More *spoken* or more *translated*?

Exploring the known unknowns of simultaneous interpreting from a multidimensional analysis perspective

Cui Xu and Dechao Li
Beijing Institute of Technology | Hong Kong Polytechnic University

This article explores the features of simultaneous interpreting (SI) from a multidimensional (MD) analysis perspective (Biber 1988), drawing on a newly built comparable intermodal corpus, the LegCo+ corpus. The corpus incorporates Cantonese speeches that are both interpreted and translated into English, as SI and written translation (WT), respectively. Additionally, a third English corpus consisting of English native speeches (NS), without mediation, serves as a benchmark comparison. We aim to examine the extent of similarities and differences between SI, NS and WT in terms of the linguistic patterns they display. Our findings show that: (1) SI is a hybrid language mode, exhibiting features that lie between those of non-mediated spoken language and mediated written language; (2) in terms of its spoken nature, SI resembles NS in certain dimensions where typical features are associated with orality, suggesting a strong modality effect; and (3) in terms of its mediated status, SI demonstrates similarities with WT, despite their perceptibly distinct modalities, pointing to a potential mediation-specific effect. These empirical findings emphasize the necessity of understanding the multidimensionality inherent in interpreted language.

Keywords: linguistic features, simultaneous interpreting, multidimensional analysis, LegCo+ corpus, mediation, modality

1. Introduction

Intuitively speaking, simultaneous interpreting (SI) in the spoken modality is perceptibly distinct from written translation due to its intrinsically different mediation mode or modality. However, Shlesinger (2008) and Shlesinger and Ordan (2012) have challenged this intuitive understanding, sparking further debates

on the impacts of modality (oral vs. written) and ontology (translated vs. non-translated; see Section 2.1) on the linguistic manifestations of interpreted outputs. Their ultimate goal has been to isolate features of SI as being both spoken and translated discourse (referred to as 'interpretese') so as to deepen our understanding of the complexities of oral translations, instead of subsuming them under the generic concept of 'translation'.

Their concern has been acknowledged gradually, thanks to the introduction of corpus-linguistic methods to Interpreting Studies (Shlesinger 1998). The compilation of comparable intermodal corpora (Shlesinger and Ordan 2012; Bernardini, Ferraresi, and Miličević 2016) in particular, such as the European Parliament Translation and Interpreting Corpus (EPTIC) and Translation and Interpreting Corpus (TIC), has facilitated research involving the intermodal comparison of translation and interpreting. Overall, linguistic manifestations of interpreted outputs are found to be more salient than translated ones, especially in terms of lexical simplification (Kajzer-Wietrzny 2015; Bernardini, Ferraresi, and Miličević 2016; Ferraresi et al. 2018), lending support to significant modality-dependent differences.

However, it should be noted that previous intermodal comparisons have mostly been unidimensional, meaning that they often consider only a single dimension such as simplification or explicitation. While salient differences may exist along one dimension between translation and interpreting, there may be hidden patterns shared between them along other dimensions. Effects of either modality or ontology from a unidimensional perspective should be interpreted with great caution. In this study, following the research of Shlesinger (1998, 2008), and Shlesinger and Ordan (2012), we explore the linguistic properties of interpreted outputs in relation to both translated and non-mediated (i.e., non-translated spontaneous) texts from a multidimensional perspective. We aim to answer the same questions as put forward in Shlesinger and Ordan (2012); that is,

> 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 translational ontology; and in what ways corpus-based translation/interpreting studies may deepen our understanding of interpreting as distinct linguistic, cognitive and textual phenomenon. (44)

Compared with their univariate method, we adopt a multivariate and multidimensional approach. It is multivariate in that we analyze seventy-nine linguistic features (see Section 3) across three language varieties, namely, spontaneous native speech (NS), SI, and written translation (WT) to explore the co-occurrence patterns identified in different dimensions. The premise for this multidimensional analysis is the acknowledgement of the multidimensional and

multifaceted nature (see De Sutter and Lefer 2020) of language in general, irrespective of its mediation status. The same opinion has also been expressed by Biber (1986) on register variation in spoken and written language varieties, in which he explains 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" (385). We hold the same view for translation and interpreting, given their constrained nature as forms of mediated language, and therefore expect more than one dimension to characterize similarities and/or differences between them. In addition, we will also investigate ontological differences and/or similarities between interpreted speech and spontaneous native speech.

In Section 2, we review studies on ontological and intermodal comparisons in Corpus-based Translation and Interpreting Studies. In Section 3, we introduce our research data and the methodology of the multidimensional analysis. Section 4 reports the findings of the multidimensional analysis, with similarities as well as differences among the three varieties highlighted. Section 5 discusses in detail the possible reasons for our findings, along with their implications. The last section summarizes our main findings concerning the multidimensional nature of interpreted language.

2. Studies on ontology versus modality

2.1 The effect of ontology

The effect of ontology (or in other words, the status of being translated — i.e., mediated — or not) has been a matter of ongoing debate in Translation Studies since the marriage between corpus linguistics and Descriptive Translation Studies (Baker 1995, 1996; Chesterman 2017). The majority of the extant research has concluded that translated language is generally more simplified, explicit and normalized than original non-translated/non-mediated texts in the target language, demonstrating a strong ontology-dependent effect (Baker 1996; Olohan and Baker 2000; Kruger and Van Rooy 2012), although conflicting results have also been reported. This conclusion has also been supported by empirical studies based on machine learning, which report that translated texts can be automatically distinguished from non-translated originals in the same language (Baroni and Bernardini 2006; Volansky, Ordan, and Wintner 2015; Avner, Ordan, and Wintner 2016), lending further support to the claim of strong ontology-dependent patterns of translated language.

The question of ontology was first taken up by Shlesinger (1998) in Interpreting Studies, in which she called for the introduction of corpus-linguistic methods to Interpreting Studies. As a form of spoken as well as mediated language, it might be expected that interpreting demonstrates linguistic patterns distinct from both non-mediated spoken language and written translated language, leading to a phenomenon that Shlesinger (1998) called 'interpretese'. Her proposal, however, was not adopted until Sandrelli and Bendazzoli (2005) provided substantial evidence for the contested and controversial simplification hypothesis, defined as "the process and/or result of making do with less words" (Blum-Kulka and Levenston 1978, 399) associated with translation. Comparing interpreted language with non-mediated spoken language, Sandrelli and Bendazzoli (2005) reported contradictory results in terms of lexical density and list heads (used as measures of complexity) in different language combinations (Spanish and Italian to English, and Spanish and English to Italian). Their mixed findings suggest that the claim of ontology-dependent patterns of written translations cannot be readily applied to interpreting, and other factors such as language pair and working direction may turn out to be more influential than ontology.

In addition to Sandrelli and Bendazzoli (2005), many others have tried to explore the effect of ontology on interpreting-specific patterns with regard to simplification, explicitation, and normalization (see, e.g., Russo, Bendazzoli, and Sandrelli 2006; Shlesinger 2008; Dayter 2018; Kajzer-Wietrzny 2012, 2015). None of these studies offer straightforward and unequivocal ontology-specific patterns characterizing interpreted language in relation to non-mediated spoken language. For example, Kajzer-Wietrzny (2015) finds that interpreted language is not only lexically denser but also more repetitive than non-mediated spoken language, a finding indicating two contradictory linguistic patterns. Dayter (2018) reports that interpreted English (from Russian) and interpreted Russian (from English) show opposite trends with respect to simplification and explicitation, respectively. That is, English-Russian translations show both simplification and explicitation, while Russian-English demonstrate neither. However, comparable analyses based on the Chinese/English language pair (Hu and Tao 2009; Li and Wang 2012) generally confirm the ontology-dependent patterns in translated language, though exceptions also exist (Qin and Wang 2009; Chen and Cui 2010).

Two major implications can be inferred from the reviewed studies. First, given its spoken and mediated status, the nature of interpreting may be much more complicated than that of translation, thus giving rise to the mixed patterns reported. Second, there may be other factors exerting stronger influence than the status of being translated/mediated or not (i.e., ontology), such as language pair (genetically distinct?), working direction (A-to-B or B-to-A?), and modality (spoken? written? signed?). In fact, a rising number of multifactorial corpus-based

studies (e.g., Kruger and De Sutter 2018; Kruger 2019; Kajzer-Wietrzny 2022) have been conducted in recent years to disentangle the possible contributors to the potentially distinct patterns characterizing translation and/or interpreting.

2.2 The effect of modality

The question of modality came to the spotlight after Shlesinger (1998) called for academic attention to the special status of oral translations, namely, interpreting, as being both spoken and translated/mediated discourse. However, early attempts (Sandrelli and Bendazzoli 2005; Russo, Bendazzoli, and Sandrelli 2006) exploring the nature of interpreted language have focused only on the effect of ontology (being translated or not) (see Section 2.1). Due to the contradictory results reported by these studies, questions have been raised concerning the applicability of the research methodology adopted in Corpus-based Translation Studies to Interpreting Studies (Kajzer-Wietrzny 2012), and doubts have also been cast upon the generalization of any recurrent patterns identifiable in translation to interpreting (Shlesinger 1998; Shlesinger and Ordan 2012). Shlesinger (2008) points out that

> few (computerized) corpus-based studies have attempted to discern features of the different modes or modalities of translation, or to pinpoint features of interpreted — as opposed to translated — texts, so as to refine our largely intuitive knowledge about the properties of interpreted outputs as such — and by extension, to shed light on the properties of constrained spoken discourse. (239)

She suggests the application of corpus-linguistic methods to analyze interpreted data, setting aside pre-existing expectations about mediated modalities. Studies that have incorporated experimental or corpus data with respect to the effects of modality have focused on SI in the spoken modality versus sight translation (Agrifoglio 2004), Consecutive Interpreting (CI) versus SI (Gile 2001; Lv and Liang 2018), SI in the spoken modality versus SI in the signed modality (Russell 2002), and most commonly, SI in the spoken modality versus WT (Kajzer-Wietrzny 2012, 2015, 2022; Shlesinger and Ordan 2012; Bernardini, Ferraresi, and Miličević 2016; Ferraresi et al. 2018; Kajzer-Wietrzny and Ivaska 2020).

