[PUBV]	.065	149	.200*	.970	149	.002
[SERE]	.426	149	.000	.553	149	.000
[SMP]	.465	149	.000	.546	149	.000
[SPAU]	.063	149	.200*	.983	149	.060
[SPIN]	.435	149	.000	.555	149	.000
[STPR]	.244	149	.000	.616	149	.000
[SUAV]	.084	149	.012	.877	149	.000
[THATD]	.076	149	.035	.966	149	.001
[WHCL]	.194	149	.000	.914	149	.000
[WHOBJ]	.507	149	.000	.386	149	.000
[WHQU]	.232	149	.000	.770	149	.000
[WHSUB]	.149	149	.000	.889	149	.000
[WZPAST]	.199	149	.000	.764	149	.000
[WZPRES]	.123	149	.000	.952	149	.000
CC	.035	149	.200*	.992	149	.581
DT	.060	149	.200*	.987	149	.197
IN	.049	149	.200*	.993	149	.716
POS	.169	149	.000	.763	149	.000
RP	.106	149	.000	.943	149	.000
WP	.111	149	.000	.930	149	.000
STTR	.073	149	.051	.948	149	.000
ASL	.199	149	.000	.872	149	.000
TOP10	.051	149	.200*	.993	149	.648
LD	.057	149	.200*	.990	149	.375
SW	.072	149	.054	.986	149	.128
LW	.169	149	.000	.899	149	.000

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Tests of Normality (for WT)

	Kolmogorov-Smirnov[a]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AWL	.104	190	.000	.961	189	.000
TTR	.089	190	.001	.969	189	.000
AMP	.127	190	.000	.927	189	.000
ANDC	.062	189	.075	.979	189	.006
CAUS	.186	189	.000	.881	189	.000
CONC	.278	189	.000	.722	189	.000
COND	.072	189	.019	.975	189	.002
CONJ	.068	189	.031	.979	189	.006
DEMO	.075	189	.012	.976	189	.003
DEMP	.072	189	.018	.964	189	.000
DPAR	.442	189	.000	.560	189	.000
DWNT	.118	189	.000	.948	189	.000
EMPH	.067	189	.036	.966	189	.000
EX	.137	189	.000	.905	189	.000
FPP1	.047	189	.200*	.990	189	.205
GER	.167	189	.000	.830	189	.000
HDG	.540	189	.000	.186	189	.000
INPR	.406	189	.000	.636	189	.000
JJ	.070	189	.027	.972	189	.001
NEMD	.094	189	.000	.925	189	.000
NN	.061	189	.084	.984	189	.031
NOMZ	.081	189	.005	.968	189	.000
OSUB	.117	189	.000	.949	189	.000
PHC	.103	189	.000	.938	189	.000
PIN	.049	189	.200*	.992	189	.359
PIT	.060	189	.099	.959	189	.000
PLACE	.136	189	.000	.941	189	.000
POMD	.065	189	.053	.985	189	.041
PRED	.073	189	.015	.984	189	.030
PRMD	.075	189	.012	.972	189	.001
RB	.053	189	.200*	.988	189	.095
SPP2	.213	189	.000	.774	189	.000
SYNE	.153	189	.000	.919	189	.000
THAC	.295	189	.000	.797	189	.000
THVC	.062	189	.076	.972	189	.001
TIME	.079	189	.006	.974	189	.001
TO	.045	189	.200*	.992	189	.365
TOBJ	.159	189	.000	.917	189	.000
TPP3	.111	189	.000	.889	189	.000
TSUB	.277	189	.000	.746	189	.000

VBD	.088	189	.001	.931	189	.000
VPRT	.040	189	.200*	.994	189	.607
XX0	.073	189	.017	.984	189	.028
[BEMA]	.053	189	.200*	.986	189	.055
[BYPA]	.180	189	.000	.904	189	.000
[CONT]	.512	189	.000	.372	189	.000
[PASS]	.076	189	.010	.972	189	.001
[PASTP]	.411	189	.000	.594	189	.000
[PEAS]	.071	189	.022	.973	189	.001
[PIRE]	.353	189	.000	.677	189	.000
[PRESP]	.105	189	.000	.940	189	.000
[PRIV]	.062	189	.071	.970	189	.001
[PROD]	.207	189	.000	.811	189	.000
[PUBV]	.093	189	.000	.944	189	.000
[SERE]	.198	189	.000	.892	189	.000
[SMP]	.456	189	.000	.553	189	.000
[SPAU]	.058	189	.200*	.980	189	.008
[SPIN]	.369	189	.000	.637	189	.000
[STPR]	.236	189	.000	.818	189	.000
[SUAV]	.095	189	.000	.896	189	.000
[THATD]	.164	189	.000	.930	189	.000
[WHCL]	.265	189	.000	.812	189	.000
[WHOBJ]	.402	189	.000	.614	189	.000
[WHQU]	.232	189	.000	.736	189	.000
[WHSUB]	.159	189	.000	.907	189	.000
[WZPAST]	.100	189	.000	.945	189	.000
[WZPRES]	.113	189	.000	.958	189	.000
CC	.103	189	.000	.967	189	.000
DT	.044	189	.200*	.990	189	.214
IN	.057	189	.200*	.987	189	.074
POS	.132	189	.000	.868	189	.000
RP	.111	189	.000	.941	189	.000
WP	.114	189	.000	.927	189	.000
STTR	.062	189	.070	.992	189	.415
ASL	.097	189	.000	.940	189	.000
TOP10	.073	189	.015	.971	189	.001
LD	.030	189	.200*	.994	189	.626
SW	.065	189	.048	.980	189	.008
LW	.086	189	.002	.973	189	.001

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

Appendix 3 79 linguistic features analyzed

(A)	TENSE AND ASPECT MARKERS	
1	VBD	Past tense
2	PEAS	Perfect aspect
3	VPRT	Present tense
(B)	PLACE AND TIME ADVERBIALS	
4	PLACE	Place adverbials
5	TIME	Time adverbials
(C)	PRONOUNS AND PROVERBS	
6	FPP1	First person pronouns
7	SPP2	Second person pronouns
8	TPP3	Third person pronouns
9	PIT	Pronoun it
10	DEMP	Demonstrative pronouns
11	INPR	Indefinite pronouns
12	PROD	Pro-verb do
(D)	QUESTIONS	
13	WHQU	WH-questions
(E)	NOMINAL FORMS	
14	NOMZ	Nominalizations
15	GER	Gerunds
16	NN	Nouns
(F)	PASSIVES	
17	PASS	Agentless passives
18	BYPA	By-passives
(G)	STATIVE FORMS	
19	BEMA	Be as main verb
20	EX	Existential there
(H)	SUBORDINATION	
21	THVC	That verb complements
22	THAC	That adjective complements
23	WHCL	WH-clauses
24	TO	Infinitives
25	PRESP	Present participial clauses
26	PASTP	Past participial clauses
27	WZPAST	Past participial WHIZ deletion relatives
28	WZPRES	Present participial WHIZ deletion relatives
29	TSUB	That relative clauses on subjection position
30	TOBJ	That relative clause on object position
31	WHSUB	WH relative clauses on subject position
32	WHOBJ	WH relative clauses on object position
33	PIRE	Pied-piping relative clauses
34	SERE	Sentence relatives
35	CAUS	Causative adverbial subordinators
36	CONC	Concessive adverbial subordinators
37	COND	Conditional adverbial subordinators
38	OSUB	Other adverbial subordinators
(I)	PREPOSITIONAL PHRASES, ADJECTIVE AND ADVERBS	
39	PIN	Total prepositional phrases
40	ATTRJ	Attributive adjectives
41	PRED	Predictive adjectives
42	R	Total adverbs
(J)	LEXICAL SPECIFICITY	
43	TTR	Type/token ratio
44	AWL	Average word length
(K)	LEXICAL CLASS	
45	CONJ	Conjuncts
46	DWNT	Downtoners
47	HDG	Hedges
48	AMP	Amplifiers
49	EMPH	Emphatics
50	DPAR	Discourse particles
51	DEMO	Demonstratives
(L)	MODALS	
52	POMD	Possibility modals
53	NEMD	Necessity modals
54	PRMD	Predictive modals
(M)	SPECIALIZED VERB CLASSES	
55	PUBV	Public verbs
56	PRIV	Private verbs
57	SUAV	Suasive verbs
58	SMP	Seem/appear
(N)	REDUCED FORMS AND DISPREFERED STRUCTURES	
59	CONT	contractions
60	THATD	Subordinator-that deletion
61	STPR	Stranded preposition
62	SPIN	Split infinitives
63	SPAU	Split auxiliaries
(O)	COORDINATION	
64	PHC	Phrasal coordination
65	ANDC	Independent clause coordination
(P)	NEGATION	
66	SYNE	Synthetic negation
67	XX0	Analytic negation
(Q)	OVERALL TEXTUAL FEATURES	
68	STTR	Standard type/token ratio
69	ASL	Average sentence length
70	TOP10	Top 10 vocabulary coverage
71	LD	Lexical density
72	SW	Shorter words (<=3)
73	LW	Longer words (>=7)
74	CC	Coordinating conjunctions
(R)	OTHER FEATURES	
75	DT	Determiner ‘the’
76	IN	Preposition or subordinating

		conjunction
77	POS	Possessive endings
78	RP	Particles
79	WP	WH-pronoun

Appendix 4 Descriptive statistics for NS, SI, and WT

Descriptive Statistics (for NS)

	N	Range	Minimum	Maximum	Mean	Std. Deviation
AWL	138	.8100	4.2700	5.0800	4.587609	.1380128
TTR	138	38	43	81	69.56	5.856
AMP	138	1.0700	.1100	1.1800	.530870	.2424554
ANDC	138	1.3200	.4800	1.8000	1.094058	.2845066
CAUS	138	.6400	.0000	.6400	.165507	.1305411
CONC	138	.1700	.0000	.1700	.025145	.0395086
COND	138	1.0400	.0000	1.0400	.250290	.1852122
CONJ	138	.5300	.0000	.5300	.132029	.1027232
DEMO	138	1.9100	.8700	2.7800	1.615000	.4151409
DEMP	138	1.9100	.3000	2.2100	.905217	.3200187
DPAR	138	.5800	.0000	.5800	.214638	.1421943
DWNT	138	.4600	.0000	.4600	.145217	.1020463
EMPH	138	1.6300	.1700	1.8000	.782681	.2924888
EX	138	.9900	.0000	.9900	.282391	.1648293
FPP1	138	4.5200	2.6300	7.1500	4.541812	.8680133
GER	138	1.4600	.0000	1.4600	.420145	.2600842
HDG	138	.2500	.0000	.2500	.019130	.0416534
INPR	138	.2400	.0000	.2400	.026884	.0465260
JJ	138	4.1600	4.1100	8.2700	6.059348	.8296101
NEMD	138	1.2200	.0000	1.2200	.241594	.1977510
NN	138	11.1400	15.3900	26.5300	21.152971	2.3299245
NOMZ	138	3.8900	1.2400	5.1300	2.957246	.8191573
OSUB	138	.5100	.0000	.5100	.147464	.1126785
PHC	138	1.3200	.1200	1.4400	.620797	.2405288
PIN	138	5.2200	7.3500	12.5700	9.876377	.9273752
PIT	138	2.2800	.3300	2.6100	1.061159	.4070135
PLACE	138	.8800	.0000	.8800	.291232	.1907092
POMD	138	1.0700	.1800	1.2500	.661377	.2270187
PRED	138	1.2400	.3100	1.5500	.898986	.2747296
PRMD	138	2.3600	.3700	2.7300	1.273986	.4158968
RB	138	3.2500	1.6900	4.9400	3.216304	.6664221
SPP2	138	1.6000	.0000	1.6000	.467826	.2827336
SYNE	138	.4500	.0000	.4500	.101159	.0905867
THAC	138	.6500	.0000	.6500	.174348	.1392732
THVC	138	1.0000	.0000	1.0000	.436594	.2011191
TIME	138	1.4800	.0600	1.5400	.481884	.2423924
TO	138	2.5400	1.0400	3.5800	2.308696	.5127638
TOBJ	138	.7200	.0000	.7200	.342899	.1674375

