EXPLORING THE TEXTUAL CHARACTERISTICS OF CONSTRAINED ENGLISH VARIETIES:

A COMPARATIVE STUDY OF TRANSLATED ENGLISH, EFL, AND NATIVE ENGLISH USING A MULTIDIMENSIONAL APPROACH

JIAXIN CHEN

PhD

The Hong Kong Polytechnic University

2024

The Hong Kong Polytechnic University

Department of Chinese and Bilingual Studies

Exploring the Textual Characteristics of Constrained English Varieties: A Comparative Study of Translated English, EFL, and Native English Using a Multidimensional Approach

Jiaxin CHEN

A thesis submitted in partial fulfilment of the requirements for the

degree of Doctor of Philosophy

August 2023

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)

__Jiaxin CHEN_____ (Name of student)

Abstract

The present study embarks on an investigation into the textual characteristics shared by translated English (TE) and English as a Foreign Language (EFL), using non-mediated, native English writing (NE) as a benchmark. The rationale behind contrasting TE with EFL stems from the assumption that both may display commonalities as they are influenced by similar cognitive, cultural, and social factors related to bilingual activation—a constraint that is absent in NE (Lanstyák & Heltai, 2012; Kotze, 2022). The study focuses on the textual peculiarities of these two constrained varieties of English, operating on the premise that language variation is multifaceted and systematic, thus a multidimensional analysis could offer a comprehensive view of how various linguistic features connect and correlate to shape variations of the constrained varieties.

Three central research objectives are posed: 1) to identify textual variations of EFL and TE compared to NE; 2) to examine the feature-level variations and their contributions to textual variations; and 3) to discuss the implications for two universal hypotheses of simplification and explicitation, and to interpret the variations in relation to the constraints that condition the two constrained English varieties.

The study utilizes a corpus-based approach, drawing data from published English news articles, including two sub-registers: editorials and news reports. Two self-compiled sub-corpora have been built, with EFL represented by English news written by native Chinese, and translated English represented by English news translated from Chinese. The non-mediated, native English writing is represented by the Press sub-corpus of an established corpus, CLOB (Xu & Liang, 2013).

To address the research questions, a multidimensional analysis of 69 lexico-grammatical features is employed. The feature selection is theoretically motivated and language-pair specific, inspired by both register-oriented and function-oriented language variation studies (Biber, 1988; Neumann, 2014; Le Foll, 2021). The multidimensional analysis is complemented by univariate statistical analysis on individual linguistic features. Based on

the results of the two-phase analysis, variations at both textual and feature levels are interpreted in relation to the shared and distinctive constraints in TE and EFL.

The multidimensional analysis identified six textual dimensions, including Dimension 1, 'Elaborated-involved versus Integrated-formal production'; Dimension 2, 'Evaluative discourse versus Reporting/retelling discourse'; Dimension 3, 'Depictive and detailed narration'; Dimension 4, 'Descriptive narration with a spatial-temporal focus'; Dimension 5, 'Activity focus versus Referential precision'; and Dimension 6, 'Information density versus Irrealis'.

Along the six dimensions, the two constrained varieties exhibit similar yet register-sensitive variations in contrast to NE. Specifically, compared to NE editorials, TE and EFL editorials are characterized by being more evaluative (D2), more integrated and formal (D1), marked by the utilization of irrealis (D6) and devoid of descriptive narration with a spatial-temporal focus (D4). Conversely, in TE and EFL reports, the language becomes more elaborated and involved (D1) than NE reports, characterized by a pronounced reporting discourse (D2), a descriptive narration with a spatial-temporal focus (D4), and an enhanced level of information density (D6). This distinction illustrates the nuanced complexity of textual variations and how they respond to different registers.

The shared tendencies in TE and EFL may stem from the interplay between two opposing forces: the influence of the source/first language (Mandarin Chinese) and the target/foreign language (English). These two forces create a continuum between interference and normalization, leading to complex textual characteristics. The competing or reinforcing nature of these forces pushes the constrained varieties to either emphasize or diverge from features that typify the non-constrained English (e.g., amplification in D2, D4, and D6; opposition in D1).