Among these studies, the one by Shlesinger and Ordan (2012) is of particular relevance to our study. Building on previous work (Shlesinger 1998, 2008), Shlesinger and Ordan (2012) set out to test "the hypothesis that modality may exert a stronger effect than ontology — i.e. that being oral (vs. written) is a more powerful influence than being translated (vs. original)" (43). Their results show that there is a tendency towards orality in interpreted language, and generally, interpreting is more spoken than translated. In other words, the very fact of inter-

preting as a form of spoken language (i.e., modality) overrides its translational/mediated ontology. This may offer one explanation for why the widely attested hypotheses about the typicalities of WT (such as increased explicitness and simplicity) do not appear to apply to interpreting.

Other corpus-based intermodal studies also report a stronger influence of modality on the linguistic patterns of interpreted outputs. For example, drawing on the EPTIC, Bernardini, Ferraresi, and Miličević (2016) and Ferraresi et al. (2018) observe that interpreters tend to simplify their outputs more than translators, lending further support to the modality-driven differences. Another study by Defrancq, Plevoets, and Magnifico (2015) also reports different usage patterns of connectives in translated and interpreted texts: while translators mostly add connectives to explicitate, interpreters do so to either fill gaps or cover misinterpretations. This suggests distinct working mechanisms underlying translation and interpreting as two modes of mediation. Nonetheless, viewed from a different perspective, these studies do reveal similarities or shared patterns between translation and interpreting, since both mediation modes are found to be characterized by more simplified and explicit language use when compared to their non-mediated counterparts, albeit to different extents.

In contrast to previous comparable studies on the effect of ontology, intermodal studies as reviewed here seem to offer more consistent evidence for the stronger influence of modality compared to ontology on the linguistic features of interpreted language. However, we argue that previous findings on the effect of ontology and/or modality should be interpreted with great caution. Specifically, the majority of previous research relies heavily on univariate and unidimensional analysis, often with a focus on one single linguistic pattern (such as simplification or explicitation). While certain differences or similarities may characterize one dimension, the opposite may be true in other unexplored dimensions, since "[s]ystematic properties of text […] are hardly ever observable on the basis of just a single feature" (Evert and Neumann 2017, 2).

2.3 The effects of ontology and/or modality from a multidimensional perspective

With the advancement of statistical tools, a rising number of translation scholars (Kruger and De Sutter 2018; Kruger 2019; Kajzer-Wietrzny and Ivaska 2020; Kajzer-Wietrzny 2022) have started to capitalize on multivariate statistical techniques to explore "the multifaceted nature" of translation and interpreting (Liang and Lv 2020). Among these methods, the multidimensional analysis (MD) approach pioneered by Biber (1986, 1988) for the analysis of register variation offers an insightful method for uncovering the multidimensionality of mediated

language varieties, more recently also referred to as one type of constrained language variety (see Kruger and Van Rooy 2012, 2016; Kotze 2022; Kotze and Van Rooy 2024).

As will be illustrated in Section 3.2, the MD approach of Biber (1986, 1988) investigates register variation in different spoken and written text types, under the assumption that the linguistic possibilities offered by a language cannot be fully accounted for by a single dimension encompassing several lexico-grammatical features. Translation and interpreting undoubtedly offer a variety of linguistic possibilities that often give rise to different output texts. However, the multidimensional nature of translation and interpreting has only recently garnered attention due to the search for "translation universals" (Baker 1993), which has produced conflicting results (Kruger and Van Rooy 2018; Xu and Li 2022). Consequently, research in Translation and Interpreting Studies utilizing MD analysis remains limited, with a few notable exceptions such as Xiao (2015); Hu, Xiao, and Hardie (2016); and Kruger and Van Rooy (2016, 2018).

Inspired by Shlesinger's MA thesis (1989), Xiao (2015) looks into the role of SI on the oral–literate continuum in Chinese–English and English–Chinese interpreting from an MD perspective. Integrating the dimensions categorized in Biber (1988) with Shlesinger's (1989) dimensions (or 'parameters of orality/literacy'), Xiao (2015) extracts altogether five dimensions for the English texts and four dimensions for the Chinese texts. Based on frequency comparison of twenty-one (of sixty-seven) linguistic features identified by Biber (1988), she reports an overall equalizing effect of interpreted language (see Shlesinger 1989), albeit with mixed findings. That is, interpreted texts show more features typical of oral language when source texts are written, or deliberately composed. By contrast, they tend to be more written-like when source texts are more oral. Exploring the effect of mediation on the linguistic patterns of the interpreted output, Xiao's study serves as the first of its kind to explore the multidimensionality of interpreting.

Apart from Xiao's study (2015) on interpreting, the works of Kruger and Van Rooy (2016, 2018) are among the earliest attempts to investigate translation within an MD framework. In Kruger and Van Rooy (2016), the authors compare translated English with a non-native indigenized variety of English to determine whether the two language varieties exhibit similarities due to shared constraints related to bilingual production. Their study reports a number of similarities between the two constrained varieties, represented particularly by their shared preference for increased explicitness and formality. They attribute these similarities to "the psycholinguistic processing constraints imposed by bilingual mode" (44) and a possible risk-aversion consideration shared by text producers, namely, translators and non-native writers. In a follow-up study, Kruger and Van Rooy (2018) expand their comparison to sixteen written contact varieties of English,

which include Native Englishes, Non-native Englishes, and Translated English, demonstrating varying degrees of contact, to explore the effects of register and variety on the underlying dimensions of register variation. The results show that there are both register differences and differences between different English varieties; however, overall, contact varieties, which share psycholinguistic and social constraints, display an avoidance of informal features and a preference for a more informational and less involved style, especially when the contact varieties are less established or institutionalized (for example, Translated English and Non-native Asian Englishes).

Another representative study closely related to the current study has been carried out by Hu, Xiao, and Hardie (2016) on translation universals. Adapting Biber's (1988) MD analysis approach, they investigate a total of ninety-six lexicogrammatical features with the aim of identifying a 'translational dimension' that distinguishes translated from non-translated texts. In line with their expectations, the MD analysis indeed reveals a 'translational dimension' incorporating linguistic features previously identified in scholarship, such as lexical simplification, and features indexing increased explicitness and normalization. Their study demonstrates the applicability of the MD approach in translation research, while also shedding light on the multidimensional nature of translated language.

In addition to studies that specifically employ Biber's (1988) MD approach, there are several other studies (Kruger and De Sutter 2018; Kruger 2019; De Sutter and Vermeire 2020; Kajzer-Wietrzny and Ivaska 2020; Kajzer-Wietrzny and Grabowski 2021; Kajzer-Wietrzny 2022) that investigate the effects of ontology and/or modality using multifactorial analysis methods. These studies are primarily focused on "constrained communication" (Lanstyák and Heltai 2012) or "constrained language varieties" (Kotze 2022; Kotze and Van Rooy 2024). Of particular relevance to our research are the studies by Kajzer-Wietrzny and Ivaska (2020), and Kajzer-Wietrzny (2022).

Starting from an intermodal perspective, Kajzer-Wietrzny and Ivaska (2020) investigate lexical diversity in written and spoken modes of constrained and non-constrained language varieties by making use of multivariate approaches (factor analysis and linear mixed-effects regression models). Through a detailed analysis of lexical density, variability evenness, dispersion, rarity, and semantic disparity across six English varieties with distinct ontological status as well as modes of mediation, their study demonstrates that constrained language varieties (i.e., translation, interpreting, non-native spoken and non-native written English) are generally less lexically diverse than non-constrained ones (i.e., native spoken and native written English). Nevertheless, intermodal differences override intervarietal ones, meaning that greater divergences are found between spoken and written language varieties, regardless of their status of mediation/constrainedness. Fur-

thermore, the mode of delivery (i.e., whether the source texts are read out or delivered impromptu) also exerts an influence on the lexical patterns of the two mediated varieties.

Using the same corpus of six English varieties, Kajzer-Wietrzny (2022) analyzes cohesion in spoken and written registers of constrained and non-constrained language varieties also with the aid of multivariate approaches (including mosaic plots, correspondence analysis, and regression modelling). She finds that, generally, constrained language varieties do diverge from non-constrained ones, but differences also exist among the former, specifically for non-native descriptive language (non-native spoken and written English) versus mediated interpretative language (translation and interpreting). Once again, source event delivery mode is shown to have an influence on the distribution patterns of cohesive devices in both spoken and written texts, regardless of their constrained/non-constrained status. As far as translation and interpreting are concerned, these two forms of mediated language resemble each other in "[t]he overall frequency of cohesive devices (excluding phrase-level coordinators)" (Kajzer-Wietrzny 2022, 155), which points to an increased level of explicitness compared to non-mediated and non-native language use. However, they differ in the distribution of specific cohesive devices. Furthermore, based on the regression analysis, the author concludes that individual speaker effects override these other effects in the distribution patterns of cohesive devices in spoken and written registers of constrained and non-constrained language varieties.