TPP3	138	3.7300	.5300	4.2600	2.169348	.7406844
TSUB	138	.6200	.0000	.6200	.197536	.1231865
VBD	138	3.9300	.4200	4.3500	1.425870	.6897357
VPRT	138	5.7700	4.0000	9.7700	7.035942	.9778299
XX0	138	1.7200	.1900	1.9100	.726739	.3391577
[BEMA]	138	1.9700	1.1500	3.1200	2.014493	.4001662
[BYPA]	138	.3200	.0000	.3200	.074275	.0649074
[CONT]	138	2.5100	.2400	2.7500	1.326087	.4874115
[PASS]	138	1.5500	.2600	1.8100	.877246	.3082107
[PASTP]	138	.2000	.0000	.2000	.021087	.0399212
[PEAS]	138	1.2300	.2000	1.4300	.682029	.2700423
[PIRE]	138	.3000	.0000	.3000	.051667	.0603588
[PRESP]	138	.3800	.0000	.3800	.099638	.0863824
[PRIV]	138	2.1400	.5400	2.6800	1.562754	.4090576
[PROD]	138	.8300	.0000	.8300	.204783	.1323246
[PUBV]	138	1.5800	.2800	1.8600	.788188	.2935194
[SERE]	138	.5700	.0000	.5700	.153043	.1070071
[SMP]	138	.2300	.0000	.2300	.030870	.0519191
[SPAU]	138	.8900	.0600	.9500	.399493	.1739969
[SPIN]	138	.1300	.0000	.1300	.014855	.0307625
[STPR]	138	.6200	.0000	.6200	.125145	.1131580
[SUAV]	138	1.1200	.1600	1.2800	.619565	.2766955
[THATD]	138	1.3200	.0600	1.3800	.427609	.1934646
[WHCL]	138	.3600	.0000	.3600	.084348	.0720093
[WHOBJ]	138	.3100	.0000	.3100	.041957	.0599009
[WHQU]	138	.3500	.0000	.3500	.054710	.0697356
[WHSUB]	138	.7600	.0000	.7600	.213986	.1454193
[WZPAST]	138	.4200	.0000	.4200	.100652	.0812243
[WZPRES]	138	.8400	.0000	.8400	.269783	.1502964
CC	138	2.1400	1.0000	3.1400	1.742246	.3838891
DT	138	6.2100	5.6200	11.8300	8.075435	1.0125549
IN	138	1.7800	.6400	2.4200	1.327754	.3546439
POS	138	.9300	.0600	.9900	.298116	.1697940
RP	138	.7400	.0000	.7400	.154275	.1382048
WP	138	1.4200	.0600	1.4800	.690290	.2973728
STTR	138	16.5000	30.7000	47.2000	39.502174	2.6900321
ASL	138	13.7328	13.9903	27.7231	20.075258	2.7193250
TOP10	138	9.1300	21.2400	30.3700	25.715507	1.6456619
LD	138	9.8300	49.2700	59.1000	54.326884	1.9478275
Valid N (listwise)	138					

Descriptive Statistics (for SI)

	N	Range	Minimum	Maximum	Mean	Std. Deviation
AWL	149	1.1900	4.2200	5.4100	4.596040	.2564698
TTR	149	28	51	79	66.59	4.881
AMP	149	.6700	.0000	.6700	.222483	.1593861
ANDC	149	2.1800	.2600	2.4400	1.086107	.4162197
CAUS	149	.6400	.0000	.6400	.197517	.1483053
CONC	149	.2200	.0000	.2200	.021141	.0456711
COND	149	1.2500	.0600	1.3100	.444362	.2493433
CONJ	149	.6600	.0000	.6600	.209128	.1494401
DEMO	149	1.9700	.3100	2.2800	1.151946	.3900064
DEMP	149	1.4700	.1900	1.6600	.843221	.3365631
DPAR	149	1.4700	.0000	1.4700	.385906	.3202702
DWNT	149	.6600	.0000	.6600	.134966	.1321068
EMPH	149	1.7000	.0700	1.7700	.733557	.3308631
EX	149	1.3700	.0000	1.3700	.534094	.2957012
FPP1	149	6.8800	.4000	7.2800	3.546913	1.2181210
GER	149	2.8600	.0000	2.8600	.638859	.5074799
HDG	149	.3400	.0000	.3400	.040403	.0640933
INPR	149	.2900	.0000	.2900	.026443	.0490184
JJ	149	6.9800	2.8900	9.8700	5.170872	1.3282222
NEMD	149	2.6500	.0000	2.6500	.421745	.3567737
NN	149	15.6000	15.0500	30.6500	21.757517	2.6224993
NOMZ	149	7.5900	1.2500	8.8400	3.741141	1.2778759
OSUB	149	.7900	.0000	.7900	.191141	.1373914
PHC	149	1.9800	.0700	2.0500	.620537	.3760351
PIN	149	7.9500	5.7300	13.6800	9.002953	1.5384572
PIT	149	2.8100	.0700	2.8800	.990604	.5371650
PLACE	149	.6500	.0000	.6500	.165638	.1397165
POMD	149	1.9200	.0700	1.9900	.791544	.3609966
PRED	149	1.5700	.0700	1.6400	.742953	.3137577
PRMD	149	2.5800	.3300	2.9100	1.238591	.4549859
RB	149	3.7300	1.2400	4.9700	3.016174	.7166434
SPP2	149	4.1000	.0000	4.1000	1.263960	1.0138872
SYNE	149	.9100	.0000	.9100	.149866	.1582292
THAC	149	.4500	.0000	.4500	.047852	.0729269
THVC	149	1.0800	.0000	1.0800	.541879	.2604813
TIME	149	1.5400	.0000	1.5400	.380604	.2385747
TO	149	3.4700	.8500	4.3200	2.223557	.5471993
TOBJ	149	.3900	.0000	.3900	.097315	.0871053
TPP3	149	5.4400	.1300	5.5700	1.673893	.8868778
TSUB	149	.2400	.0000	.2400	.053691	.0613513
VBD	149	5.1700	.0700	5.2400	1.522953	.7507982

VPRT	149	7.9400	2.0800	10.0200	6.705436	1.4312882
XX0	149	2.8100	.1300	2.9400	1.193490	.5171882
[BEMA]	149	2.9900	.4600	3.4500	1.793289	.5967688
[BYPA]	149	.4300	.0000	.4300	.103221	.1008496
[CONT]	149	3.0700	.0600	3.1300	1.205906	.6353835
[PASS]	149	2.2500	.2600	2.5100	1.038389	.3842509
[PASTP]	149	.3500	.0000	.3500	.012550	.0389066
[PEAS]	149	2.1800	.1400	2.3200	.759866	.3380668
[PIRE]	149	.3200	.0000	.3200	.020738	.0457829
[PRESP]	149	.4600	.0000	.4600	.099195	.1030843
[PRIV]	149	2.1700	.2600	2.4300	1.246577	.3662092
[PROD]	149	.8300	.0000	.8300	.207718	.1702308
[PUBV]	149	2.1200	.0600	2.1800	.848993	.3884928
[SERE]	149	.3300	.0000	.3300	.022550	.0444205
[SMP]	149	.1900	.0000	.1900	.018121	.0373148
[SPAU]	149	.9500	.0600	1.0100	.414832	.1891575
[SPIN]	149	.2400	.0000	.2400	.021946	.0443219
[STPR]	149	1.2900	.0000	1.2900	.118054	.1704790
[SUAV]	149	2.3200	.0600	2.3800	.571409	.3139053
[THATD]	149	.8400	.0000	.8400	.318389	.1897084
[WHCL]	149	.3400	.0000	.3400	.096711	.0802651
[WHOBJ]	149	.2000	.0000	.2000	.009732	.0282830
[WHQU]	149	.6200	.0000	.6200	.092282	.1190737
[WHSUB]	149	.4900	.0000	.4900	.127718	.1107598
[WZPAST]	149	.9900	.0000	.9900	.137718	.1548239
[WZPRES]	149	.7500	.0000	.7500	.230805	.1612054
CC	149	2.5900	.7400	3.3300	1.925906	.4616027
DT	149	8.4300	5.1800	13.6100	8.423289	1.3927653
IN	149	2.2700	.5000	2.7700	1.549933	.4448588
POS	149	1.8200	.0000	1.8200	.290940	.2668291
RP	149	1.0300	.0000	1.0300	.299530	.2038360
WP	149	1.2900	.0000	1.2900	.396779	.2773262
STTR	149	19.9000	23.4000	43.3000	36.557382	2.6980372
ASL	149	19.0875	11.0350	30.1224	16.802661	3.5462877
TOP10	149	11.8100	20.3900	32.2000	26.189664	2.1236988
LD	149	9.9400	49.9400	59.8800	54.972081	2.0564717
Valid N (listwise)	149					

Descriptive Statistics (for WT)

	N	Range	Minimum	Maximum	Mean	Std. Deviation
AWL	190	1.1300	4.4800	5.6100	4.935053	.2167723
TTR	190	36	45	81	68.61	5.852
AMP	190	.8600	.0000	.8600	.232053	.1541747
ANDC	190	1.0500	.0700	1.1200	.488474	.2067442
CAUS	190	.5300	.0000	.5300	.118368	.1133799
CONC	190	.3700	.0000	.3700	.047474	.0656585
COND	190	1.0200	.0000	1.0200	.330474	.1804060
CONJ	190	1.2100	.0700	1.2800	.526474	.2310333
DEMO	190	1.7200	.3700	2.0900	1.088579	.3565763
DEMP	190	1.4200	.0000	1.4200	.444158	.2393511
DPAR	190	.2400	.0000	.2400	.021105	.0421429
DWNT	190	.7200	.0000	.7200	.233474	.1537005
EMPH	190	1.6500	.0600	1.7100	.571105	.2985509
EX	190	.9700	.0000	.9700	.227842	.1661391
FPP1	190	4.9600	.1900	5.1500	2.454000	.9580479
GER	190	3.2300	.0600	3.2900	.685316	.5368768
HDG	190	.0700	.0000	.0700	.002263	.0116216
INPR	190	.2400	.0000	.2400	.028000	.0485580
JJ	190	8.9600	3.4200	12.3800	6.388421	1.4447990
NEMD	190	1.6200	.0000	1.6200	.389211	.2440897
NN	190	11.3200	18.7900	30.1100	23.655895	2.1623199
NOMZ	190	7.7600	1.6100	9.3700	4.426895	1.2500791
OSUB	190	.7200	.0000	.7200	.217000	.1378518
PHC	190	2.3000	.1800	2.4800	.890053	.4484229
PIN	190	7.3700	6.9900	14.3600	10.413579	1.2808243
PIT	190	2.1700	.0700	2.2400	.787947	.3995891
PLACE	190	.6600	.0000	.6600	.179263	.1259608
POMD	190	1.8100	.0600	1.8700	.759263	.3374123
PRED	190	1.1700	.1200	1.2900	.624053	.2517198
PRMD	190	2.1100	.1900	2.3000	1.053105	.4321122
RB	190	5.2500	1.0700	6.3200	3.188579	.7687822
SPP2	190	2.4300	.0000	2.4300	.381947	.4825537
SYNE	190	.7100	.0000	.7100	.186368	.1490001
THAC	190	.2500	.0000	.2500	.048684	.0568192
THVC	190	1.2900	.0000	1.2900	.524737	.2624784
TIME	190	1.5100	.0000	1.5100	.601421	.3059871
TO	190	2.4300	.8000	3.2300	1.970105	.4565375
TOBJ	190	.5500	.0000	.5500	.137895	.1146321
TPP3	190	6.0400	.1300	6.1700	1.485895	.8350737
TSUB	190	.3300	.0000	.3300	.049632	.0656943
VBD	190	4.7300	.2500	4.9800	1.407947	.7361637

VPRT	190	5.5000	2.0900	7.5900	5.081947	1.0664558
XX0	190	2.2800	.0000	2.2800	.848211	.3772849
[BEMA]	190	2.2600	.5100	2.7700	1.352947	.4327465
[BYPA]	190	.5000	.0000	.5000	.132368	.1122222
[CONT]	190	.1900	.0000	.1900	.009474	.0289816
[PASS]	190	2.3000	.2500	2.5500	1.067000	.4064232
[PASTP]	190	.3000	.0000	.3000	.024579	.0453275
[PEAS]	190	2.0100	.2500	2.2600	.988368	.3455951
[PIRE]	190	.3000	.0000	.3000	.040737	.0653479
[PRESP]	190	.6800	.0000	.6800	.227105	.1462134
[PRIV]	190	2.4200	.3800	2.8000	1.256474	.3992118
[PROD]	190	.4700	.0000	.4700	.081053	.0893714
[PUBV]	190	2.1600	.0600	2.2200	.729579	.3573445
[SERE]	190	.3800	.0000	.3800	.098579	.0828627
[SMP]	190	.2000	.0000	.2000	.018632	.0380089
[SPAU]	190	1.3200	.1900	1.5100	.708368	.2736680
[SPIN]	190	.3400	.0000	.3400	.032211	.0544074
[STPR]	190	.3200	.0000	.3200	.062053	.0702941
[SUAV]	190	1.9800	.1200	2.1000	.624053	.3273105
[THATD]	190	.7200	.0000	.7200	.186789	.1275770
[WHCL]	190	.3200	.0000	.3200	.057579	.0665430
[WHOBJ]	190	.3600	.0000	.3600	.028474	.0504717
[WHQU]	190	.7100	.0000	.7100	.085316	.1163495
[WHSUB]	190	.5900	.0000	.5900	.167947	.1376465
[WZPAST]	190	1.0600	.0000	1.0600	.323263	.2090464
[WZPRES]	190	.8400	.0000	.8400	.278105	.1581794
CC	190	2.5300	.7900	3.3200	1.906211	.4950179
DT	190	7.3100	5.8200	13.1300	9.136895	1.3286793
IN	190	2.2200	.6500	2.8700	1.569421	.3740005
POS	190	1.8000	.0000	1.8000	.350842	.2732120
RP	190	.9000	.0000	.9000	.224263	.1566252
WP	190	1.3800	.0000	1.3800	.328632	.2274688
STTR	189	15.3000	31.5000	46.8000	39.278307	2.8525475
ASL	189	26.8706	13.7544	40.6250	23.274905	4.4103728
TOP10	189	10.2700	21.5300	31.8000	25.705608	1.9976572
LD	190	9.1600	52.5600	61.7200	56.938105	1.8875371
Valid N (listwise)	189					