Distinctions are also found between TE and EFL at the textual level (for instance, D3 and D5), and EFL generally exhibits more variations, evident by a larger and inconsistent divergence from NE. This may be tentatively attributed to the mediated production mode

unique to translation. Translators, compared to EFL writers, may be more experienced in playing the role of mediating between languages and cultures, thus produce translations that are more align with the native production. Another significant finding is the increased variation found in editorials compared to reports. This trend may be related to the constraint of register – greater variation emerges in response to the larger gap in argumentative writing conventions between Chinese and English.

Univariate feature-level analysis yielded mixed results, demonstrating register-sensitive variations. While TE and EFL share tendencies on certain textual dimensions, their realization at the feature level varies. A consistent pattern identified through univariate analysis is that adjectival noun phrases are overrepresented in both TE and EFL. This may be attributed to a combination of interference and normalization, i.e., the preference for left-branching premodification of nouns in Chinese and the trend toward using more adjective-noun sequences in contemporary English (Huang, 1998; Leech et al., 2009). The register-sensitive characteristics challenge prevailing hypotheses regarding simplification and explicitation in translation and EFL studies. These findings again underscore the importance of a multidimensional approach to fully comprehend the multifaceted nature of language variation.

The study highlights the complex interplay of various constraints in shaping language use, particularly language activation and mediation across different communicative contexts. By adopting an interdisciplinary perspective, it extends translation studies beyond simple comparisons between translations and non-translations in the target language. The analysis reveals how constrained language varieties characterized by bilingual activation, such as translation and foreign language writing, intersect and diverge. By highlighting the nuanced variations of constrained English varieties, this study not only enriches translation studies but also fosters its connections with the wider field of bilingual communication, contributing to cohesive understanding of language phenomena.

Publication Arising from the Thesis

Chen, J., Li, D., & Liu, K. (2024). Unraveling cognitive constraints in constrained languages: a comparative study of syntactic complexity in translated, EFL, and native varieties. *Language Sciences*, *2024*, *volume 102*, Article 101612. (published)

Acknowledgements

I am deeply thankful to a remarkable circle of mentors, colleagues, and loved ones, whose steadfast support and encouragement have been the cornerstone of my research journey.

First and foremost, I express my profound gratitude to Prof. Li Dechao. Being a diligent and openminded scholar, Prof. Li's mentorship transcended academic learning. Through every book club and lecture, Prof. Li imparted not just scholarship but also invaluable life lessons, which has indelibly shaped my personal and professional growth.

Special thanks are due to Prof. Haidee Kotze, who generously welcomed me into the academic community at Utrecht University. Her insightful guidance and critical feedback were instrumental in navigating the complexities of my project.

I am indebted to Dr. Liu Kanglong and Dr. Andrew Cheung whose generous support throughout this journey was indispensable. Dr. Liu's detailed advice, in particular, was crucial in enhancing the quality of my research. I am equally thankful to Dr. Sun-A Kim for her insightful feedback during my registration confirmation.

I also express my sincere gratitude to Prof. Zhang Wei and Prof. Kong Lei, my external examiners, for their rigorous review and constructive criticism, which have greatly enhanced the rigor of my thesis.

To the AG518 Marvel Team, becoming part of such an inspiring and supportive community is a great privilege. My deepest thanks go to Bonnie Kwok, Chen Zhuojia, Han Tianyi, Li Juanjuan, Li Nannan, Liu Tingyu, Liu Yanjin, Liu Yi, Liu Yufeng, Qin Yan, Sha Lei, Su Yanfang, Yao Yao, Zeng Weixin, and the many others. They are the most cherished friends I could have hoped for. The time we spent together at the beach, in the mountains, and in our windowless office room will brightly shine in my memory.

A special mention to Chen Xinyi, my roommate of nearly two years, whose companionship and intellectual conversations have been a source of joy for me. Along with Chen Jing, our

Zoom sessions brimming with encouragement were a beacon of support through the challenges of thesis writing.

My heartfelt thanks to Li Xiaohuan and Jiao Linsen, friends I met in Utrecht, whose friendship offered me solace during trying times. The shared meals, the bright sunflowers, and long walks in cozy summer nights are memories I hold dear.