These multivariate studies attempt to disentangle the possible factors contributing to the specific linguistic patterns of constrained (including mediated) language varieties. Thought-provoking as they are, they have also tended to focus on the distribution patterns of a small number of linguistic features (such as optional 'that' or cohesive devices), albeit from a multifactorial perspective. In this study, we make use of corpus-linguistic methods and the MD approach (Biber 1986, 1988) to investigate the nature of SI in relation to spontaneous NS and WT based on the distribution patterns of seventy-nine lexico-grammatical features.

3. Methodology

We aim to explore the multidimensional nature of SI in relation to NS and WT by incorporating corpus-linguistic methods with Biber's (1986, 1988) MD analysis approach. By doing so, we hope to isolate possible patterns in language dimension realizations specific to SI as both *spoken* and *mediated* discourse. Specifically, we focus on interpreting into a B language, which is the mainstream practice in the Chinese context (Chinese–English, by native Chinese speakers). However, we are

also fully aware that this may add further complexity to the interpretations of the linguistic manifestations of the interpreted language, and we will be particularly cautious when interpreting our results.

Overall, we aim to address the following four questions:

1. How many dimensions can be extracted to isolate the three language varieties under discussion; that is, SI, NS, and WT?
2. How can these dimensions be interpreted in functional terms?
3. What are the differences and/or similarities among the three language varieties along different dimensions?
4. What implications do they have for the spoken or translated/mediated nature of SI under discussion?

3.1 Introducing the LegCo+ corpus

Our research data are drawn from a newly built comparable intermodal corpus named the LegCo+ (see Table 1; Xu 2021). It contains four sub-corpora: Cantonese speeches as source texts (ST), simultaneously interpreted speeches into English (SI), written translation into English (WT), and English native speeches (NS), from two legislative settings — the Legislative Council of Hong Kong and the UK Parliament. These sub-corpora cover three genres labeled as Genre A, Genre B, and Genre C (see Table 1). They were chosen based on the primary roles shared by Members of Parliament (MPs) in the two legislative settings — specifically the oversight of governmental actions. However, our current investigation focuses solely on the analysis of three English subcorpora (SI, WT, and NS) with different mediation statuses. The primary objective is to examine the overall patterns of linguistic variation within these subcorpora, with particular emphasis on understanding the linguistic traits of SI when interpreting into a B language.

In terms of the mediated language varieties of SI and WT, both were produced from the same Cantonese speeches (ST), but under different conditions: while SI was conducted by interpreters recruited from the Official Languages Division (OLD) at the Civil Service Bureau of the Government of Hong Kong Special Administrative Region of the People's Republic of China, translations were carried out by in-house translators of the Legislative Council of Hong Kong (or HK LegCo). Consequently, it can be assumed that translations and interpretations were conducted independently from each other. Nevertheless, it is important to note that SI in Chinese institutional settings, such as the HK LegCo or the United Nations General Assembly for Chinese speakers, differs substantially from the interpreting practices in other contexts outside of China, such as the European Parliament. In these settings, "many oral statements in Chinese and their SI

renditions are based on written scripts," and if the Chinese interpreters receive these scripts in advance, most of them "will translate [the scripts] into English to be read out as the SI renditions" (Wu, Cheung, and Xing 2021,503). This is occasionally the case in the interpreting practice at the HK LegCo, where interpreters may receive pre-prepared translations and read them out when the speakers deliver their speeches. In our SI and WT datasets, we have observed some overlap between the two, as certain parts of the orthographic transcriptions from SI are identical to WT. To determine the extent to which each subset contributes to this overlap, we performed a manual comparison between the SI and WT subsets. Our analysis revealed that SI_B (representing Genre B of the SI dataset) exclusively consists of scripted interpretations (i.e., interpretations that are translations read out by simultaneous interpreters), accounting for 7.2% of the whole SI subcorpus. We believe that this influence is not very significant. However, we acknowledge that this observation introduces some complexity to the interpretation of linguistic patterns in the interpreted language, thereby necessitating caution in the discussion of our findings.

To mitigate the possible influence of "linguistic convergence" within the same discourse community (Defrancq 2018), we incorporated a NS subset from a distinct setting — the UK Parliament. However, the inclusion of this subset introduces additional intricacies to the interpretation of patterns of linguistic variation due to cultural and procedural differences. We will take account of this aspect in our subsequent analysis and discussion.

Table 1. Overview 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 the Prime Minster/Chief Executive Genre B: Questions to the ministers/secretaries Genre C: Debates on motions and bills			
Setting	The HK LegCo			The UK Parliament
Participants and power relations	Chief Executive, President, legislative members (including secretaries) representing different political parties			Prime Minister, Speaker, MPs (including ministers) representing different political parties
Time period	2015–2017			

3.2 The MD analysis approach

One prominent merit of an MD analysis lies in its capacity to reduce a large number of observed variables into several underlying constructs known as factors, or dimensions in functional terms, based on the idea that "statistical linguistic co-occurrence patterns reflect underlying shared communicative functions" (Biber 1988, 332). The term 'multidimensional' in Biber (1992) assumes that "multiple parameters of variation will be operative in any discourse domain" (385). Since translation and interpreting have been increasingly acknowledged as multifaceted activities (Evert and Neumann 2017; De Sutter and Lefer 2020), a multidimensional perspective is conducive to uncovering the underlying constraints of mediated activities.

Biber (1988, 63–64) outlines three methodological steps for an MD analysis. The first step deals with the selection, retrieval, and standardization of linguistic features to be investigated. This is followed by two quantitative steps: factor analysis and calculation of factor scores. Factor analysis reduces 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 text" (64). Then, based on the communicative functions most widely shared by these co-occurring linguistic features, textual dimensions are specified. The final step identifies factor scores with the operational representation of textual dimensions, which allows a comparison of genre variation.

In the current study, seventy-nine linguistic features (see Table A1 in the Addendum) were investigated, including the original sixty-seven in Biber's model (1988), and twelve more based on previous research, as reviewed in Section 2. After the selection, the linguistic features were tagged and retrieved using the Multidimensional Analysis Tagger (Nini 2014), a program designed to replicate "Biber's (1988) tagger for the multidimensional functional analysis of English texts" (Multidimensional Analysis Tagger (v.1.2) — Manual, 1), and WordSmith Tools v.6 (Scott 2012). Prior to statistical analysis, the raw frequencies of these features were normalized for each text segment[1] to frequency per 100 words.

A new factor analysis was then carried out on the frequency scores utilizing IBM SPSS 20. The analysis yielded a twenty-factor solution based on the shared variance among the variables (i.e., 79 linguistic features), but we decided to keep seven based on the scree plot which showed a considerable drop in eigenvalues — standardized measures of the amount of variance accounted for by a given factor

1. Given the uneven duration of each proceeding session, transcriptions of the three genres were segmented into 477 text segments altogether, with the total running words ranging between 1300 to 2000 per text segment.

— between Factor 4 and 5, and also a flattening of the line of the contribution of eigenvalues after the first seven dimensions. A follow-up rotated factor extraction was done to determine the total variance explained by the seven factors, each having several linguistic features with a factor loading larger than 0.30 (a statistically significant cut-off) loaded on it. This was then followed by interpretations of factors as textual dimensions based on the communicative functions shared by most of the co-occurring linguistic features.

In order to establish the extent of the possible influence of ontology (translated or not) and modality (spoken or written), we calculated the factor scores for each dimension. Computation of factor scores was done by summing the number of the standardized scores of linguistic features having salient loadings larger than 0.30 on that dimension (Biber 1988). In cases where a linguistic feature loaded onto multiple dimensions, we adopted Hu, Xiao, and Hardie's (2016) methodology instead of Biber's original approach. Following this approach, every linguistic feature was taken into consideration as long as its factor loading was equal to or larger than 0.30. This decision was made because we concur with Hu, Xiao, and Hardie's (2016) assertion that it is "unwise to prejudge the issue of whether or not a feature might be important to more than one dimension" (20). For a detailed overview of the descriptive dimension scores for NS, SI, and WT, see Table A2 in the Addendum.

4. Findings

4.1 Dimensions extracted from factor analysis

4.1.1 *Dimension 1: Involved versus Informational Production*

Linguistic features loading onto this dimension (see Table 2) exhibit considerable overlap with features that also load on Biber's (1988) Dimension 1. Specifically, linguistic features with positive loadings are associated with affective, involved, fragmented, and informal language use produced under high cognitive and time constraints (e.g., contractions, shorter words, demonstratives, independent clause coordination). By contrast, the negative loadings indicate a high information density and "a careful integration of information in a text" (Biber 1988, 104) that is not subject to temporal constraints (e.g., longer words, total prepositional phrases, attributive adjectives, and nouns). As a result, we have adopted Biber's (1988) original designation and label Dimension 1 as 'Involved versus Informational Production'. Examples (1), (2) and (3)[2] are extracts from the SI and WT

2. For all the examples illustrated in the article, we provide a clean version; that is, we delete the paralinguistic features annotated previously based on the transcription rules.

subsets of Genre A, namely, Questions to the Prime Minister/Chief Executive. These examples exhibit high and low scores on this dimension. Positive features are italicized, while negative features are underlined.