Appendix 5 Mann-Whitney U Test between NS and SI, and between SI and WT

Test Statistics[a] (between NS and SI)

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
AWL	9137.500	20312.500	-1.628	.103
TTR	6443.000	17618.000	-5.473	.000
AMP	2841.500	14016.500	-10.594	.000
ANDC	9770.000	20945.000	-.727	.467
CAUS	8771.500	18362.500	-2.154	.031
CONC	9472.500	20647.500	-1.465	.143
COND	5267.500	14858.500	-7.140	.000
CONJ	6743.500	16334.500	-5.047	.000
DEMO	4316.000	15491.000	-8.492	.000
DEMP	9190.000	20365.000	-1.553	.120
DPAR	7167.500	16758.500	-4.436	.000
DWNT	9393.500	20568.500	-1.270	.204
EMPH	9194.500	20369.500	-1.547	.122
EX	4691.500	14282.500	-7.959	.000
FPP1	4940.500	16115.500	-7.602	.000
GER	7546.000	17137.000	-3.894	.000
HDG	8457.000	18048.000	-3.185	.001
INPR	10190.500	21365.500	-.159	.874
JJ	5406.000	16581.000	-6.940	.000
NEMD	6666.000	16257.000	-5.149	.000
NN	9364.500	18955.500	-1.305	.192
NOMZ	6356.000	15947.000	-5.587	.000
OSUB	8081.000	17672.000	-3.139	.002
PHC	9177.000	20352.000	-1.572	.116
PIN	5721.000	16896.000	-6.491	.000
PIT	8808.000	19983.000	-2.097	.036
PLACE	6110.500	17285.500	-5.945	.000
POMD	8098.000	17689.000	-3.108	.002
PRED	7262.500	18437.500	-4.297	.000
PRMD	9906.500	21081.500	-.533	.594
RB	8883.000	20058.000	-1.990	.047
SPP2	5583.000	15174.000	-6.688	.000
SYNE	8122.500	17713.500	-3.097	.002
THAC	4081.500	15256.500	-9.023	.000
THVC	7856.000	17447.000	-3.453	.001
TIME	7469.500	18644.500	-4.003	.000
TO	9169.500	20344.500	-1.582	.114

TOBJ	2023.000	13198.000	-11.784	.000
TPP3	6395.500	17570.500	-5.531	.000
TSUB	3011.500	14186.500	-10.483	.000
VBD	9284.500	18875.500	-1.419	.156
VPRT	9033.000	20208.000	-1.777	.076
XX0	4622.500	14213.500	-8.055	.000
[BEMA]	8026.500	19201.500	-3.209	.001
[BYPA]	8660.000	18251.000	-2.347	.019
[CONT]	9052.000	20227.000	-1.750	.080
[PASS]	7771.000	17362.000	-3.573	.000
[PASTP]	9122.500	20297.500	-2.354	.019
[PEAS]	9104.000	18695.000	-1.676	.094
[PIRE]	7116.500	18291.500	-5.118	.000
[PRESP]	10177.500	21352.500	-.150	.881
[PRIV]	5791.000	16966.000	-6.392	.000
[PROD]	10023.000	21198.000	-.368	.713
[PUBV]	9228.000	18819.000	-1.499	.134
[SERE]	2441.000	13616.000	-11.620	.000
[SMP]	9130.000	20305.000	-2.068	.039
[SPAU]	9808.500	19399.500	-.673	.501
[SPIN]	9635.500	19226.500	-1.242	.214
[STPR]	9326.000	20501.000	-1.375	.169
[SUAV]	9199.000	20374.000	-1.540	.123
[THATD]	7020.000	18195.000	-4.643	.000
[WHCL]	8981.000	18572.000	-1.878	.060
[WHOBJ]	6942.000	18117.000	-5.976	.000
[WHQU]	8536.000	18127.000	-2.609	.009
[WHSUB]	6569.000	17744.000	-5.298	.000
[WZPAST]	8953.500	18544.500	-1.907	.057
[WZPRES]	8782.000	19957.000	-2.135	.033
CC	7739.500	17330.500	-3.618	.000
DT	8687.500	18278.500	-2.268	.023
IN	7120.000	16711.000	-4.500	.000
POS	8982.500	20157.500	-1.850	.064
RP	5314.500	14905.500	-7.082	.000
WP	4621.500	15796.500	-8.057	.000
STTR	4237.500	15412.500	-8.604	.000
ASL	3983.000	15158.000	-8.965	.000
TOP10	8659.500	18250.500	-2.308	.021
LD	8741.000	18332.000	-2.192	.028

a. Grouping Variable: New_variety

Test Statistics (between SI and WT)

	Mann-Whitney U	Wilcoxon W	Z	Asymp. Sig. (2-tailed)
AWL	4137.500	15312.500	-11.186	.000
TTR	10887.500	22062.500	-3.655	.000
AMP	13792.500	24967.500	-.405	.685
ANDC	2422.000	20567.000	-13.102	.000
CAUS	9239.000	27384.000	-5.522	.000
CONC	10686.500	21861.500	-4.483	.000
COND	10410.500	28555.500	-4.182	.000
CONJ	3410.000	14585.000	-12.002	.000
DEMO	12653.500	30798.500	-1.677	.094
DEMP	4673.500	22818.500	-10.588	.000
DPAR	2302.000	20447.000	-13.942	.000
DWNT	8677.500	19852.500	-6.134	.000
EMPH	10145.000	28290.000	-4.478	.000
EX	4829.500	22974.500	-10.419	.000
FPP1	6417.500	24562.500	-8.640	.000
GER	13436.000	24611.000	-.803	.422
HDG	9235.500	27380.500	-8.100	.000
INPR	14002.500	25177.500	-.210	.833
JJ	7258.500	18433.500	-7.700	.000
NEMD	14081.500	25256.500	-.082	.935
NN	7798.000	18973.000	-7.098	.000
NOMZ	9064.000	20239.000	-5.684	.000
OSUB	12792.000	23967.000	-1.525	.127
PHC	8590.500	19765.500	-6.213	.000
PIN	6304.000	17479.000	-8.766	.000
PIT	11286.500	29431.500	-3.203	.001
PLACE	13111.000	24286.000	-1.169	.242
POMD	13664.500	31809.500	-.548	.584
PRED	11182.000	29327.000	-3.320	.001
PRMD	10626.500	28771.500	-3.940	.000
RB	12583.000	23758.000	-1.755	.079
SPP2	5589.500	23734.500	-9.579	.000
SYNE	11973.500	23148.500	-2.446	.014
THAC	13613.500	24788.500	-.656	.512
THVC	13389.500	31534.500	-.855	.393
TIME	7846.500	19021.500	-7.045	.000
TO	10131.500	28276.500	-4.493	.000
TOBJ	11733.500	22908.500	-2.730	.006
TPP3	12047.500	30192.500	-2.353	.019

TSUB	12997.500	31142.500	-1.382	.167
VBD	12678.500	30823.500	-1.649	.099
VPRT	4997.000	23142.000	-10.226	.000
XX0	8454.000	26599.000	-6.366	.000
[BEMA]	7870.500	26015.500	-7.017	.000
[BYPA]	12290.500	23465.500	-2.103	.036
[CONT]	44.000	18189.000	-16.813	.000
[PASS]	13694.500	24869.500	-.514	.607
[PASTP]	12091.000	23266.000	-3.145	.002
[PEAS]	8826.500	20001.500	-5.950	.000
[PIRE]	12117.500	23292.500	-2.758	.006
[PRESP]	6649.000	17824.000	-8.416	.000
[PRIV]	14111.500	25286.500	-.049	.961
[PROD]	6875.500	25020.500	-8.216	.000
[PUBV]	11258.500	29403.500	-3.234	.001
[SERE]	5909.000	17084.000	-9.662	.000
[SMP]	14133.500	25308.500	-.033	.974
[SPAU]	5464.500	16639.500	-9.705	.000
[SPIN]	12815.500	23990.500	-1.831	.067
[STPR]	10820.500	28965.500	-3.845	.000
[SUAV]	12959.500	24134.500	-1.335	.182
[THATD]	8042.500	26187.500	-6.835	.000
[WHCL]	9801.500	27946.500	-5.005	.000
[WHOBJ]	11509.500	22684.500	-4.011	.000
[WHQU]	13405.000	31550.000	-.873	.383
[WHSUB]	12204.000	23379.000	-2.189	.029
[WZPAST]	5965.500	17140.500	-9.160	.000
[WZPRES]	11876.000	23051.000	-2.547	.011
CC	13402.500	31547.500	-.840	.401
DT	9990.500	21165.500	-4.650	.000
IN	13925.000	25100.000	-.257	.797
POS	11907.500	23082.500	-2.511	.012
RP	10899.000	29044.000	-3.639	.000
WP	12217.000	30362.000	-2.165	.030
STTR	6674.500	17849.500	-8.304	.000
ASL	3037.000	14212.000	-12.382	.000
TOP10	11683.000	29638.000	-2.688	.007
LD	6824.000	17999.000	-8.186	.000

a. Grouping Variable: New_Variety

Appendix 6 Total variance explained based on Principal Axis Factoring

Total Variance Explained

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	15.374	19.461	19.461	15.063	19.067	19.067
2	5.310	6.721	26.182	4.836	6.121	25.188
3	4.001	5.064	31.246	3.563	4.510	29.698
4	2.900	3.671	34.916	2.546	3.222	32.920
5	2.513	3.181	38.097	1.990	2.518	35.439
6	2.051	2.597	40.694	1.572	1.990	37.429
7	1.720	2.177	42.871	1.253	1.586	39.015
8	1.648	2.085	44.957	1.188	1.504	40.519
9	1.516	1.919	46.876	1.047	1.326	41.845
10	1.482	1.876	48.751	.958	1.213	43.058
11	1.390	1.760	50.511	.896	1.134	44.192
12	1.370	1.734	52.245	.810	1.026	45.218
13	1.301	1.647	53.892	.762	.965	46.182
14	1.230	1.557	55.449	.740	.936	47.118
15	1.218	1.541	56.990	.681	.862	47.980
16	1.159	1.468	58.458	.653	.827	48.807
17	1.138	1.441	59.899	.614	.777	49.584
18	1.103	1.396	61.294	.553	.701	50.284
19	1.070	1.354	62.648	.532	.674	50.958
20	1.053	1.333	63.982	.470	.595	51.554
21	.989	1.252	65.233			
22	.970	1.228	66.461			
23	.938	1.187	67.648			
24	.928	1.175	68.823			
25	.909	1.150	69.973			
26	.889	1.126	71.099			
27	.865	1.095	72.193			
28	.858	1.086	73.279			
29	.812	1.028	74.307			
30	.803	1.017	75.324			
31	.785	.993	76.317			
32	.755	.955	77.273			
33	.738	.934	78.207			
34	.707	.895	79.102			
35	.690	.874	79.975			
36	.680	.861	80.836			
37	.665	.841	81.678			

38	.652	.825	82.502			
39	.641	.811	83.314			
40	.615	.778	84.091			
41	.605	.766	84.858			
42	.594	.752	85.610			
43	.580	.734	86.344			
44	.575	.728	87.071			
45	.572	.724	87.795			
46	.532	.674	88.469			
47	.527	.668	89.137			
48	.503	.636	89.773			
49	.488	.618	90.391			
50	.472	.597	90.988			
51	.468	.592	91.581			
52	.448	.567	92.148			
53	.427	.540	92.688			
54	.417	.528	93.216			
55	.391	.494	93.710			
56	.388	.491	94.201			
57	.373	.473	94.674			
58	.354	.449	95.123			
59	.343	.434	95.557			
60	.321	.406	95.963			
61	.311	.393	96.356			
62	.304	.385	96.741			
63	.273	.346	97.087			
64	.260	.329	97.416			
65	.254	.322	97.738			
66	.242	.306	98.044			
67	.226	.286	98.330			
68	.218	.276	98.606			
69	.195	.247	98.853			
70	.167	.212	99.065			
71	.139	.176	99.241			
72	.133	.169	99.410			
73	.121	.153	99.564			
74	.107	.135	99.699			
75	.097	.122	99.821			
76	.054	.068	99.889			
77	.044	.056	99.945			
78	.026	.033	99.978			
79	.017	.022	100.000			

Extraction Method: Principal Axis Factoring.