To my dance buddies, particularly Wang Jue, whose love and support has become a constant source of strength for me. My sense of belonging to this city is deeply tied to the memories we shared.

Lastly, to Dr. Chen Zirong, my boyfriend and steadfast supporter, witnessing his dedication and achievements has been a profound source of inspiration. And to my parents who offer me unconditional love and patience, my gratitude knows no bounds. I love you all deeply, beyond words.

Table of Contents

List of Tables

List of Figures

List of Appendices

Chapter 1 Introduction

1.1 Research motivation

In translation studies, there is an enduring interest in uncovering the unique features that distinguish translations from non-mediated language production (House, 2008; Chesterman, 2014). Studies dedicated to the identification of such characteristics have found translated language tend to be simpler both lexically and syntactically, more explicit, and more aligned with the target language's standards or norms (Zanettin, 2012). These observed traits are termed "translation universals", denoted as attributes that "typically occur in translated texts rather than original utterances" (Baker, 1993, p.243). In fact, the quest for linguistic patterns characterizing a language variety is not limited to translation studies. Scholars in areas such as second language acquisition (SLA) and language contact are also interested in unique features of the language varieties of their concern, which often display diverging features compared to native language use, e.g., hyperclarity, anti-deletion, regularization, and simplification (Filppula et al., 2009; Mesthrie, 2006; Williams,1987).

A striking parallel emerges in the linguistic patterns exhibited in translation and contact language varieties in general. Correspondingly, proposals have been made suggesting that such language varieties may share recurring features with translation, thus warranting study within a unified framework. At the heart of this argument lies the hypothesis that all communication is inherently constrained (Lanstyák & Heltai, 2012; Kotze, 2022). In bilingualism-influenced communication, language use is shaped by various intra- and extra-linguistic factors. Translation can be viewed as a specialized form of bilingualism-influenced communication, with similar constraints likely to result in language patterns distinct from native language use. This understanding has led to the proposal of "constrained communication universals" to describe shared features across multiple constrained language varieties (Lanstyák & Heltai, 2012, p.100), with several analogous theoretical frameworks emerging in fields such as SLA, bilingual communication, and

contact linguistics (Chesterman, 2004; Kolehmainen et al., 2014; Granger, 2015; Kotze, 2022).

In recent years, empirical studies guided by the theoretical framework of constrained communication have begun employing a corpus-based approach to compare constrained language varieties. Rather than relying solely on source or target language texts for comparison, translations are now analyzed alongside indigenized varieties, non-native and learner writings, and edited texts across various genres using multifaceted statistical methods (Gaspari & Bernardini, 2010; Granger, 2018; Ivaska & Bernardini, 2020; Kruger & De Sutter, 2018; Kruger & Van Rooy, 2016a, 2016b; Bisiada, 2017; Kruger, 2012). The primary language investigated is constrained English, with source languages predominantly from European and African countries. These studies partially support the hypothesis that constrained language varieties display similar patterns such as reduced lexical diversity and increased formality (Rabinovitch et al., 2016; Kajzer-Wietrzny & Ivaska, 2020; Kruger & De Sutter, 2018; De Sutter & Lefer, 2020; Kruger & Van Rooy, 2016a).

These pioneering investigations have marked significant advancements in extending the exploration of features typifying translation to constrained language. However, the multifaceted nature of linguistic phenomena involved in constrained communication indicates that current empirical inquiries remain limited. For one thing, there is a growing recognition that constrained language demonstrate systematic properties necessitating a multidimensional examination, which remains underrepresented in this research line. Additionally, existing studies contrasting translations with other constrained varieties often neglects typologically distant language pairs such as Chinese-English. This study seeks to bridge these gaps by broadening the research on constrained communication represented by translated English from Chinese and EFL produced by native Chinese and incorporating a multidimensional perspective. By doing so, it seeks to offer a more nuanced understanding of characteristics of translation and constrained language use in broader linguistic contexts.