(1) *We* should listen to the views of all sectors of society. *We have* representatives from the employers, employees, scholars, experts, *etc.*. *We'll* listen to their advice. (D1_02)

(2) Now *you* probably *haven't* put a question here. I*'d* like to say *that* the hands-on approach for officials apply to to all kinds of work. (D6_05)

(3) Therefore, regarding how standard working hours should be formulated, we ought to listen to views of all walks of life in society, including representatives of employers and employees in the SWHC, community leaders and experts in the academia. We should listen to more views. (G1_03)

Table 2. Linguistic features loading onto Dimension 1

Positive features		**Negative features**	
Present tense	.792	Average word length	−.897
Contractions	.787	Longer words	−.831
Shorter words	.731	Average sentence length	−.706
BE as main verb	.697	Total prepositional phrases	−.703
Discourse particles	.684	Total other nouns	−.623
Demonstrative pronouns	.649	Attributive adjectives	−.608
First person pronouns	.645	Past participial WHIZ deletion relatives	−.581
Second person pronouns	.607	Phrasal coordination	−.571
Pro-verb DO	.571	Nominalizations	−.539
Independent clause coordination	.543	Present participial clauses	−.529
Analytic negation	.537	Conjuncts	−.456
Subordinator *that* deletion	.513	Lexical density	−.434
Pronoun *it*	.500	Determiner *the*	−.402
Emphatics	.479	Split auxiliaries	−.317
Existential *there*	.429		
Causative adverbial subordinators	.426		
Wh-pronouns	.417		
Hedges	.408		
Predicative adjectives	.389		
Wh-clauses	.383		

Table 2. *(continued)*

Positive features		Negative features
Public verbs	.379	
Demonstratives	.324	
Total adverbs	.304	

4.1.2 *Dimension 2: On-line Information Elaboration with Stancetaking Concerns*

Linguistic features loading onto Dimension 2 (see Table 3) include another subset of features categorized in Biber's (1988) Dimension 6 (On-line Information Elaboration), which signals "informal, unplanned types of discourse" (113). These features encompass subordination structures, such as *that* relative clauses on object position and *that* clauses as adjective complements, as well as demonstratives. The nature of Dimension 6 in Biber's (1988) model has been elaborated by Van Rooy et al. (2010), who assert 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. (343)

Hence, in contrast to the focus on informational production in Dimension 1, Dimension 2 captures the tension between information density and on-line production strain. An additional significant function shared by these co-occurring features is the expression of personal stance. Biber (1988) notes that "discourse tasks which involve the explicit marking of an individual's stance are frequently also tasks that demand informational production under real-time constraints" (160). Conversely, features with negative loadings convey an objective sense of meaning. Therefore, we designate Dimension 2 as 'On-line Information Elaboration with Stancetaking Concerns', as exemplified in Example (4) and (5),[3] respectively (with positive features italicized and negative features underlined).

(4) Does *my* right honorable Friend share *my* concern *that*, if the other place were to vote against working tax credits, *this* would be a serious challenge to the privilege of this House, a privilege codified as long ago as sixteen seventy-eight? (A4_04)

3. Example (4) is an excerpt taken from the NS subset of Genre A, while Example (5) contains extracts from the SI subset of Genre B.

(5) If an applicant comes to work in Hong Kong, of course we require written employment contracts, setting out in detail the duties, conditions of work, remuneration. (E3_06)
It's just a couple of examples of cited out of the many different measures to verify whether there is a genuine need, and whether that person is actually working in that company. (E3_06)

Table 3. Linguistic features loading onto Dimension 2

Positive features		**Negative features**	
That relative clauses on object position	.647	Conditional adverbial subordinators	–.437
Amplifiers	.613	Analytic negation	–.436
That relative clauses on subject position	.574	Second person pronouns	–.322
First person pronouns	.524	Existential *there*	–.310
That adjective complements	.486		
Demonstratives	.439		
Private verbs	.431		
Sentence relatives	.431		
Standardized type-token ratio	.375		
Wh-pronouns	.355		
Place adverbials	.323		

4.1.3 *Dimension 3: Precise versus Simplified Description*

Dimension 3 (see Table 4) captures positive features associated with elaborate, precise and explicit description, such as split auxiliaries, conjuncts, total adverbs, downtoners, and time adverbials. Some of these features overlap with Biber's (1988) Dimension 7 labeled as 'Academic Qualification or Hedging', and include total adverbs, downtoners and concessive adverbial subordinators that are used to "qualify the extent to which an assertion is 'known' in academic discourse" (114). Although Biber eventually discarded Dimension 7 due to small factor loadings of the three co-occurring features, in our study, these features have relatively high weights. Together with other positive features, they signal precise and elaborate description. By contrast, the negative weights indicate simplified, informal, and repetitive language use. Example (6) and (7)[4] illustrate this distinction.

4. Example (6) and (7) are taken from the WT subset of Genre A and the SI subset of Genre B, respectively.

(6) I *have already* answered Dr Leung Ka-lau's question *just now*. (G3_09)
In the *past*, I *have repeatedly* stressed in public *that* one of the important social benefits of tourism to Hong Kong was the provision of employment opportunities for the large working population in Hong Kong. (G3_09)

(7) And also it's not the case that we would implement the the programme after after the discussion. (E2_04)
Well this is an important point, as I said we'll get in touch with schools from time to time. (E2_04)

Table 4. Linguistic features loading onto Dimension 3

Positive features		**Negative features**	
Split auxiliaries	.590	Contractions	–.408
Conjuncts	.489	Top-10 coverage	–.321
Total adverbs	.485	Independent clause coordination	–.315
Downtoners	.462		
Time adverbials	.419		
Concessive adverbial subordinators	.359		
Perfect aspect	.353		
Synthetic negation	.339		
That verb complements	.309		

4.1.4 *Dimension 4: Narrative versus Abstract Focus*

Features with positive loadings on Dimension 4 (see Table 5) seem to prioritize narrative focus, as indicated by the co-occurrence of third person pronouns, *wh*-pronouns, and *wh*-relative clauses on subject position. These features are used to index referents that may not be in the immediate co-text or context. On the contrary, the negative features on this dimension are all nominal features that exhibit high information density. However, only two out of the five negative features, namely, nominalizations and gerunds, have relatively larger factor loadings. According to Biber (1988), when interpreting factors as dimensions, greater attention should be given to features with the largest loadings. In our study, this pertains specifically to nominalizations and gerunds, which have been regarded as "markers of conceptual abstractness" (225). Therefore, we refer to Dimension 4 as 'Narrative versus Abstract Focus'. Example (8), for instance, demonstrates more frequent usage of third person pronouns indicating a narrative perspective, while

Example (9)[5] is characterized by the co-occurrence of nominal features, such as nominalizations and gerunds, which shows a high level of abstraction.

(8) Otherwise there will *be* people *who* have died we'll never know about, and too many people *who* need help *who* will not get *it*. (B6_04)
Well I thank the right honorable Lady for that. *She is* right. (B6_04)

(9) Two, whether the authorities have grasped the latest situation of Mainland residents obtaining approval to come to work in Hong Kong by falsely claiming that they have already got an offer of employment from a Hong Kong employer or by other illegal means. (H3_05)

Table 5. Linguistic features loading onto Dimension 4

Positive features		**Negative features**	
Third person pronouns	.533	Nominalizations	–.466
Wh-pronouns	.477	Gerunds	–.408
Wh-relative clauses on subject position	.377	Lexical density	–.319
BE as main verb	.372	Prepositions or subordinating conjunctions	–.313
		Longer words	–.302

4.1.5 *Dimension 5: Lexical versus Functional Concerns*

Only four features loaded on Dimension 5 (see Table 6), and the communicative functions associated with them are relatively straightforward. Lexical density cues a lexical focus, while the features with negative loadings are consistently related to functional or generic usage. As a result, we interpret Dimension 5 as 'Lexical versus Functional Concerns'. However, due to the 'abstract' nature of the linguistic features presented in Table 6, we refrain from providing specific examples for this particular dimension.

Table 6. Linguistic features loading onto Dimension 5

Positive features		**Negative features**	
Lexical density	.460	Top-10 coverage	–.776
		Determiner *the*	–.669
		Shorter words	–.485

5. Example (8) and (9) are taken from the NS and WT subsets of Genre B, respectively.

4.1.6 *Dimension 6: On-line Persuasion*

The positive features on Dimension 6 (see Table 7) appear to capture a widely discussed lexical pattern in translation, specifically the optional usage of *that* after reporting verbs. However, a closer examination reveals that verb complements with *that* have much lower loadings than suasive verbs and public verbs, which are frequently used to imply intentions to cause specific future events (Biber 1988). Predictive modals are also closely linked to future predictions and are employed to persuade the audience or readers by predicting potential outcomes in the future. Additionally, the negative weight of standardized type-token ratio (STTR) may suggest that these persuasive strategies are employed under time constraints. Therefore, Dimension 6 is tentatively labeled 'On-line Persuasion'. Example (10) and (11)[6] illustrate this dimension.