Appendix 7 Rotated factorial structure based on Varimax

Factor 1	
present tense	.792
contractions	.787
shorter words	.731
Be as main verb	.697
discourse particles	. 684
demonstrative pronoun	. 649
first person pronoun	.645
second person pronoun	.607
pro-verb *do*	.571
independent clause coordination	.543
analytic negation	.537
subordinator that deletion	.513
pronoun it	.500
emphatics	.479
existential there	.429
causative adverbial subordinators	.426
wh-pronouns	.417
hedges	.408
predicative adjectives	.389
wh-clauses	.383
public verbs	.379
(demonstratives	.324)
(total adverbs	.304)
……	
average word length	-.897
……	
conditional adverbial subordinators	-.437
analytic negation	-.436
(second person pronouns	-.322)
(existential 'there'	-.310)
Factor 3	
split auxiliaries	.590
conjunct	.489
total adverbs	.485
downtoners	.462
time adverbials	.419
concessive adverbial subordinators	.359
perfect aspect	.353
(synthetic negation	.339)
(that verb complements	.309)
……	
contractions	-.408
(top 10 coverage	-.321)
(independent clause coordination	-.315)
Factor 4	
longer words	-.831
average sentence length	-.706
total prepositional phrases	-.703
total other nouns	-.623
attributive adjectives	-.608
past participial WHIZ deletion relatives	-.581
phrasal coordination	-.571
nominalization	-.539
present participial clauses	-.529
conjuncts	-.456
lexical density	-.434
determiner 'the'	-.402
(split auxiliaries	-.317)
Factor 2	
that relative clauses on object position	.647
amplifiers	.613
that relative clauses on subject position	.574
first person pronoun	.524
that adjective complements	.486
demonstratives	.439
private verbs	.431
sentence relatives	.431
standardized type-token ratio	.375
wh-pronouns	.355
(place adverbials	.323)
nominalization	-.466
gerunds	-.408
(lexical density	-.319)
(preposition or subordinating conjunction	-.313)
(longer words	-.302)
Factor 5	
lexical density	.460
…….	
top 10 coverage	-.776
determiner 'the'	-.669
shorter words	-.485
Factor 6	
suasive verbs	.588
public verbs	.434
(predictive modals	.341)
(that verb complements	.341)
…….	
standardized type-token ratio	-.403

third person pronoun	.533	**Factor 7**	
wh-pronoun	.477	coordinating conjunction	.352
wh relative clauses on subject position	.377	…….	
be as main verb	.372	total other nouns	-.507
……..		possessive endings	-.361
(lexical density	-.340)		
Factor 8			
(pied-piping relative clauses	.326)		
(average sentence length	.317)		
……			
particles	-.371		

References

Adolphs, S., Atkins, S., & Harvey, K. (2007). Caught between professional requirements and interpersonal needs: Vague language in healthcare contexts. In J. Cutting (Ed.), *Vague language explored* (pp. 62–78). Palgrave Macmillan. https://doi.org/10.1057/9780230627420_4

Adolphs, S., & Knight, D. (2010). Building a spoken corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics*. Routledge. http://orca.cf.ac.uk/78849/

Aijmer, K. (2002). English discourse particles. In *Scl.10* (Vol. 10). John Benjamins Publishing Company. https://benjamins.com/catalog/scl.10

Akinnaso, F. N. (1982). On the differences between spoken and written language: *Language and Speech*, *25*(2), 97–125. https://doi.org/10.1177/002383098202500201

Al-Salman, S., & Al-Khanji, R. (2002). The native language factor in simultaneous interpretation in an Arabic/English context. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *47*(4), 607–626. https://doi.org/10.7202/008040ar

Al-Surmi, M. (2012). Authenticity and TV shows: A multidimensional analysis perspective. *TESOL Quarterly*, *46*(4), 671–694. https://doi.org/10.1002/tesq.33

Altenberg, B. (1986). *Contrastive linking in spoken and written English*.

Altenberg, B. (1998). On the phraseology of spoken English: The evidence of recurrent word-combinations. In A. P. Cowie & MyiLibrary (Eds.), *Phraseology: Theory, analysis, and applications: Vol. Oxford studies in lexicography and lexicology* (pp.

101–122). Oxford University Press. http://www.myilibrary.com?id=81483

Altenberg, B., & Granger, S. (2001). The grammatical and lexical patterning of MAKE in native and non-native student writing. *Applied Linguistics*, *22*(2), 173–195. https://doi.org/10.1093/applin/22.2.173

Anthony, L. (2015). *TagAnt* (1.1.0) [Computer software]. Waseda University. https://www.laurenceanthony.net/software/tagant/

Avner, E. A., Ordan, N., & Wintner, S. (2016). Identifying translationese at the word and sub-word level. *Digital Scholarship in the Humanities*, *31*(1), 30–54. https://doi.org/10.1093/llc/fqu047

Baker, M. (1993). Corpus linguistics and translation studies—Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins Publishing Company. https://benjamins.com/catalog/z.64.15bak

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target. International Journal of Translation Studies*, *7*(2), 223–243. https://doi.org/10.1075/target.7.2.03bak

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and translation: Studies in language engineering in honour of Juan C. Sager* (Vol. 18, pp. 175–186). John Benjamins Publishing Company. https://benjamins.com/catalog/btl.18.17bak

Baker, M. (1999). The role of corpora in investigating the linguistic behaviour of professional translators. *International Journal of Corpus Linguistics*, *4*(2), 281–298.

https://doi.org/10.1075/ijcl.4.2.05bak

Baker, M. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics*, *9*(2), 167–193. https://doi.org/10.1075/ijcl.9.2.02bak

Baker, M. (2007). Patterns of idiomaticity in translated vs. Non-translated text. *Belgian Journal of Linguistics*, *21*(1), 11–21. https://doi.org/10.1075/bjl.21.02bak

Bamford, J., Cavalieri, S., & Diani, G. (2013). *Variation and change in spoken and written discourse: Perspectives from corpus linguistics*. John Benjamins Publishing Company.

Barbieri, F. (2008). Patterns of age-based linguistic variation in American English1. *Journal of Sociolinguistics*, *12*(1), 58–88. https://doi.org/10.1111/j.1467-9841.2008.00353.x

Baroni, M., & Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), 259–274. https://doi.org/10.1093/llc/fqi039

Bartłomiejczyk, M. (2004). Simultaneous interpreting A-B vs. B-A from the interpreters' standpoint. In G. Hansen, K. Malmkjær, & D. Gile (Eds.), *Claims, changes and challenges in translation studies: Selected contributions from the EST congress* (Vol. 50, pp. 239–249). John Benjamins Publishing Company. https://doi.org/10.1075/btl.50.20bar

Bartłomiejczyk, M. (2006). Strategies of simultaneous interpreting and directionality. *Interpreting*, *8*, 149–174. https://doi.org/10.1075/intp.8.2.03bar

Baumgarten, N., Meyer, B., & Özçetin, D. (2008). Explicitness in translation and interpreting: A critical review and some empirical evidence (of an elusive concept). *Across Languages and Cultures*, *9*(2), 177–203. https://doi.org/10.1556/Acr.9.2008.2.2

Becher, V. (2010). Abandoning the notion of "translation-inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures*, *11*(1), 1–28. https://doi.org/10.1556/ACR.11.2010.1.1

Becher, V. (2011). *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts* [Doctoral dissertation, Universität Hamburg]. https://ediss.sub.uni-hamburg.de/volltexte/2011/5321/pdf/Dissertation.pdf

Bendazzoli, C., & Sandrelli, A. (2005). An approach to corpus-based interpreting studies: Developing EPIC (European Parliament Interpreting Corpus). In S. Bazzanella, S. Buhl, L. Jiang, & K. Mysak (Eds.), *Challenges of multidimensional translation* (pp. 1–36). St. Jerome Publishing.

Bendazzoli, C., & Sandrelli, A. (2009). Corpus-based interpreting studies: Early work and future prospects. *Revista Tradumàtica*, *7*, 1–9.

Bernardini, S. (2015). Translation. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 515–536). Cambridge University Press.

Bernardini, S., & Baroni, M. (2005). Spotting translationese: A corpus-driven approach using support vector machines. *Proceedings of Corpus Linguistics Conference Series 2005*, *1*, 1–12.

Bernardini, S., Collard, C., Defrancq, B., Ferraresi, A., & Russo, M. (2018). Building Interpreting and Intermodal Corpora: A How-to for a Formidable Task. In C. Bendazzoli, M. Russo, & B. Defrancq (Eds.), *Making Way in Corpus-based Interpreting Studies* (pp. 21–42). Springer. https://doi.org/10.1007/978-981-10-6199-8_2

Bernardini, S., & Ferraresi, A. (2011). Practice, description and theory come together – Normalization or interference in Italian technical translation? *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *56*(2), 226–246. https://doi.org/10.7202/1006174ar

Bernardini, S., Ferraresi, A., & Miličević, M. (2016). From EPIC to EPTIC — Exploring simplification in interpreting and translation from an intermodal perspective. *Target. International Journal of Translation Studies*, *28*(1), 61–86. https://doi.org/10.1075/target.28.1.03ber

Bernardini, S., Ferraresi, A., Russo, M., Collard, C., & Defrancq, B. (2018). Building interpreting and Intermodal corpora: A how-to for a formidable task. In C. Bendazzoli, M. Russo, & B. Defrancq (Eds.), *Making way in corpus-based interpreting studies* (pp. 21–42). Springer. https://doi.org/10.1007/978-981-10-6199-8_2

Bernardini, S., & Zanettin, F. (2004). When is a universal not a universal? Some limits of current corpus-based methodologies for the investigation of translation universals. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (Vol. 48, pp. 51–62). John Benjamins Publishing Company.

https://benjamins.com/catalog/btl.48.05ber

Bernstein, B. (1964). Elaborated and restricted codes: Their social origins and some consequences. *American Anthropologist*, *66*(6_PART2), 55–69. https://doi.org/10.1525/aa.1964.66.suppl_3.02a00030

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language*, *62*(2), 384–414. https://doi.org/10.2307/414678

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (1992a). Using computer-based text corpora to analyze the referential strategies of spoken and written texts. In J. Svartvik (Ed.), *Directions in corpus linguistics proceedings of Nobel Symposium 82 Stockholm, 4-8 August 1991* (pp. 213–252). De Gruyter Mouton.

Biber, D. (1992b). On the complexity of discourse complexity: A multidimensional analysis. *Discourse Processes*, *15*(2), 133–163. https://doi.org/10.1080/01638539209544806

Biber, D. (1992c). The multi-dimensional approach to linguistic analyses of genre variation: An overview of methodology and findings. *Computers and the Humanities*, *26*(5–6), 331–345. https://doi.org/10.1007/BF00136979

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge University Press. https://doi.org/10.1017/CBO9780511519871

Biber, D. (1999). A register perspective on grammar and discourse: Variability in the form and use of English complement clauses. *Discourse Studies*, *1*(2), 131–150. https://doi.org/10.1177/1461445699001002001

Biber, D. (2006a). *University language: A corpus-based study of spoken and written registers*. John Benjamins Publishing Company.

Biber, D. (2006b). Stance in spoken and written university registers. *Journal of English for Academic Purposes*, *5*(2), 97–116. https://doi.org/10.1016/j.jeap.2006.05.001

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics*, *14*(3), 275–311.

Biber, D. (2010). Corpus-based and corpus-driven analyses of language variation and use. In B. Heine & N. Heiko (Eds.), *The Oxford handbook of linguistic analysis* (pp. 159–191).

Biber, D. (2014). Using multi-dimensional analysis to explore cross-linguistic universals of register variation. *Languages in Contrast*, *14*(1), 7–34. https://doi.org/10.1075/lic.14.1.02bib

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, D., & Hared, M. (1992). Dimensions of register variation in Somali. *Language Variation and Change*, *4*(1), 41–75. https://doi.org/10.1017/S095439450000065X

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Longman.

Biber, D., & Kim, Y. (1994). A corpus-based analysis of register variation in Korean. In E. Finegan & D. Biber (Eds.), *Sociolinguistic perspectives on register* (pp. 157–181).

Oxford University Press.

Biber, D., & Reppen, R. (2015). *The Cambridge handbook of English corpus linguistics*. Cambridge University Press. https://doi.org/10.1017/CBO9781139764377

Blankenship, J. (1962). A linguistic analysis of oral and written style. *Quarterly Journal of Speech*, *48*(4), 419–422. https://doi.org/10.1080/00335636209382571

Blankenship, J. (1974). The influence of mode, sub-mode, and speaker predilection on style. *Communications Monographs*, *41*(2), 85–118. https://doi.org/10.1080/03637757409375826

Blum-Kulka, S. (1986). Jyvaskyla: Jyvaskyla Cross-Language Studies no. II. *Department of English, University of Jyvaskyla*, *8*(2), 237–238.

Blum-Kulka, S., & Levenston, S. (1983). Universals of lexical simplification. In G. Kasper & C. Faerch (Eds.), *Strategies in interlanguage communication* (pp. 119–139). Longman.

Bolinger, D. (1975). *Aspects of language* (2nd ed.). Harcourt Brace Jovanovich, Inc.

Bowker, J. (2013). Variation across spoken and written registers in internal corporate communication. In J. Bamford, S. Cavalieri, & G. Diani (Eds.), *Variation and change in spoken and written discourse: Perspectives from corpus linguistics* (Vol. 21, pp. 47–64). John Benjamins Publishing Company. https://benjamins.com/catalog/ds.21.08bow

Bros-Brann, E. (1976). Critical comments on H.C. Barik's article "Interpreters Talk a Lot, among other things." *AIIC Bulletin*, *4*(1), 16–18.

Burnard, L. (2005). Metadata for corpus work. In M. Wynne (Ed.), *Developing linguistic*

corpora: A guide to good practice (pp. 30–46).

Cappelle, B. (2012). English is less rich in manner-of-motion verbs when translated from French. *Across Languages and Cultures*, *13*(2), 173–195. https://doi.org/10.1556/Acr.13.2012.2.3

Carter, R., & Sánchez-Macarro, A. (1998). *Linguistic choice across genres: Variation in spoken and written English* (Vol. 158). John Benjamins Publishing Company.

Cattell, R. B. (1966). The scree test for the number of factors. *Multivariate Behavioral Research*, *1*(2), 245–276. https://doi.org/10.1207/s15327906mbr0102_10

Cencini, M. (2002). On the importance of an encoding standard for corpus-based interpreting studies extending the TEI scheme. *InTRAlinea*, *Special issue: CULT2K*. http://www.intralinea.org/specials/article/1678

Chafe, W. (1982). Integration and involvement in speaking, writing, and oral literature. *Spoken and Written Language: Exploring Orality and Literacy*, 35–54.

Chafe, W., & Danielewicz, J. (1986). Properties of spoken and written language. In *Comprehending oral and written language* (pp. 83–113). Academic Press.

Chafe, W., & Tannen, D. (1987). The relation between written and spoken language. *Annual Reviews*, *16*(1), 383–407. https://doi.org/10.1146/annurev.an.16.100187.002123

Chang, C., & Schallert, D. L. (2007). The impact of directionality on Chinese/English simultaneous interpreting. *Interpreting*, *9*(2), 137–176. https://doi.org/10.1075/intp.9.2.02cha

Cheng, W. (2007). The use of vague language across spoken genres in an intercultural

Hong Kong corpus. In J. Cutting (Ed.), *Vague language explored* (pp. 161–181). Palgrave MacMillan. https://doi.org/10.1057/9780230627420_9

Cheng, W., Greaves, C., & Warren, M. (2005). The creation of a prosodically transcribed intercultural corpus: The Hong Kong corpus of spoken English (prosodic). *International Computer Archive of Modern English*, *29*, 47–68.

Chernov, G. V. (1994). Message redundancy and message anticipation in simultaneous interpreting. In B. Moser-Mercer & S. Lambert (Eds.), *Bridging the gap: Empirical research in simultaneous interpretation* (Vol. 3, pp. 139-). John Benjamins Publishing Company. https://benjamins.com/catalog/btl.3.13che

Chesterman, A. (2004). Hypotheses about translation universals. In G. Hansen, K. Malmkjær, & D. Gile (Eds.), *Claims, changes and challenges in translation studies: Selected contributions from the EST congress, Copenhagen 2001* (Vol. 50, pp. 1–13). John Benjamins Publishing Company. https://benjamins.com/catalog/btl.50.02che

Chesterman, A. (2011). Translation universals. In Y. Gambier & L. van Doorslaer (Eds.), *Handbook of translation studies* (Vol. 2, pp. 175–179). John Benjamins Publishing Company. https://benjamins.com/catalog/hts.2.tra12

Chesterman, A. (2015). Models of what processes. In M. Ehrensberger-Dow, B. E. Dimitrova, S. Hubscher-Davidson, & U. Norberg (Eds.), *Describing cognitive processes in translation: Acts and events* (Vol. 77, pp. 7–20). John Benjamins Publishing Company.

Chesterman, A. (2017). *Reflections on translation theory: Selected papers 1993 - 2014*. John Benjamins Publishing Company.

Cook, V. J. (2014). *The English writing system*. Routledge.

Corpas Pastor, G. (2008). *Investigar con corpus en traducción: Los retos de un nuevo paradigma*. https://www.peterlang.com/view/title/52112

Dai, G., & Xiao, R. (2011). "SL shining through" in translational language: A corpus-based study of Chinese translation of English passives. *Translation Quarterly*, *62*, 85–108.

Dayter, D. (2018). Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). *FORUM*, *16*(2), 241–264. https://doi.org/10.1075/forum.17004.day

De Sutter, G., & Lefer, M.-A. (2020). On the need for a new research agenda for corpus-based translation studies: A multi-methodological, multifactorial and interdisciplinary approach. *Perspectives*, *28*(1), 1–23. https://doi.org/10.1080/0907676X.2019.1611891

De Sutter, G., Lefer, M.-A., & Delaere, I. (2017). Empirical translation studies: New methodological and theoretical traditions. In *Empirical Translation Studies* (Vol. 300). De Gruyter Mouton. https://www.degruyter.com/view/title/517065

De Sutter, G., & Vermeire, E. (2020). Grammatical optionality in translations: A multifactorial corpus analysis of that/zero alternation in English using the MuPDAR approach. In L. Vandevoorde, J. Daems, & B. Defrancq (Eds.), *New empirical perspectives on translation and interpreting* (pp. 24–51). Routledge.

Defrancq, B. (2018). The European parliament as a discourse community: Its role in comparable analyses of data drawn from parallel interpreting corpora. *EUT Edizioni*

Università Di Trieste, *The Interpreters' Newsletter n. 23*, 115–132. https://doi.org/10.13137/2421-714X/22401

Defrancq, B., Plevoets, K., & Magnifico, C. (2015). Connective items in interpreting and translation: Where do they come from? In J. Romero-Trillo (Ed.), *Yearbook of corpus linguistics and pragmatics 2015* (Vol. 3, pp. 195–222). Springer International Publishing. https://doi.org/10.1007/978-3-319-17948-3_9

Déjean Le Féal, K. (2005). Can and should interpretation Into a second language be taught? *Communication and Cognition. Monographies*, *38*(1–2), 167–194.

Delaere, I. (2015). *Do translations walk the line?: Visually exploring translated and non-translated texts in search of norm conformity* [Doctoral dissertation, Ghent University]. http://hdl.handle.net/1854/LU-5888594

Delaere, I., & De Sutter, G. (2013). Applying a multidimensional, register-sensitive approach to visualize normalization in translated and non-translated Dutch. *Belgian Journal of Linguistics*, *27*(1), 43–60. https://doi.org/10.1075/bjl.27.03del

Delaere, I., De Sutter, G., & Plevoets, K. (2012). Is translated language more standardized than non-translated language?: Using profile-based correspondence analysis for measuring linguistic distances between language varieties. *Target. International Journal of Translation Studies*, *24*(2), 203–224. https://doi.org/10.1075/target.24.2.01del

Denissenko, J. (1989). Communicative and interpretative linguistics. In J. Dodds & L. Gran (Eds.), *The theoretical and practical aspects of teaching conference interpretation: First International Symposium on Conference Interpreting at the*

University of Trieste (pp. 155–158). Campanotto Editore.

Deshors, S. C., & Gries, S. Th. (2014). Using regressions to explore deviations between corpus data and a standard/target: Two suggestions. *Corpora*, *9*(1), 109–136. https://doi.org/10.3366/cor.2014.0053

DeVito, J. A. (1966). Psychogrammatical factors in oral and written discourse by skilled communicators. *Speech Monographs*, *33*(1), 73–76. https://doi.org/10.1080/03637756609375483

DeVito, J. A. (1967). A linguistic analysis of spoken and written language. *Central States Speech Journal*, *18*(2), 81–85. https://doi.org/10.1080/10510976709362867

Dimitrova, B. E. (2005). *Expertise and explicitation in the translation process*. John Benjamins Publishing Company.

Donavan, D. T., Mishra, S., Patil, V. H., & Surendra, N. S. (2007). *Parallel analysis engine to aid in determining number of factors to retain using R* (Version 1) [Computer software]. https://analytics.gonzaga.edu/parallelengine/

Donovan, C. (2003). Teaching simultaneous interpretation into B. In D. Kelly, A. Martin, M. Nobs, & C. Sanchez (Eds.), *La direccionalidad en traduccion e interpretacion [Directionality in translation and interpretation]* (pp. 367–380). Editorial Atrio.

Donovan, C. (2004). European masters project Group: Teaching simultaneous interpretation into a B language. *Interpreting*, *6*(2), 205–216. https://doi.org/10.1075/intp.6.2.06don

Donovan, C. (2005). Teaching simultaneous interpretation into B: A challenge for responsible interpreter training. *Communication & Cognition. Monographies*, *38*(1–

2), 147–166.

Drieman, G. H. J. (1962). Differences between written and spoken language: An exploratory study. *Acta Psychologica*, *20*, 36–57. https://doi.org/10.1016/0001-6918(62)90006-9

Duff, A. (1981). *Third language: Recurrent problems of translation into English*. Pergamon Press.

Ellis, N. C., & Simpson-Vlach, R. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics*, *31*, 487–512. https://doi.org/10.1093/APPLIN/AMP058

Elsness, J. (1984). That or zero? A look at the choice of object clause connective in a corpus of American English. *English Studies*, *65*(6), 519–533.

Eskola, S. (2004). Untypical frequencies in translated language: A corpus-based study on a literary corpus of translated and non-translated Finnish. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (Vol. 48, pp. 83–99). John Benjamins Publishing Company. https://doi.org/10.1075/btl.48.08esk

Even-Zohar, I. (1978). The position of translated literature within the literary polysystem. In J. S. Holmes, J. Lambert, & R. van den Broeck (Eds.), *Literature and translation: New perspectives in literary studies* (pp. 117–127). Acco.

Evert, S., & Neumann, S. (2017). The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In G. De Sutter, I. Delaere, & M.-A. Lefer (Eds.), *Empirical translation studies: New theoretical and methodological traditions* (pp. 47–80). De Gruyter Mouton.

https://doi.org/10.1515/9783110459586-003

Evison, J., McCarthy, M., & O'Keeffe, A. (2007). 'Looking out for love and all the rest of It': Vague category markers as shared social space. In J. Cutting (Ed.), *Vague language explored* (pp. 138–157). Palgrave Macmillan UK. https://doi.org/10.1057/9780230627420_8

Fernández, J. (2013). A corpus-based study of vague language use by learners of Spanish in a study abroad context. In C. Kinginger (Ed.), *Social and cultural aspects of language learning in study abroad* (pp. 299–332). John Benjamins Publishing Company. https://doi.org/10.1075/lllt.37.12fer

Ferragne, E. (2013). Automatic suprasegmental parameter extraction in learner corpora. In N. Ballier, A. Díaz-Negrillo, & P. Thompson (Eds.), *Automatic treatment and analysis of learner corpus data* (Vol. 59, pp. 151–168). John Benjamins Publishing Company. https://doi.org/10.1075/scl.59.10fer

Ferraresi, A., Bernardini, S., Petrović, M. M., & Lefer, M.-A. (2018). Simplified or not simplified? The different guises of mediated English at the European parliament. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *63*(3), 717–738. https://doi.org/10.7202/1060170ar

Fidell, L. S., & Tabachnick, B. G. (2007). *Using multivariate statistics* (5th ed.). Allyn and Bacon.

Field, A. P. (2005). *Discovering statistics using SPSS* (2nd ed.). Sage Publications.

Forchini, P. (2011). *Movie language revisited: Evidence from multi-dimensional analysis and corpora* (Vol. 1). Peter Lang.