(10) That's why I *believe* my first Policy Address is in line with public expectations and I *hope that* we could take forward those issues as soon as possible. (D7_01)

(11) Well I for one *believe that* every life is equal. (F6_04)

Table 7. Linguistic features loading onto Dimension 6

Positive features		**Negative features**	
Suasive verbs	.588	Standardized type-token ratio	−.403
Public verbs	.434		
Predictive modals	.341		
That verb complements	.341		

4.1.7 *Dimension 7: Coordinating versus Possessive Functions*

The last dimension (see Table 8) captures four co-occurring features, but their factor loadings are relatively small, with absolute values equal to or lower than 0.5. The presence of coordinating conjunctions, such as *and*, *but*, *so*, *either*, and *or*, indicates a focus on coordinating or conjoining two sentences, phrases or words, thereby indicating greater usage of coordinated language use. In contrast, the features with negative loadings signify a more concise representation of information, emphasizing possessive relationships. This is exemplified by phrases like "people's income," "patients' conditions," and "a person's talent." Therefore, we name Dimension 7 'Coordinating versus Possessive Functions'.

6. Example (10) and (11) are extracted from the SI subset of Genre A and Genre C, respectively.

Table 8. Linguistic features loading onto Dimension 7

Positive features		Negative features	
Coordinating conjunctions	.352	Total other nouns	−.507
		Possessive endings	−.361
		Lexical density	−.340

4.2 Language varieties across dimensions

Following the calculation of dimension scores described in Section 3.2, and considering the functional descriptions of the dimensions presented in Section 4.1, this section explores the potential similarities and differences among the three language varieties (NS, SI, and WT) across the seven dimensions. The focus is specifically on linguistic variations observed in the SI language variety. Three key aspects need to be borne in mind when interpreting linguistic variation patterns, as suggested by Biber (1988,129): (1) the similarities and differences in the mean dimension scores among language varieties; (2) the co-occurring features within the dimension being discussed, including both features with positive weights and negative weights; and (3) the communicative functions these features serve.

Figure 1 presents the mean scores of all seven dimensions for the language varieties NS, SI, and WT, using a heatmap generated with Python. A darker color indicates a higher dimension score, while the color contrast within each dimension signifies the level of similarity among the three language varieties. A smaller color contrast suggests a higher level of similarity, and vice versa. The mixed linguistic production of SI as *spoken* and *mediated* discourse is clearly evident: across Dimension 1, Dimension 3, and Dimension 7, SI displays greater resemblance to NS, potentially due to a shared modality; in contrast, it exhibits more similarities with WT across Dimension 2, Dimension 4, and Dimension 5, possibly owing to their shared mediation status. Dimension 6 presents a slightly different scenario where the three language varieties are relatively indistinguishable from one another. This is reflected in their intermediate dimension scores, typically falling within an absolute value smaller than or approximately equal to 0.5. In the following sections, we will examine closely the overall influence of ontology versus modality on the patterns of linguistic variation specific to SI.

4.2.1 *SI — More spoken than translated/mediated?*

As summarized in Section 4.2, three dimensions reveal shared similarities between SI and NS to varying degrees: Dimension 1 (Involved vs. Informational Production), Dimension 3 (Precise vs. Simplified Description), and Dimension 7 (Coordinating vs. Possessive Functions). Specifically, Dimension 1 suggests a less-

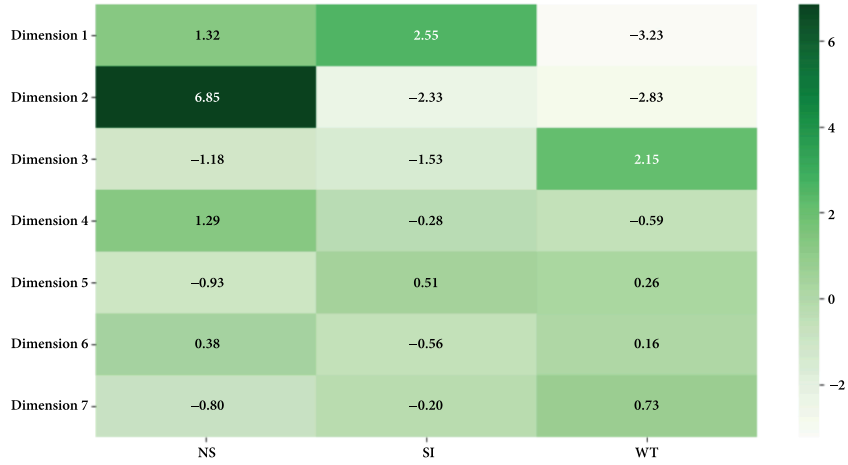

Figure 1. Mean dimension scores[7] for the three language varieties across the seven dimensions

er resemblance between SI and NS (mean score difference = 1.23) compared to the other two dimensions. Dimension 3 and 7 demonstrate a closer similarity, with mean score differences of 0.35 and 0.6, respectively.

Starting with Dimension 1, a high positive score in this dimension indicates involved and informal language use produced 'on-line', characterized by the presence of co-occurring features such as contractions, personal pronouns (first and second), discourse particles, and demonstrative pronouns, among others. By contrast, a high negative score indexes informative and carefully integrated language, marked by frequent co-occurrences of longer words, prepositional phrases, total other nouns and adjectives, to name a few. Figure 1 shows that both SI and NS exhibit a tendency towards 'involved production', which reflects the typical features of spoken language, irrespective of their mediation status. In comparison, the mediated written language (WT) is positioned more towards the 'informational' end of the continuum. However, SI demonstrates an even greater inclination towards 'involvedness', as evident in its higher dimension score. This finding lends some support to the lexical simplification hypothesis, as reviewed in Section 2.1 (Hu and Tao 2009; Kajzer-Wietrzny 2012, 2015; Li and Wang 2012). To illustrate this distinction, Example (12) shows the interpretation (F6_01) and translation (I6_01) of the same source speech.

7. A high positive score indicates that a language variety is marked by a frequent use of co-occurring features with positive weights, and infrequent use of those with negative weights; by contrast, a high negative score indicates that a language variety is marked by a frequent use of co-occurring features with negative weights, and infrequent use of positive ones.

(12) Animals may not understand what *I'm* saying, *but it* proves the famous quote of Gandhi, "the the civilization of a society depends on how people treat the animals." And nowadays, in Hong Kong, in twenty seventeen, many Hong Kong people *are* willing to listen to the needs of earth and the voice of animals because lives *are* valuable *and we* should protect their lives *and* should *and we* should not pleasure ourselves by killing animals. *We* have had enough of these news reported. Why *can't* the government enact legislation in a comprehensive manner to protect animals? (F6_01)
I know that animals may not understand this motion that I propose. However, my motion has also borne testimony to a famous saying of Mr Mahatma GANDHI: "The degree to which a city is civilized depends on the way how the people in the city treat animals." In Hong Kong nowadays, many local people are willing to squat down to listen to the thoughts of animals because all lives are precious and there should be no distinction between human and animal lives. It is improper to deprive animals of their right to live, not to mention taking pleasure in abusing and killing them. I have had enough of the news on animal abuse. Why can't the Hong Kong Government learn a lesson from the bitter past experience and truly draw up a holistic set of legislation for this group of silent, disadvantaged animals? (I6_01)

In comparison, Dimension 3 exhibits greater similarities between mediated and non-mediated spoken language (SI and NS). The co-occurring features with positive loadings on Dimension 3 are indicative of more elaborate and precise description, while the negative features are characteristic of spoken language and include contractions, top-10 vocabulary coverage and independent clause coordination, marking more simplified language use. Both SI and NS exhibit simplified language use in this dimension, as reflected in their mean dimension scores of −1.18 and −1.53, respectively. Neither group shows a preference for more precise and elaborated description. In contrast, WT demonstrates an opposite trend in terms of features related to precision and elaboration. This clear modality difference between interpreted texts and translated texts is evident in Dimension 3. In fact, Dimension 3 captures the second largest difference between SI and WT, with a mean dimension score difference of 3.69, the first being Dimension 1 (with a difference of 5.78).

Dimension 7 also suggests some similarities between SI and NS, but it is not as distinct as Dimension 1 and Dimension 3, as indicated by the relatively small dimension scores. Dimension 7 isolates linguistic co-occurrence patterns associated with coordinating functions (e.g., coordinating conjunctions) from possessive functions (e.g., total other nouns, possessive endings, and lexical density). In this dimension, SI is unmarked, meaning it does not display a clear preference for either coordinating or possessive functions (mean dimension score = −0.2). On

the other hand, the NS subset shows a slight preference for possessive functions (mean dimension score = −0.8). In WT, there is a slight preference for coordinating functions.

4.2.2 *SI — More translated/mediated than spoken?*

In relation to their shared mediation status, there is evidence of similarities between SI and WT in Dimension 2 (On-line Information Elaboration with Stancetaking Concerns), Dimension 4 (Narrative vs. Abstract Focus), and Dimension 5 (Lexical vs. Functional Concerns). The extent of similarity is indexed mostly by the differences in the dimension scores for the varieties.

Dimension 2 reveals interesting patterns of linguistic variation, where SI and WT display a considerable level of homogeneity (mean score difference = 0.5) despite their modality differences, while remarkable ontological differences are observed between SI and NS (mean score difference = 9.18). Specifically, both SI and WT, as mediated language varieties, are highly unmarked in the co-occurrence of features associated with stancetaking expressions produced under time constraints (e.g., *that* relative clauses on object position, *that* relative clauses on subject position, *that* adjective complements), and features highlighting immediate referents (e.g., first person pronouns, demonstratives, and private verbs), and marked in the co-occurrence of conditional adverbial subordinators and analytic negation, indicative of an objective language use. By contrast, non-mediated NS shows a very high degree of markedness in on-line expression of personal stance. Section 5 will explore the possible reasons for these dramatic differences.