Forchini, P. (2012). *Movie language revisited*. Peter Lang CH. https://doi.org/10.3726/978-3-0351-0325-0

Frawley, W. (1984). Prolegomenon to a theory of translation. In *Translation: Literary, linguistic and philosophical perspectives* (pp. 159–175). Associated University Press.

Friginal, E. (2009). *The language of outsourced call centers: A corpus-based study of cross-cultural interaction*. John Benjamins Publishing Company.

Gaspari, F., & Bernardini, S. (2010). *Comparing non-native and translated language: Monolingual comparable corpora with a twist*.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (Eds.), *Translation studies in Scandinavia* (pp. 88–95). CWK Gleerup.

Gellerstam, M. (2005). Fingerprints in translation. In M. Rogers & G. Anderman (Eds.), *In and out of English* (pp. 201–213). Multilingual Matters. https://doi.org/10.21832/9781853597893-016

Gibson, J. W., Gruner, C. R., Kibler, R. J., & Kelly, F. J. (1966). A quantitative examination of differences and similarities in written and spoken messages. *Speech Monographs*, *33*(4), 444–451. https://doi.org/10.1080/03637756609375510

Gile, D. (1995). *Basic concepts and models for interpreter and translator training* (Vol. 8). John Benjamins Publishing Company.

Gile, D. (1999). Testing the Effort Models' tightrope hypothesis in simultaneous interpreting-A contribution. *HERMES-Journal of Language and Communication in*

Business, *23*, 153–172.

Gile, D. (2004). Translation research versus interpreting research: Kinship, differences and prospects for partnership. In C. Schäffner (Ed.), *Translation research and interpreting research: Traditions, gaps and synergies* (pp. 10–34). Multilingual Matters Ltd. https://doi.org/10.21832/9781853597350-003

Gile, D. (2005). *Directionality in conference interpreting: A cognitive view*. 9–26.

Gile, D. (2009). *Basic concepts and models for interpreter and translator training: Revised edition* (Vol. 8). John Benjamins Publishing Company.

Gile, D. (1997). Conference interpreting as a cognitive management problem. *Cognitive Processes in Translation and Interpreting*, *3*, 196–214.

Goldman-Eisler, F. (1958). Speech production and the predictability of words in context. *The Quarterly Journal of Experimental Psychology*, *10*, 96–106. https://doi.org/10.1080/17470215808416261

Gorsuch, R. L. (1983). *Factor analysis (2nd edition)*. Lawrence Erlbaum Associates.

Grbic, N. (2015). 'Settings.' In F. Pöchhacker (Ed.), *Routledge encyclopedia of interpreting studies* (pp. 370–371). Routledge.

Gumul, E. (2006). Explicitation in simultaneous interpreting: A strategy or a by-product of language mediation? *Across Languages and Cultures*, *7*(2), 171–190. https://doi.org/10.1556/Acr.7.2006.2.2

Gumul, E. (2007). Explicitation in conference interpreting. *Translation and Meaning. Part 7*.

Gumul, E. (2008). Explicitation in simultaneous interpreting—The quest for optimal

relevance? In E. Wałaszewska, M. Kisielewska-Krysiuk, A. Korzeniowska, & M. Grzegorzewska (Eds.), *Relevant worlds: Current perspectives on language, translation and relevance theory* (pp. 188–205). Cambridge Scholars Publishing.

Gumul, E. (2012). Variability of cohesive patterns. Personal reference markers in simultaneous and consecutive interpreting. *Linguistica Silesiana*, *33*, 147–172.

Gumul, E. (2015). Explicitation. In F. Pöchhacker (Ed.), *Routledge encyclopedia of interpreting studies* (p. 156). Routledge. https://www.researchgate.net/publication/322318145_Explicitation_-_in_The_Routledge_Encyclopedia_of_Interpreting_Studies_2015

Gumul, E. (2017). Explicitation and directionality in simultaneous interpreting. *Linguistica Silesiana*, *38*, 311–329.

Gumul, E. (2020). Explicitation and cognitive load in simultaneous interpreting: Product- and process-oriented analysis of trainee interpreters' outputs. *Interpreting. International Journal of Research and Practice in Interpreting*, 1–31. https://doi.org/10.1075/intp.00051.gum

Gut, U. (2009). *Non-native speech: A corpus-based analysis of phonological and phonetic properties of L2 English and German*. Peter Lang.

Hair, J. F., Tatham, R. L., Anderson, R. E., & Black, W. (1998). *Multivariate data analysis* (5th edition). Prentice Hall.

Halliday, M. A. K., & Matthiessen, C. (2004). *An introduction to functional grammar*.

Halverson, S. L. (2003). The cognitive basis of translation universals. *Target. International Journal of Translation Studies*, *15*(2), 197–241.

https://doi.org/10.1075/target.15.2.02hal

Halverson, S. L. (2007). A cognitive linguistic approach to translation shifts. *Belgian Journal of Linguistics*, *21*(1), 105–121. https://doi.org/10.1075/bjl.21.08hal

Halverson, S. L. (2009). Elements of doctoral training. *The Interpreter and Translator Trainer*, *3*(1), 79–106. https://doi.org/10.1080/1750399X.2009.10798782

Halverson, S. L. (2010). Cognitive translation studies: Developments in theory and method. In E. Angelone & G. M. Shreve (Eds.), *Translation and cognition: Vol. XV* (pp. 349–369). John Benjamins Publishing Company. https://doi.org/10.1075/ata.xv.18hal

Halverson, S. L. (2017). Gravitational pull in translation. Testing a revised model. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies* (pp. 9–46). De Gruyter Mouton. https://doi.org/10.1515/9783110459586-002

Hansen-Schirra, S. (2011). Between normalization and shining-through: Specific properties of English-German translations and their influence on the target language. In S. Kranich, V. Becher, S. Höder, & J. House (Eds.), *Multilingual discourse production: Diachronic and synchronic perspectives* (Vol. 12, pp. 133–162). John Benjamins Publishing Company. https://benjamins.com/catalog/hsm.12.07han

Hansen-Schirra, S., Neumann, S., & Steiner, E. (2007). Cohesive explicitness and explicitation in an English-German translation corpus. *Languages in Contrast*, *7*(2), 241–265. https://doi.org/10.1075/lic.7.2.09han

Hansen-schirra, S., & Teich, E. (2001). Multi-layer analysis of translation corpora: Methodological issues and practical implications. *In Proceedings of EUROLAN*

2001 Workshop on Multi-Layer Corpus-Based Analysis, 44–55.

Hareide, L. (2017a). The translation of formal source-language lacunas: An empirical study of the over-representation of target-language specific features and the unique items hypotheses. In M. Ji, M. Oakes, D. Li, & L. Hareide (Eds.), *Corpus methodologies explained*. Routledge.

Hareide, L. (2017b). Is there gravitational pull in translation? A corpus-based test of the gravitational pull hypothesis on the language pairs Norweigian-Spanish and English-Spanish. In M. Ji, M. Oakes, D. Li, & L. Hareide (Eds.), *Corpus methodologies explained*. Routledge.

He, H., Boyd-Graber, J., & Daumé III, H. (2016). Interpretese vs. translationese: The uniqueness of human strategies in simultaneous interpretation. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 971–976. https://doi.org/10.18653/v1/N16-1111

He, L., Xiao, R., & Yue, M. (2010). In pursuit of the "third code": Using the ZJU Corpus of Translational Chinese in translation studies. *Using Corpora in Contrastive and Translation Studies*, 182–214.

Helt, M. (2001). A multi-dimensional comparison of British and American spoken English. In D. Biber & S. Conrad (Eds.), *Variation in English: Multi-dimensional studies*. Routledge. https://doi.org/10.4324/9781315840888

Herbert, J. (1952). *The interpreter's handbook: How to become a conference interpreter*. Librairie de l'Université.

House, J. (2008). Beyond intervention: Universals in translation? *Trans-Kom*, *1*(1), 6–19.

Hu, K. B. (2012). Corpus translation studies: Connotations and implications. *Journal of Foreign Languages*, *35*(5), 59–70.

Hu, K. B. (2016). *Introducing corpus-based translation studies*. Shanghai Jiao Tong University Press. https://doi.org/10.1007/978-3-662-48218-6

Hu, K. B., & Tao, Q. (2009). A corpus-based study of explicitation of textual meaning in Chinese-English conference interpreting. *PLA Foreign Studies University Journal*, *4*, 67–73.

Hu, K. B., & Tao, Q. (2010). The compilation and use of the Chinese-English conference interpreting corpus. *Chinese Translators Journal*, *5*, 49–56.

Hu, K. B., & Tao, Q. (2013). The Chinese-English conference interpreting corpus: Uses and limitations. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *58*(3), 626–642. https://doi.org/10.7202/1025055ar

Hu, K. B., Tao, Q., & Wu, Y. (2007). Corpora and translation studies: Trend and problems? A critical review of the international symposium of corpora and translation studies. *Journal of Foreign Languages*, *5*, 64–69.

Hu, X., Xiao, R., & Hardie, A. (2016). How do English translations differ from non-translated English writings? A multi-feature statistical model for linguistic variation analysis. *Corpus Linguistics and Linguistic Theory*, *15*(2), 347–382. https://doi.org/10.1515/cllt-2014-0047

Huang, Libo. (2005).

Huang, Li-bo. (2008). Explicitation of personal pronoun subjects in English-Chinese

translation——A corpus-based investigation. *Journal of Foreign Language Teaching and Research*, *40*(6), 454–459.

Hyland, K. (2015). Corpora and written academic English. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 292–308). Cambridge University Press. https://doi.org/10.1017/CBO9781139764377.017

Iglesias Fernández, E. (2005). Bidirectionality in interpreting training in Spanish universities: An empirical study. *Communication & Cognition*, *Directionality in Interpreting. The "retour" or the "native"?*, 101–125.

Ilisei, I., & Inkpen, D. (2011). Translationese traits in Romanian newspapers: A machine learning approach. *International Journal of Computational Linguistics and Applications*, *2*(1–2), 319–332.

Ilisei, I., Inkpen, D., Pastor, G. C., & Mitkov, R. (2010). Identification of translationese: A machine learning approach. *Computational Linguistics and Intelligent Text Processing: 11th International Conference*, 503–511. https://www.academia.edu/406952/Identification_of_Translationese_A_Machine_Learning_Approach

Ishikawa, L. (1999). Cognitive explicitation in simultaneous interpreting. In A. Álvarez Lugris & A. Fernández Ocampo (Eds.), *Anovar/Anosar estudios de traducción e interpretación* (pp. 231–257). Universidade de Vigo.

Jaeger, F. (2010). Redundancy and reduction: Speakers manage syntactic information density. *Cognitive Psychology*, *61*(1), 23–62.

Jänis, M. (2002). From the A language to the B language and from the B language to the

A language—What is the difference? *CLUEB*, 1000–1012. https://doi.org/10.1400/34353

Jantunen, J. H. (2001). Synonymity and lexical simplication in translations: A corpus-based approach. *Across Languages and Cultures*, *2*(1), 97–112.

Jantunen, J. H. (2004). Untypical patterns in translations. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (Vol. 48, pp. 101–126). John Benjamins Publishing Company. https://benjamins.com/catalog/btl.48.09jan

Kade, O. (1968). *Zufall und gesetzmäßigkeit in der Übersetzung* (Vol. 1).

Kaiser, H. F. (1960). The application of electronic computers to factor analysis. *Educational and Psychological Measurement*, *20*, 141–151. https://doi.org/10.1177/001316446002000116

Kajzer-Wietrzny, M. (2012). *Interpreting universals and interpreting style* [Doctoral dissertation]. Adama Mickiewicza.

Kajzer-Wietrzny, M. (2015). Simplification in interpreting and translation. *Across Languages and Cultures*, *16*(2), 233–255. https://doi.org/10.1556/084.2015.16.2.5

Kajzer-Wietrzny, M. (2018). Interpretese vs. non-native language use: The case of optional that. In M. Russo, C. Bendazzoli, & B. Defrancq (Eds.), *Making way in corpus-based interpreting studies* (pp. 97–113). Springer. https://doi.org/10.1007/978-981-10-6199-8_6

Kenny, D. (1998). *Corpora in translation studies*. /paper/Corpora-in-translation-studies-Kenny/dd4c703fe93203b0a5911123b250e3a9a65c99fd

Kenny, D. (2001). *Lexis and creativity in translation: A corpus-based study*. St. Jerome

Publishing.

Klaudy, K. (2009). The asymmetry hypothesis in translation research. In R. Dimitriu & M. Shlesinger (Eds.), *Translators and their readers. In homage to Eugene A. Nida. Brussels: Les éditions du Hazard* (pp. 283–303).

Klaudy, K., & Károly, K. (2005). Implicitation in translation: Empirical evidence for operational asymmetry in translation. *Across Languages and Cultures*, *6*(1), 13–28. https://doi.org/10.1556/Acr.6.2005.1.2

Kruger, H. (2012). A corpus-based study of the mediation effect in translated and edited language. *Target. International Journal of Translation Studies*, *24*(2), 355–388. https://doi.org/10.1075/target.24.2.07kru

Kruger, H. (2017). The effects of editorial intervention. Implications for studies of the features of translated language. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical translation studies* (pp. 113–156). De Gruyter Mouton. https://doi.org/10.1515/9783110459586-005

Kruger, H. (2018). Expanding the third code: Corpus-based studies of constrained communication and language mediation. In L. Aguiar de Souza Penha Marion, S. Granger, & M.-A. Lefer (Eds.), *Book of abstracts. Using corpora in contrastive and translation studies conference* (5th ed., pp. 9–12). CECL Papers. https://alfresco.uclouvain.be/alfresco/service/guest/streamDownload/workspace/SpacesStore/c850523d-1953-4204-964d-c6d1ee174bfe/UCCTS2018_book_of_abstracts_with%20correction.pdf?guest=true#page=24

Kruger, H. (2019). That again: A multivariate analysis of the factors conditioning syntactic explicitness in translated English. *Across Languages and Cultures*, *20*(1), 1–33. https://doi.org/10.1556/084.001

Kruger, H., & De Sutter, G. (2018). Alternations in contact and non-contact varieties: Reconceptualising that-omission in translated and non-translated English using the MuPDAR approach. *Translation, Cognition & Behavior*, *1*(2), 251–290. https://doi.org/10.1075/tcb.00011.kru

Kruger, H., & Van Rooy, B. (2012). Register and the features of translated language. *Across Languages and Cultures*, *13*(1), 33–65. https://doi.org/10.1556/Acr.13.2012.1.3

Kruger, H., & Van Rooy, B. (2016a). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide*, *37*(1), 26–57. https://doi.org/10.1075/eww.37.1.02kru

Kruger, H., & Van Rooy, B. (2016b). Syntactic and pragmatic transfer effects in reported-speech constructions in three contact varieties of English influenced by Afrikaans. *Language Sciences*, *56*, 118–131. https://doi.org/10.1016/j.langsci.2016.04.003

Kujamäki, P. (2004). What happens to "unique items" in learners' translations?: "Theories" and "concepts" as a challenge for novices' views on "good translation." In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (Vol. 48, pp. 187–204). John Benjamins Publishing Company. https://benjamins.com/catalog/btl.48.16kuj

Lam, P. W. Y. (2010). Discourse particles in corpus data and textbooks: The case of well.

Applied Linguistics, *31*(2), 260–281. https://doi.org/10.1093/applin/amp026

Lanstyák, I., & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures*, *13*(1), 99–121. https://doi.org/10.1556/Acr.13.2012.1.6

Lapshinova-Koltunski, E., & Vela, M. (2015). Measuring `registerness' in human and machine translation: A text classification approach. *Proceedings of the Second Workshop on Discourse in Machine Translation*, 122–131. https://doi.org/10.18653/v1/W15-2517

Laviosa, S. (1997). How comparable can "comparable corpora" be? *Target. International Journal of Translation Studies*, *9*(2), 287–317.

Laviosa, S. (1998a). The English comparable corpus: A resource and a methodology. In M. Cronin, L. Bowker, D. Kenny, & J. Pearson (Eds.), *Unity in diversity: Current trends in translation studies.* (pp. 101–112). St. Jerome Publishing.

Laviosa, S. (1998b). The corpus-based approach: A new paradigm in translation studies. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *43*(4), 474–479. https://doi.org/10.7202/003424ar

Laviosa, S. (1998c). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta : Journal Des Traducteurs / Meta: Translators'Journal*, *43*(4), 557–570. https://doi.org/10.7202/003425ar

Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, applications* (Vol. 17). Rodopi.

Laviosa, S. (2008). Universals. In M. Baker & G. Saldanha (Eds.), *Routledge encyclopedia of translations studies* (2nd ed., pp. 306–310). Routledge.

Laviosa-Braithwaite, S. (1996). *The English comparable corpus (ECC): A resource and a methodology for the imperical study of translation* [Doctoral dissertation]. University of Manchester.

Li, D. C., & Wang, K. F. (2012). A corpus-based study on lexical patterns in simultaneous interpreting from Chinese into English. *Modern Foreign Languages*.

Lim, H.-O. (2003). Interpreting into B: To B or not to B? *FORUM. Revue Internationale d'interprétation et de Traduction / International Journal of Interpretation and Translation*, *1*(2), 151–171. https://doi.org/10.1075/forum.1.2.07lim

Lim, H.-O. (2005). Working into the B language: The condoned taboo? *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *50*(4). https://doi.org/10.7202/019870ar

Lin, P. M. S. (2013). The prosody of formulaic expression in the IBM/Lancaster Spoken English Corpus. *International Journal of Corpus Linguistics*, *18*(4), 561–588.

Lindemann, S., & Mauranen, A. (2001). "It's just real messy": The occurrence and function of just in a corpus of academic speech. *English for Specific Purposes*, *20*(1), 459–475. https://doi.org/10.1016/S0889-4906(01)00026-6

Liu, D. Z., Yang, P. X., & Zhou, X. Y. (2009). Translationese: Fanyiti? Fanyizheng? Fanyiqiang? *Chinese Terminology*, *3*, 52–54.

Lv, Q., & Liang, J. (2018). Is consecutive interpreting easier than simultaneous interpreting? – A corpus-based study of lexical simplification in interpretation. *Perspectives*, *27*(1), 91–106. https://doi.org/10.1080/0907676X.2018.1498531

Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*(3),

299–320. https://doi.org/10.1093/applin/ams010

Marzocchi, C., & Pöchhacker, F. (2015). “Parliamentary settings.” In *Routledge encyclopedia of interpreting studies* (pp. 297–299). Routledge.

Mauranen, A. (2000). Strange strings in translated language a study on corpora. In M. Olohan (Ed.), *Intercultural faultlines: Research models in translation studies v.1: Textual and cognitive aspects* (1st ed., pp. 119–141). Routledge. https://doi.org/10.4324/9781315759951-9

Mauranen, A. (2004). Corpora, universals and interference. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (Vol. 48, pp. 65–82). John Benjamins Publishing Company. https://doi.org/10.1075/btl.48.07mau

Mauranen, A. (2008). Universal tendencies in translation. In M. Rogers & G. Anderman (Eds.), *Incorporating corpora: The linguist and the translator* (pp. 32–48). Multilingual matters. https://researchportal.helsinki.fi/en/publications/universal-tendencies-in-translation

McDavid, V. (1964). *The alternation of “that” and zero in noun clauses*. *39*(2), 102–113.

Morselli, N. (2018). Interpreting universals: A study of explicitness in the intermodal EPTIC corpus. *InTRAlinea*, *Special issue: New findings in corpus-based interpreting studies*, 12.

Muller, S. (2005). Discourse markers in native and non-native English discourse. In *Pbns.138* (Vol. 138). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.138

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. OUP

Oxford.

Newmark, P. (1988). *A textbook of translation* (Vol. 66). Prentice-Hall International.

Nicodemus, B., & Emmorey, K. (2013). Direction asymmetries in spoken and signed language interpreting. *Bilingualism (Cambridge, England)*, *16*(3), 624–636. https://doi.org/10.1017/S1366728912000521

Nini, A. (2014). *Multidimensional analysis tagger*. https://sites.google.com/site/multidimensionaltagger

Niska, H. (1999). *Text linguistic models for the study of simultaneous interpreting*. 76.

Ochs, E. (1979). Planned and unplanned discourse. *Discourse and Syntax*. https://doi.org/10.1163/9789004368897_004

Olohan, M. (2003). How frequent are the contractions?: A study of contracted forms in the translational English corpus. *Target. International Journal of Translation Studies*, *15*(1), 59–89. https://doi.org/10.1075/target.15.1.04olo

Olohan, M. (2004). *Introducing corpora in translation studies*. Routledge.

Olohan, M., & Baker, M. (2000). Reporting that in translated English. Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, *1*(2), 141–158. https://doi.org/10.1556/Acr.1.2000.2.1

Øverås, L. (1998). In search of the third code: An investigation of norms in literary translation. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *43*(4), 557–570. https://doi.org/10.7202/003775ar

Padilla, P. (2005). Cognitive implications of the English-Spanish direction for the quality and the training of simultaneous interpreting. *Communication & Cognition.*

Monographies, *38*(1–2), 47–62.

Pápai, V. (2004). A universal of translated text? In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (Vol. 48, pp. 143–164). John Benjamins Publishing Company. https://benjamins.com/catalog/btl.48.12pap

Pawley, A., & Syder, F. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 191–226). Longman.

Perego, E. (2003). Evidence of explicitation in subtitling: Towards a categorisation. *Across Languages and Cultures*, *4*(1), 63–88. https://doi.org/10.1556/Acr.4.2003.1.4

Pöchhacker, F. (2004). *Introducing interpreting studies*. Routledge.

Pöchhacker, F. (2016). *Introducing interpreting studies* (2nd ed.). Routledge.

Pöchhacker, F. (2019). Moving boundaries in interpreting. In M. N. Brøgger, H. V. Dam, & K. K. Zethsen (Eds.), *Moving boundaries in translation studies* (pp. 45–63). Routledge. https://doi.org/10.4324/9781315121871-4

Popescu, M. (2011). Studying translationese at the character level. *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, 634–639.

Prieels, L., Delaere, I., Plevoets, K., & De Sutter, G. (2015). A corpus-based multivariate analysis of linguistic norm-adherence in audiovisual and written translation. *Across Languages and Cultures*, *16*(2), 209–231. https://doi.org/10.1556/084.2015.16.2.4

Puurtinen, T. (2003). Genre-specific features of translationese? Linguistic differences between translated and non-translated Finnish children's literature. *Literary and*

Linguistic Computing, *18*(4), 389–406. https://doi.org/10.1093/llc/18.4.389

Puurtinen, T. (2004). Explicitation of clausal relations: A corpus-based analysis of clause connectives in translated and non-translated Finnish children's literature. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (pp. 165–176). John Benjamins Publishing Company. https://www.jbe-platform.com/content/books/9789027295835-13puu

Pym, A. (2004). Text and risk in translation. *Choice and Difference in Translation. The Specifics of Transfer*, 27–42.

Pym, A. (2005). Explaining explicitation. In A. Fóris & K. Károly (Eds.), *New trends in translation studies: In honour of Kinga Klaudy* (pp. 29–34). Akademiai Kiado. https://pdfs.semanticscholar.org/1458/0f1d2f4a733d1b8f58238bc2e90464b1825b.pdf

Pym, A. (2007). On Shlesinger's proposed equalizing universal for interpreting. In A. L. Jakobsen, I. M. Mees, & F. Pöchhacker (Eds.), *Interpreting studies and beyond: A tribute to Miriam Shlesinger* (Vol. 35, pp. 175–190). Samfundslitteratur.

Pym, A. (2008a). *On Toury's laws of how translators translate*. https://doi.org/10.1075/btl.75.24pym

Pym, A. (2008b). On omission in simultaneous interpreting: Risk analysis of a hidden effort. *Efforts and Models in Interpreting and Translation Research*, 83–105. https://doi.org/10.1075/btl.80.08pym

Pym, A. (2011). Translation research terms: A tentative glossary for moments of perplexity and dispute. *Translation Research Projects 3: Intercultural Studies Group*,

75–99.

Pym, A. (2015). Translating as risk management. *Journal of Pragmatics*, *85*, 67–80. https://doi.org/10.1016/j.pragma.2015.06.010

Quaglio, P. (2009). *Television dialogue: The sitcom friends vs. natural conversation*. John Benjamins Publishing Company.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J., Leech, E. P. of E. L. G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. Pearson Longman.

Rabadán, R., Labrador, B., & Ramon, N. (2009). Corpus-based contrastive analysis and translation universals A tool for translation quality assessment English → Spanish. *Babel*, *55*(4), 303–328. https://doi.org/10.1075/babel.55.4.01rab

Rabinovich, E., Wintner, S., & Lewinsohn, O. (2015). *The Haifa corpus of translationese*.

Rayson, P. (2009). *Wmatrix: A web-based corpus processing environment* (Web based) [Computer software]. Computing department, Lancaster University. http://ucrel.lancs.ac.uk/wmatrix/

Redeker, G. (1984). On differences between spoken and written language. *Discourse Processes*, *7*(1), 43–55. https://doi.org/10.1080/01638538409544580

Redelinghuys, K. (2016). Levelling-out and register variation in the translations of experienced and inexperienced translators: A corpus-based study. *Stellenbosch Papers in Linguistics*, *45*. https://doi.org/10.5774/45-0-198

Renouf, A., & Sinclair, J. (1991). Collocational frameworks in English. In K. Aijmer & B. Altenberg (Eds.), *Advances in corpus linguistics* (pp. 128–143). Rodopi.