Dimension 4 and Dimension 5 also appear to distinguish mediated language varieties (SI and WT) from the non-mediated variety (NS), but it is unclear whether the observed patterns in these two dimensions are specific to mediation. Dimension 4, labeled 'Narrative versus Abstract Focus', reveals that both SI and WT display relatively neutral characteristics, as indicated by their intermediate dimension scores (Biber 1988). In other words, they do not lean towards either a narrative or abstract focus. By contrast, the NS data show a clear preference for linguistic features associated with narration, such as the co-occurrence of third person pronouns, *wh*-pronouns, and *wh*-relative clauses, while being unmarked in abstract expression. However, this distinction is more likely attributed to procedural differences rather than inherent ontological distinctions, especially concerning the use of third person pronouns, as will be further elaborated in Section 5.

Likewise, Dimension 5 captures potential mediation-specific patterns, in that both translation and interpreting demonstrate a lack of markedness in either lexical or functional language usage. In comparison, the NS subset exhibits a moderately high dimension score (−0.93), which suggests a focus on functional aspects

realized through the more frequent co-occurrence of top-10 vocabulary coverage, the use of determiner *the*, and shorter words (shorter than 3 letters), while not displaying markedness in lexical density.

5. Discussion

Based on the above results, it appears that English retour SI from Cantonese is a hybrid language mode, exhibiting features that are somewhat between original non-interpreted native speeches and WT. Being forms of both *spoken* and *translated* discourse, SI demonstrates typical features of spoken language, as observed in Dimension 1, Dimension 3, and Dimension 7; meanwhile, it also exhibits certain co-occurrence patterns that are commonly found in mediated written language, as witnessed across Dimension 2, Dimension 4, and Dimension 5. We did not include Dimension 6 in our interpretation, given that the three language varieties are relatively difficult to differentiate on this dimension.

Beginning with the divergences between SI and WT, which share a similar ontology but differ in modality, our intuitive knowledge, along with previous intermodal studies (Shlesinger 2008; Shlesinger and Ordan 2012; Kajzer-Wietrzny 2015), leads us to anticipate noticeable differences between interpreted language and translated language. This expectation is supported by our MD analysis, specifically with respect to Dimension 1 (Involved vs. Informational Production), Dimension 3 (Precise vs. Simplified Description), and Dimension 7 (Coordinating vs. Possessive Functions). In these dimensions, SI exhibits linguistic features typically associated with spoken language, indexing involved, fragmented, simplified, and informal language use (Chafe 1982; Chafe and Danielewicz 1987). Consequently, it can be inferred that SI leans more towards *spoken* language than translated language. This suggests that the influence of modality (spoken or written) may outweigh that of ontology (translated or not), thus validating the findings of Shlesinger and Ordan (2012). Additionally, based on mean dimension score differences, Dimension 1 and Dimension 3 (with values of 5.78 and 3.68, respectively) are prominent discriminators between SI and WT, further supporting the notion that modality is a stronger predictor in distinguishing interpreting from translation.

However, despite the differences in modality, there are notable similarities between SI and WT as forms of mediated language that should not be overlooked, as reported in Section 4.2.2. Notably, Dimension 2 (On-line Information Elaboration with Stancetaking Concerns) emerges as the strongest discriminator among all seven dimensions (with a mean dimension score difference of 9.18 between SI and NS). Our findings reveal that both translated and interpreted texts exhibit a

significant underuse of co-occurrence features associated with stancetaking linguistic practices, whereas native spoken texts display a high prevalence of stancetaking expressions. We argue that there are two potential contributors to such dramatic differences. Firstly, the eschewal of stancetaking expressions in the SI and WT data could be attributed to risk-avoidance strategies employed by translators and interpreters (Pym 2015). This aligns with the findings reported in Kruger and Van Rooy (2016), where both translated and non-native varieties of English exhibit a tendency to avoid employing stancetaking expressions in favor of maintaining an objective style. Another possible explanation could be the existence of different rules of procedures for Members of the UK Parliament and the HK LegCo in terms of explicitly expressing stance, but this information is not clear. A parallel analysis comparing stance markers in Cantonese source speeches with translated and interpreted speeches could be conducted to test this assumption, which falls outside the scope of our study.

In addition to Dimension 2, Dimension 4 (Narrative vs. Abstract Focus) and Dimension 5 (Lexical vs. Functional Concerns) also reveal potential mediation-specific patterns, as discussed in Section 4.2.2. However, these patterns are likely attributable to other influencing factors rather than the effect of mediation. One possible factor is the procedural differences between the two legislative settings. This is particularly evident in Dimension 4, where NS demonstrates a preference for a more prominent and marked narrative focus compared to SI (as shown in Example (8) to (10)). According to Ilie (2015), the default address in the UK Parliament is third person pronouns. This form of address demonstrates deference and creates a distance between MPs, allowing them "to make straightforward and forceful statements in their interactions with fellow MPs" while maintaining "a safe institutional distance from one another" (9). In contrast, in the HK LegCo, MPs often directly address the speakers using second person pronouns (*you*).

Another potential factor that could contribute to the observed patterns is the influence of a mixed mode of delivery in the SI outputs. This is particularly evident in Dimension 5, where SI exhibits a closer resemblance to WT (with a mean dimension score difference of 0.25), while significantly differing from NS (with a mean dimension score difference of 1.44). As explained in Section 3.1, we found that 7.2% of SI consisted of scripted interpretations, which are pre-prepared translations read out by interpreters when the source speakers deliver their speeches. Although this proportion is relatively small, their presence could have contributed to balancing out the markedness in either direction of the co-occurrence patterns of SI in Dimension 5. As a result, this may have led to a greater resemblance between SI and WT in this dimension.

One final interpretation that needs particular attention is for Dimension 1 (Involved vs. Informational Production). As previously discussed, this dimension

demonstrates the second largest disparity between SI and WT as distinct (mediated) modalities, while SI shares many co-occurrence features that are typical of spoken language with NS. However, we have also observed discrepancies between the two forms of spoken language, as SI exhibits more pronounced patterns of involved, fragmented, and informal language production compared to NS. This finding provides support to the lexical simplification hypothesis (Laviosa 1998; Bernardini, Ferraresi, and Miličević 2016; Ferraresi et al. 2018), and additionally demonstrates that the presence of scripted interpretations in the SI outputs does not significantly impact the overall variation patterns. To explain the more 'involved' nature of SI, taking into account the significant constraints (such as linear, cognitive, and socio-cultural) faced by simultaneous interpreters (Shlesinger 1995; Gumul 2006), we argue that the practice of interpreting into a B language can further accentuate this pattern. Interpreting into a B language, or retour interpreting, is widely recognized as being more cognitively demanding than native interpreting (Chang and Schallert 2007; Gile 2009). Consequently, interpreters may unconsciously simplify their language usage by employing certain interpreting strategies (such as generalization or even omission) to alleviate cognitive load, which ultimately contributes to the more 'involved' (Dimension 1) and 'simplified' (Dimension 3) patterns as observed in the interpreted outputs.

In general, there are both similarities and discrepancies between SI and NS, as well as between SI and WT. The impact of ontology (translated or not) on the linguistic characteristics of interpreted language does not consistently outweigh that of modality (spoken or written), and vice versa, across the seven dimensions examined. Interestingly, the influence of ontology does not necessarily distinguish SI from NS in terms of the well-established translation universals of explicitation and normalization, with the exception of lexical simplification. This is due to the fact that stancetaking emerges as the most prominent factor that sets mediated and non-mediated language varieties apart. On the other hand, the role of modality does appear to differentiate spoken language from written language, as evident in Dimension 1, Dimension 3, and Dimension 7. Overall, when adopting a multidimensional perspective, interpreted language tends to exhibit a hybrid nature that lies between non-translated spoken originals and written translations.

6. Summary and conclusion

The analysis presented in this article yielded three key findings. Firstly, it identified seven dimensions based on seventy-nine linguistic features. Overall, SI was found to exhibit a hybrid language mode, displaying features that fall between those of NS and WT across different dimensions. Contrary to our initial expec-

tations, the second finding revealed similarities between interpreting and translation, suggesting that these modes of mediation may be less distinct than anticipated. This was particularly evident in Dimension 2 (On-line Information Elaboration with Stancetaking Concerns), indicating potential shared co-occurrence patterns of mediated (and constrained) language varieties. Thirdly, in dimensions associated with oral features, SI displayed a greater resemblance to NS than WT. This suggests a strong modality effect, implying that interpreting can be viewed as an extreme case of both spoken and translated/mediated discourse.

The findings resulting from our corpus-based MD analysis further challenge the intuitive belief that SI and WT must differ due to their distinct modalities, irrespective of their shared ontological/mediated status. Our overall observation of the high resemblance between SI and WT (with similar ontology yet different modality) across multiple dimensions lends support to the perspective that mediated texts may possess "universal features that set them apart from original texts" (Shlesinger and Olohan 2012, 55). In our study, these features or patterns are particularly evident in the avoidance of stancetaking expressions by both translators and interpreters. They also provide insights into the potential shared norms and constraints encountered by translators and interpreters during translation, factors that are important in descriptive analysis. However, further well-designed research should be conducted to determine whether these shared patterns are truly specific to mediation.