Reppen, R. (1994). *Variation in elementary student language: A multi-dimensional*

perspective. [Doctoral dissertation, Northern Arizona University]. https://www.elibrary.ru/item.asp?id=5692488

Reppen, Randi. (2010). *Using corpora in the language cassroom*. Cambridge University Press.

Rodriguez-Castro, M. (2011). Translationese and punctuation: An empirical study of translated and non-translated international newspaper articles (English and Spanish). *Translation and Interpreting Studies*, *6*(1), 40–61.

Rohdenburg, G. (1996). Cognitive complexity and increased grammatical explicitness in English. *Cognitive Linguistics*, *7*(2), 149–182.

Russo, M., Bendazzoli, C., & Sandrelli, A. (2006). Looking for lexical patterns in a trilingual corpus of source and interpreted speeches: Extended analysis of EPIC (European Parliament Interpreting Corpus). *Forum*, *4*(1), 221–254. https://doi.org/10.1075/forum.4.1.10rus

Sandrelli, A., & Bendazzoli, C. (2005). *Lexical patterns in simultaneous interpreting: A preliminary investigation of EPIC (European Parliament Interpreting Corpus)*. 18.

Sandrelli, A., Bendazzoli, C., & Russo, M. (2010). European Parliament Interpreting Corpus (EPIC): Methodological issues and preliminary results on lexical patterns in simultaneous interpreting. *International Journal of Translation*, *22*, 165–203.

Santos, D. (1995). On grammatical translationese. *Short Papers Presented at the Tenth Scandinavian Conference on Computational Linguistics*, 59–66.

Schäffner, C., & Adab, B. (2001). The idea of the hybrid text in translation: Contact as conflict. *Across Languages and Cultures*, *2*(2), 167–180.

https://doi.org/10.1556/Acr.2.2001.2.1

Schjoldager, A. (1995). An exploratory study of translational norms in simultaneous interpreting: Methodological reflections. *HERMES - Journal of Language and Communication in Business*, *8*(14), 65. https://doi.org/10.7146/hjlcb.v8i14.25096

Schmid, H. (1994). Probabilistic part-of-speech tagging using decision trees. *Proceedings of International Conference on New Methods in Language Processing*. https://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/data/tree-tagger1.pdf

Scott, M. (1999). *WordSmith Tools* (Version 3) [Computer software]. Oxford University Press. https://lexically.net/wordsmith/downloads/

Scott, M. (2012). *WordSmith Tools* (Version 6) [Computer software]. Lexical Analysis Software. https://lexically.net/wordsmith/downloads/

Seel, O. I. (2005). Non-verbal means as culture-specific determinants that favour directionality into foreign language in simultaneous interpreting. In R. Godijns & M. Hinderdael (Eds.), *Directionality in interpreting: The "retour" or the native?* (pp. 63–82). Communications & Cognition.

Seleskovitch, D. (1978). Language and cognition. In D. Gerver & H. W. Sinaiko (Eds.), *Language interpretation and communication* (pp. 333–341). Springer US. https://doi.org/10.1007/978-1-4615-9077-4_29

Seleskovitch, D. (1987). La traduction interprétative. *Palimpsestes. Revue de Traduction*, *1*, 41–50.

Seleskovitch, D. (1999). The teaching of conference interpretation in the course of the last 50 years. *History of Interpreting*, *4*(1), 55–66.

Seleskovitch, D., & Lederer, M. (1989). *Pédagogie raisonnée de l'interprétation*. Didier erudition.

Setton, R. (2011). Corpus-based interpreting studies (CIS): Overview and prospects. In A. Kruger, K. Wallmach, & J. Munday (Eds.), *Corpus-based translation studies: Research and applications* (pp. 14–32). Continuum International Publishing Group.

Shlesinger, M. (1989). *Simultaneous interpretation as a factor in effecting shifts in the position of texts in the oral-literate continuum* [MA thesis, Tel Aviv University]. https://doi.org/10.13140/RG.2.2.31471.69285

Shlesinger, M. (1991). Interpreter latitude vs. due process: Simultaneous and consecutive interpretation in multilingual trials. In S. Tirkkonen-Condit (Ed.), *Empirical research in translation and intercultural studies* (pp. 147–155). Narr.

Shlesinger, M. (1995). Shifts in cohesion in simultaneous interpreting. *The Translator*, *1*(2), 193–214. https://doi.org/10.1080/13556509.1995.10798957

Shlesinger, M. (1998). Corpus-based interpreting studies as an offshoot of vorpus-based translation studies. *Meta : Journal Des Traducteurs / Meta: Translators' Journal*, *43*(4), 486–493. https://doi.org/10.7202/004136ar

Shlesinger, M. (2008). Towards a definition of interpretese: An intermodal, corpus-based study. In G. Hansen, A. Chesterman, & H. Gerzymisch-Arbogast (Eds.), *Efforts and models in interpreting and translation research: A tribute to Daniel Gile* (Vol. 80, pp. 237–253). John Benjamins Publishing Company. https://doi.org/10.1075/btl.80.18shl

Shlesinger, M., & Ordan, N. (2012). More spoken or more translated?: Exploring a known

unknown of simultaneous interpreting. In E. Brems, R. Meylaerts, & L. van Doorslaer (Eds.), *Benjamins current topics* (Vol. 69, pp. 47–64). John Benjamins Publishing Company. https://doi.org/10.1075/bct.69.04shl

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Staples, S. (2015). Spoken discourse. In D. Biber & R. Reppen (Eds.), *The Cambridge handbook of English corpus linguistics* (pp. 271–291). Cambridge University Press.

Staples, S., & Biber, D. (2014). The expression of stance in nurse-patient interactions: An ESP perspective. *Linguistic Insights*, *200*, 123–142.

Steiner, E. (2008). Empirical studies of translations as a mode of language contact—"explicitness" of lexicogrammatical encoding as a relevant dimension. In *Language contact and contact languages* (pp. 317–341). https://doi.org/10.1075/hsm.7.18ste

Steiner, E. (2012). A characterization of the resource based on shallow statistics. In S. Hansen-Schirra, S. Neumann, & E. Steiner (Eds.), *Crosslinguistic corpora for the study of translations: Insights from the language pair English–German* (pp. 71–89). Walter de Gruyter.

Stewart, D. (2000). Conventionality, creativity and translated text: The implications of electronic corpora in translation. In O. Maeve (Ed.), *Intercultural faultlines: Research models in translation studies: V. 1: Textual and cognitive aspects* (pp. 73–92). Routledge.

Straniero Sergio, F., & Falbo, C. (2012). *Breaking ground in corpus-based interpreting studies*. Peter Lang.

Swales, J., & Burke, A. (2003). “It’s really fascinating work”: Differences in evaluative

adjectives across academic registers. *Language and Computers*, 1–18.

Swales, J. M., & Malczewski, B. (2001). Discourse management and new-episode flags in MICASE. In J. M. Swales & R. Simpson-Vlach (Eds.), *Corpus linguistics in North America* (pp. 145–163). The University of Michigan Press.

Szymor, N. (2018). Translation: Universals or cognition? *Target. International Journal of Translation Studies*, *30*(1), 53–86.

Tagliamonte, S., & Smith, J. (2005). No momentary fancy! The zero 'complementizer' in English dialects. *English Language & Linguistics*, *9*(2), 289–309. https://doi.org/10.1017/S1360674305001644

Tang, F. (2014). *An empirical study of explicitation patterns in consecutive interpreting: A comparison between professional and novice interpreters* [Doctoral dissertation]. The Hong Kong Polytechnic University.

Tang, F. (2018). *Explicitation in consecutive interpreting* (Vol. 135). John Benjamins Publishing Company. https://doi.org/10.1075/btl.135

Tannen, D. (1982). Spoken and written language: Exploring orality and literacy. *Norwood, N.J.: Ablex*.

Tannen, D. (1985). Relative focus on involvement in oral and written discourse. In D. R. Olson, N. Torrance, & A. Hildyard (Eds.), *Literacy, language, and learning: The nature and consequences of reading and writing* (pp. 124–147). CUP Archive.

Teich, E. (2003). *Cross-linguistic wariation in system and text* (Reprint 2011 ed. edition, Vol. 5). De Gruyter Mouton.

Thompson, S., & Mulac, A. (1991). The discourse conditions for the use of the

complementizer that in conversational English. *Journal of Pragmatics*, *15*(3), 237–251. https://doi.org/10.1016/0378-2166(91)90012-M

Tirkkonen-Condit, S. (2004). Unique items—Over- or under-represented in translated language? In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist?* (Vol. 48, pp. 177–184). John Benjamins Publishing Company. https://benjamins.com/catalog/btl.48.14tir

Tommola, J., & Helevä, M. (1998). Language direction and source text complexity: Effects ontrainee performance in simultaneous interpretin. In L. Bowker, M. Cronin, D. Kenny, & J. Pearson (Eds.), *Unity in diversity? Current trends in translation studies.* (pp. 177–186). St. Jerome Publishing.

Toury, G. (1980). *In search of a theory of translation*. Porter Institute for Poetics and Semiotics, Tel Aviv University.

Toury, G. (1995). *Descriptive translation studies and beyond*. John Benjamins Publishing Company.

Toury, G. (2012). *Descriptive translation studies-and beyond: Revised edition* (Vol. 100). John Benjamins Publishing Company.

Toutanova, K., Klein, D., Manning, C. D., & Singer, Y. (2003). Feature-rich part-of-speech tagging with a cyclic dependency network. *Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology - NAACL '03*, *1*, 173–180. https://doi.org/10.3115/1073445.1073478

Van Rooy, B., Haase, C., Schmied, J., & Terblanche, L. (2010). Register differentiation

in East African English: A multidimensional study. *English World-Wide*, *31*(3), 311–349. https://doi.org/10.1075/eww.31.3.04van

Vanderauwera, R. (1985). *Dutch novels translated into English: The transformation of a "minority" literature* (Vol. 6). Rodopi.

Vandevoorde, L., De Sutter, G., & Plevoets, K. (2016). On semantic differences between translated and non-translated Dutch: Using bidirectional corpus data for measuring and visualizing distances between lexemes in the semantic field of inceptiveness. In J. Meng (Ed.), *Empirical translation studies: Interdisciplinary methodologies explored* (pp. 128–146). Equinox Publishing. https://doi.org/10.1558/equinox.24835

Vilinsky, B. (2012). On the lower frequency of occurrence of Spanish verbal periphrases in translated texts as evidence for the unique items hypothesis. *Across Languages and Cultures*, *13*(2), 197–210. https://doi.org/10.1556/Acr.13.2012.2.4

Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, *30*(1), 98–118. https://doi.org/10.1093/llc/fqt031

Wang, B., & Feng, D. (2018). A corpus-based study of stance-taking as seen from critical points in interpreted political discourse. *Perspectives*, *26*(2), 246–260. https://doi.org/10.1080/0907676X.2017.1395468

Williams, D. A. (2005). *Recurrent features of translation in Canada: A corpus-based study* [Doctoral dissertation, University of Ottawa (Canada)]. https://doi.org/10.20381/ruor-12864

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language & Communication - LANG COMMUN*, *20*(1), 1–28. https://doi.org/10.1016/S0271-5309(99)00015-4

Wu, M. (2001). The importance of being strategic: A strategic approach to the teaching of simultaneous interpreting. *Studies of Translations and Interpretation*, *6*, 79–92.

Xiao, R. (2009). Multidimensional analysis and the study of world Englishes. *World Englishes*, *28*(4), 421–450. https://doi.org/10.1111/j.1467-971X.2009.01606.x

Xiao, R. (2011). Word clusters and reformulation markers in Chinese and English: Implications for translation universal hypotheses. *Languages in Contrast*, *11*(2), 145–171. https://doi.org/10.1075/lic.11.2.01xia

Xiao, X. Y. (2015). *On the oral-literate continuum: A corpus-based study of interpretese*. Xiamen University Press. https://core.ac.uk/download/pdf/41391426.pdf

Zanettin, F. (2013). Corpus methods for descriptive translation studies. *Procedia - Social and Behavioral Sciences*, *95*, 20–32. https://doi.org/10.1016/j.sbspro.2013.10.618

Zipf, G. K. (1949). *Human behavior and the principle of least effort* (pp. xi, 573). Addison-Wesley Press.

Zufferey, S., & Cartoni, B. (2014). A multifactorial analysis of explicitation in translation. *Target. International Journal of Translation Studies*, *26*(3), 361–384. https://doi.org/10.1075/target.26.3.02zuf