Despite the mixed results, the fundamental aim of the MD analysis is to uncover the intricate nature of language in use and to increase awareness about the "processes (linguistic, pragmatic, practical or cognitive) [...] engaged during an interpretation" (Cencini 2002, 1). Our findings suggest that the influence of ontology versus modality on the linguistic manifestations of interpreted outputs is not always straightforward. This complexity arises because interpreters (and translators) face additional constraints that go beyond the direct influence of these two factors. These factors include cultural and procedural differences between legislative settings, the working direction (in our case, interpreting into a B language), and possibly the mixed delivery modes of SI, among others. By examining the similarities and differences among the three language varieties, we aim to enhance our understanding of the very nature of mediated language and its relationship to other language varieties. After all, as Chesterman (2017, 264) points out, "we can currently do little more than speculate as rationally as possible," which is also the case with interpreting the linguistic patterns we have observed.

One limitation of the current study is the inclusion of a small portion of scripted interpretations in our SI Genre B dataset, which deviates from the *de facto* interpreting often examined in the European contexts. However, it is important to note that in Chinese contexts it is common practice for interpreters to

occasionally read out pre-prepared translations. Moreover, this inclusion does not have a significant impact on the overall patterns of linguistic variation reported in this study, such as in Dimension 1 and Dimension 3. Nevertheless, we acknowledge the importance of designing additional detailed research that specifically isolates and examines different modes of delivery to effectively address this issue.

References

Agrifoglio, Marjorie. 2004. "Sight Translation and Interpreting: A Comparative Analysis of Constraints and Failures." *Interpreting* 6 (1): 43–67.

Avner, Ehud Alexander, Noam Ordan, and Shuly Wintner. 2016. "Identifying Translationese at the Word and Sub-word Level." *Digital Scholarship in the Humanities* 31 (1): 30–54.

Baker, Mona. 1993. "Corpus Linguistics and Translation Studies: Implications and Applications." In *Text and Technology: In Honour of John Sinclair*, edited by Mona Baker, Gill Francis, and Elena Tognini-Bonelli, 233–250. Amsterdam: John Benjamins.

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

Baker, Mona. 1996. "Corpus-based Translation Studies: The Challenges That Lie Ahead." In *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, edited by Harold Somers, 175–186. Amsterdam: John Benjamins.

Baroni, Marco, and Silvia Bernardini. 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.

Bernardini, Silvia, Adriano Ferraresi, and Maja Miličević. 2016. "From EPIC to EPTIC: Exploring Simplification in Interpreting and Translation from an Intermodal Perspective." *Target* 28 (1): 61–86.

Biber, Douglas. 1986. "Spoken and Written Textual Dimensions in English: Resolving the Contradictory Findings." *Language* 62 (2): 384–414.

Biber, Douglas. 1988. *Variation Across Speech and Writing*. Cambridge: Cambridge University Press.

Biber, Douglas. 1992. "The Multidimensional Approach to Linguistic Analyses of Genre Variation: An Overview of Methodology and Findings." *Computers and the Humanities* 26 (5–6): 331–345.

Blum-Kulka, Shoshana, and Eddie A. Levenston. 1978. "Universals of Lexical Simplification." *Language Learning* 28 (2): 399–415.

Cencini, Marco. 2002. "On the Importance of an Encoding Standard for Corpus-based Interpreting Studies Extending the TEI Scheme." In *CULT2K*, edited by Silvia Bernardini and Federico Zanettin, special issue of *InTRAlinea*. https://www.intralinea.org/specials/article/1678

Chafe, Wallace, and Jane Danielewicz. 1987. "Properties of Spoken and Written Language." In *Comprehending Oral and Written Language*, edited by Rosalind Horowitz and S. Jay Samuels, 83–113. London: Academic Press.

Chafe, Wallace. 1982. "Integration and Involvement in Speaking, Writing, and Oral Literature." In *Spoken and Written Language: Exploring Orality and Literacy*, edited by Deborah Tannen, 35–54. Westport: Praeger.

Chang, Chia-chien, and Diane L. Schallert. 2007. "The Impact of Directionality on Chinese/English Simultaneous Interpreting." *Interpreting* 9 (2): 137–176.

Chen, Jiansheng, and Yani Cui. 2010. "A Corpus-based Study on Lexical Features in the English Translation of Report on the Work of the Government." *Contemporary Foreign Languages Studies* 6: 39–43.

Chesterman, Andrew. 2017. *Reflections on Translation Theory: Selected Papers 1993–2014*. Amsterdam: John Benjamins.

Dayter, Daria. 2018. "Describing Lexical Patterns in Simultaneously Interpreted Discourse in a Parallel Aligned Corpus of Russian-English Interpreting (SIREN)." *Forum* 16 (2): 241–264.

De Sutter, Gert, and Eline Vermeire. 2020. "Grammatical Optionality in Translations: A Multifactorial Corpus Analysis of *That*/Zero Alternation in English Using the MuPDAR Approach. In *New Empirical Perspectives on Translation and Interpreting*, edited by Lore Vandevoorde, Joke Daems, and Bart Defrancq, 24–51. London: Routledge.

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

Defrancq, Bart, Koen Plevoets, and Cédric Magnifico. 2015. "Connective Items in Interpreting and Translation: Where Do They Come From?" In *Yearbook of Corpus Linguistics and Pragmatics: Current Approaches to Discourse and Translation Studies*, edited by Jesús Romero-Trillo, 195–222. New York: Springer.

Defrancq, Bart. 2018. "The European Parliament as a Discourse Community: Its Role in Comparable Analyses of Data Drawn from Parallel Interpreting Corpora." *The Interpreters' Newsletter* 23: 115–132.

Evert, Stefan, and Stella Neumann. 2017. "The Impact of Translation Direction on Characteristics of Translated Texts: A Multivariate Analysis for English and German." In *Empirical Translation Studies: New Theoretical and Methodological Traditions*, edited by Gert de Sutter and Marie-Aude Lefer, 47–80. Berlin: De Gruyter Mouton.

Ferraresi, Adriano, Silvia Bernardini, Maja Petrović, and Marie-Aude Lefer. 2018. "Simplified or Not Simplified? The Different Guises of Mediated English at the European Parliament." *Meta* 63 (3): 717–738.

Gile, Daniel. 2001. "Consecutive vs. Simultaneous: Which is More Accurate?" *Interpretation Studies* 1 (1): 8–20.

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

Gumul, Ewa. 2006. "Explicitation in Simultaneous Interpreting: A Strategy or a By-product of Language Mediation?" *Across Languages and Cultures* 7 (2): 171–190.

Hu, Kaibao, and Qing Tao. 2009. "A Corpus-based Study of Explicitation of Textual Meaning in Chinese–English Conference Interpreting." *PLA International Studies University Journal* 32 (5): 67–73.

Hu, Xianyao, Richard Xiao, and Andrew Hardie. 2016. "How do English Translations Differ from Nontranslated English Writings? A Multi-feature Statistical Model for Linguistic Variation Analysis." *Corpus Linguistics and Linguistic Theory* 15 (2): 347–382.

Ilie, Cornelia. 2015. "Parliamentary Discourse". In *Parliamentary Discourse*, edited by Karen Tracy, 1–15. Hoboken, NJ: John Wiley & Sons.

Kajzer-Wietrzny, Marta, and Ilmari Ivaska. 2020. "A Multivariate Approach to Lexical Diversity in Constrained Language." *Across Languages and Cultures* 21 (2): 169–194.

Kajzer-Wietrzny, Marta, and Łukasz Grabowski. 2021. "Formulaicity in Constrained Communication: An Intermodal Approach." In *Reflexión crítica en los estudios de traducción basados en corpus / CTS Spring-cleaning: A Critical Reflection*, edited by María Calzada and Sara Laviosa, special issue of *MonTI* 13: 148–183.

Kajzer-Wietrzny, Marta. 2012. *Interpreting Universals and Interpreting Style*. PhD diss. Adam Mickiewicz University.

Kajzer-Wietrzny, Marta. 2015. "Simplification in Interpreting and Translation". *Across Languages and Cultures* 16 (2): 233–255.

Kajzer-Wietrzny, Marta. 2022. "An Intermodal Approach to Cohesion in Constrained and Unconstrained Language." *Target* 34 (1): 130–162.

Kotze, Haidee. 2022. "Translation as Constrained Communication: Principles, Concepts and Methods." In *Extending the Scope of Corpus-based Translation Studies*, edited by Sylviane Granger and Marie-Aude Lefer, 67–98. London: Bloomsbury.

Kotze, Haidee, and Bertus van Rooy. 2024. "Introduction: The Constrained Communication Framework for Studying Contact-influenced Varieties." In *Constraints on Language Variation and Change in Complex Multilingual Contact Settings*, edited by Bertus van Rooy and Haidee Kotze, 1–28. Amsterdam: John Benjamins.

Kruger, Haidee, and Bertus van Rooy. 2012. "Register and the Features of Translated Language." *Across Languages and Cultures* 13 (1): 33–65.

Kruger, Haidee, and Bertus van Rooy. 2016. "Constrained Language: A Multidimensional Analysis of Translated English and a Non-native Indigenised Variety of English." *English World-Wide* 37 (1): 26–57.

Kruger, Haidee, and Bertus van Rooy. 2018. "Register Variation in Written Contact Varieties of English: A Multidimensional Analysis." *English World-Wide* 39 (2): 214–242.

Kruger, Haidee, and Gert de Sutter. 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.

Kruger, Haidee. 2019. "*That* Again: A Multivariate Analysis of the Factors Conditioning Syntactic Explicitness in Translated English." *Across Languages and Cultures* 20 (1): 1–33.

Lanstyák, István, and Pál Heltai. 2012. "Universals in Language Contact and Translation." *Across Languages and Cultures* 13 (1): 99–121.

Laviosa, Sara. 1998. "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose." *Meta* 43 (4): 557–570.

Li, Dechao, and Kefei Wang. 2012. "A Corpus-based Study on Lexical Patterns in Simultaneous Interpreting from Chinese into English." *Modern Foreign Languages* 4: 409–415.

Liang, Junying, and Qianxi Lv. 2020. "Converging Evidence in Empirical Interpreting Studies: Peculiarities, Paradigms and Prospects." In *New Empirical Perspectives on Translation and Interpreting*, edited by Lore Vandevoorde, Joke Daems, and Bart Defrancq, 303–332. London: Routledge.

Lv, Qianxi, and Junying Liang. 2018. "Is Consecutive Interpreting Easier than Simultaneous Interpreting? A Corpus-based Study of Lexical Simplification in Interpretation." *Perspectives* 27 (1): 91–106.

Nini, Andrea. 2014. The Multidimensional Analysis Tagger. https://sites.google.com/site/multidimensionaltagger

Olohan, Maeve, and Mona Baker. 2000. "Reporting *That* in Translated English: Evidence for Subconscious Processes of Explicitation?" *Across Languages and Cultures* 1 (2): 141–158.

Pym, Antony. 2015. "Translating as Risk Management." *Journal of Pragmatics* 85: 67–80.

Qin, Hongwu, and Kefei Wang. 2009. "A Parallel Corpus-based Study of Chinese as Target Language in EC Translation." *Foreign Language Teaching and Research* 2: 131–136.

Russell, Debra. 2002. *Interpreting in Legal Contexts: Consecutive and Simultaneous Interpretation*. Burtonsville, MD: Linstok.

Russo, Mariachiara, Claudio Bendazzoli, and Annalisa Sandrelli. 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.

Sandrelli, Annalisa, and Claudio Bendazzoli. 2005. "Lexical Patterns in Simultaneous Interpreting: A Preliminary Investigation of EPIC (European Parliament Interpreting Corpus)." *Proceedings from the Corpus Linguistics Conference Series* 1 (1): 1–18. Birmingham: University of Birmingham. https://www.birmingham.ac.uk/Documents/college-artslaw/corpus/conference-archives/2005-journal/ContrastiveCorpusLinguistics/lexicalpatternsinsimultaneousinterpreting.doc

Scott, Mike. 2012. WordSmith Tools (Version 6). *Lexical Analysis Software*. https://lexically.net/wordsmith/downloads/

Shlesinger, Miriam, and Noam Ordan. 2012. "More Spoken or More Translated? Exploring a Known Unknown of Simultaneous Interpreting." *Target* 24 (1): 43–60.

Shlesinger, Miriam. 1989. *Simultaneous Interpretation as a Factor in Effecting Shifts in the Position of Texts in the Oral-Literate Continuum*. MA thesis. Tel Aviv University.

Shlesinger, Miriam. 1995. "Shifts in Cohesion in Simultaneous Interpreting." *The Translator* 1 (2): 193–214.

Shlesinger, Miriam. 1998. "Corpus-based Interpreting Studies as an Offshoot of Corpus-based Translation Studies." *Meta* 43 (4): 486–493.

Shlesinger, Miriam. 2008. "Towards a Definition of Interpretese: An Intermodal, Corpus-based Study." In *Efforts and Models in Interpreting and Translation Research: A Tribute to Daniel Gile*, edited by Gyde Hansen, Andrew Chesterman, and Heidrun G. Arbogast, 237–253. Amsterdam: John Benjamins.

Van Rooy, Bertus, Lize Terblanche, Christoph Haase, and Joseph Schmied. 2010. "Register Differentiation in East African English: A Multidimensional Study." *English World-Wide* 31 (3): 311–349.

Volansky, Vered, Noam Ordan, and Shuly Wintner. 2015. "On the Features of Translationese." *Digital Scholarship in the Humanities* 30 (1): 98–118.

Wu, Baimei, Andrew K. F. Cheung, and Xing Jie. 2021. "Learning Chinese Political Formulaic Phraseology from a Self-built Bilingual United Nations Security Council corpus: A pilot study." *Babel* 67 (4): 500–521.

Xiao, Xiaoyan. 2015. *On the Oral-Literate Continuum: A Corpus-based Study of Interpretese*. Xiamen: Xiamen University Press.

Xu, Cui. 2021. *Identification of L2 Interpretese: A Corpus-based, Intermodal, and Multidimensional Analysis*. PhD diss. The Hong Kong Polytechnic University.

Xu, Cui, and Dechao Li. 2022. "Exploring Genre Variation and Simplification in Interpreted Language from Comparable and Intermodal Perspectives." *Babel* 68 (5): 742–770.

Addendum

Table A1 presents the seventy-nine linguistic features analyzed in this study. These linguistic features have been grouped into eighteen major categories: (A) tense and aspect markers, (B), place and time adverbials, (C) pronouns and pro-verbs, (D) questions, (E) nominal forms, (F) passives, (G) stative forms, (H) subordination features, (I) adjectives and adverbs, (J) lexical specificity, (K) specialized verb classes, (N) reduced or dispreferred forms, (O) coordination, (P) negation, (Q) overall textual features, and (R) other features. The first sixteen groups are adopted from Biber (1988), while the remaining two groups are selected based on previous scholarship on 'universal' features/patterns of translated and interpreted language.

Table A1. Seventy-nine linguistic features investigated in the MD analysis

(A)	TENSE AND ASPECT MARKERS		30	TOBJ	*That* relative clause on object position
1	VBD	Past tense	31	WHSUB	*Wh*-relative clauses on subject position
2	PEAS	Perfect aspect	32	WHOBJ	*Wh*-relative clauses on object position
3	VPRT	Present tense	33	PIRE	Pied-piping relative clauses
(B)	PLACE AND TIME ADVERBIALS		34	SERE	Sentence relatives
4	PLACE	Place adverbials	35	CAUS	Causative adverbial subordinators
5	TIME	Time adverbials	36	CONC	Concessive adverbial subordinators
(C)	PRONOUNS AND PRO-VERBS		37	COND	Conditional adverbial subordinators
6	FPP1	First person pronouns	38	OSUB	Other adverbial subordinators

Table A1. *(continued)*

7	SPP2	Second person pronouns	(I)	PREPOSITIONAL PHRASES, ADJECTIVES AND ADVERBS	
8	TPP3	Third person pronouns	39	PIN	Total prepositional phrases
9	PIT	Pronoun *it*	40	ATTRJ	Attributive adjectives
10	DEMP	Demonstrative pronouns	41	PRED	Predictive adjectives
11	INPR	Indefinite pronouns	42	RB	Total adverbs
12	PROD	Pro-verb DO	(J)	LEXICAL SPECIFICITY	
(D)	QUESTIONS		43	TTR	Type/token ratio
13	WHQU	*Wh*-questions	44	AWL	Average word length
(E)	NOMINAL FORMS		(K)	LEXICAL CLASS	
14	NOMZ	Nominalizations	45	CONJ	Conjuncts
15	GER	Gerunds	46	DWNT	Downtoners
16	NN	Nouns	47	HDG	Hedges
(F)	PASSIVES		48	AMP	Amplifiers
17	PASS	Agentless passives	49	EMPH	Emphatics
18	BYPA	*By*-passives	50	DPAR	Discourse particles
(G)	STATIVE FORMS		51	DEMO	Demonstratives
19	BEMA	BE as main verb	(L)	MODALS	
20	EX	Existential *there*	52	POMD	Possibility modals
(H)	SUBORDINATION		53	NEMD	Necessity modals
21	THVC	*That* verb complements	54	PRMD	Predictive modals
22	THAC	*That* adjective complements	(M)	SPECIALIZED VERB CLASSES	
23	WHCL	*Wh*-clauses	55	PUBV	Public verbs
24	TO	Infinitives	56	PRIV	Private verbs
25	PRESP	Present participial clauses	57	SUAV	Suasive verbs
26	PASTP	Past participial clauses	58	SMP	SEEM/APPEAR
27	WZPAST	Past participial WHIZ deletion relatives	(N)	REDUCED FORMS AND DISPREFERED STRUCTURES	
28	WZPRES	Present participial WHIZ deletion relatives	59	CONT	Contractions

Table A1. *(continued)*

29	TSUB	*That* relative clauses on subjection position	60	THATD	Subordinator *that* deletion
61	STPR	Stranded prepositions			
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*-pronouns			

Table A2 presents the descriptive statistics for the dimension scores for the three language varieties (NS, SI, and WT) examined in this study. The table includes mean dimension scores, minimum and maximum values of dimension scores within each language variety, as well as their range and standard deviation.

Table A2. Descriptive statistics for dimension scores for the three language varieties across seven dimensions

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