1.2 Research questions

The present study seeks to situate translation within an expanded scope of constrained communication by juxtaposing two constrained varieties of English—translated English (TE) and English as a Foreign Language (EFL)—against non-mediated, native English (NE). The general hypothesis is that TE and EFL, as constrained language varieties, exhibit shared textual similarities stemming from the common constraint of bilingual activation which is absent in NE. Simultaneously, differences may arise between these two constrained varieties, attributable to the interplay of constraints that are either exclusive to or exert varying degrees of influence on each variety.

Drawing on data collected from published news writing consisting of two sub-registers, i.e., editorials and news reports, the study focuses on the textual characteristics of two constrained English varieties and explores whether the characteristics observed are consequent to their common and unique constraints. Specifically, the research is guided by the pursuit of three main questions:

RQ1 What are the textual variations of the constrained English varieties under examination?

RQ1.1 Based on a multidimensional analysis of a collection of 69 linguistic features, what underlying dimensions of variation can be identified?

RQ1.2 Along the identified dimensions, how do constrained varieties compare to non-constrained English? Are there shared similarities or divergent patterns between the constrained varieties?

RQ2 How do textual variations of the constrained English varieties realize at the level of individual linguistic features?

RQ2.1 Concerning the distributions of the 69 linguistic features, how do constrained varieties compare to non-constrained English? Are there shared similarities or divergent patterns between the constrained varieties?

RQ2.2 How do feature-level variations contribute to textual variations of the constrained varieties?

RQ3 What implications emerge from the linguistic variations of the constrained English varieties?

RQ3.1 Do feature-level variations of the constrained varieties have any implications for constrained language "universals", i.e., simplification and explicitation?

RQ3.2 How do the textual and feature-level variations of constrained varieties relate to their shared and distinctive constraints?

1.3 Data and methodology

This study utilizes a corpus-based approach to investigate translated English and EFL writing, contrasting them with non-mediated, native English as a baseline. The corpus representing native English writing is drawn from the Press sub-corpus of CLOB, a balanced contemporary written English corpus (Xu & Liang, 2013). The EFL and TE corpora, which represent EFL writing and translation respectively, are derived from online English media archives in China: translations from *Global Times* and *Sixth Tone*, and EFL writings from *China Daily*. In total, the corpora consist of about 400,000 tokens, encapsulating a wide variety of domains.

The study emphasizes data comparability, which is vital to bridge the gap between various language varieties within a unified framework. The deliberate choice of news writing produced by professionals is part of the efforts made to ensure comparability. News writing serves as a convenient source for studying language variations, reflecting language use across different periods and regions (Leech et al., 2009). Recognizing the complexity and heterogeneity of newspaper language, the study specifically distinguishes between news reports and editorials as two sub-registers, which are distinctive in terms of their functions and language use. By concentrating on professional news writings and translations in the

Chinese-English context, the study efficiently controls for other constraint dimensions, thus provides an ideal platform for exploring the effects of bilingual activation in TE and EFL.

Methodologically, the three research questions are addressed using two-phase analysis. First, a multidimensional analysis is employed to examine the textual variations of the two constrained varieties. The analysis is based on a comprehensive set of 69 lexico-grammatical features whose selection is theoretically driven with a consideration of language pair specific characteristics, inspired by both register-oriented and function-oriented variational studies (Biber, 1988; Neumann, 2014; Le Foll, 2021). This multidimensional analysis is further enhanced by a univariate statistical examination of individual linguistic features, which elucidates how these features contribute to textual level variations and pinpoints distinctively distributed features in the two constrained varieties. Finally, the study synthesizes the results from both multidimensional and univariate analyses to interpret textual and feature-level variations of constrained English varieties. These interpretations are contextualized within their shared and distinctive constraints, enabling a nuanced understanding of the underlying linguistic phenomena and their interplay.

1.4 Structure of the thesis

The thesis consists of seven chapters.

Chapter One lays the foundation for the thesis by delineating the core motivation behind this study: an exploration of translation as a constrained language variety that shares features with other bilingualism-influenced constrained varieties. The chapter outlines the research questions, offers a succinct overview of the data and methodology employed, describes the organization of the thesis, and clarifies essential terminologies.

Chapter Two commences with an overview of corpus-based translation studies, focusing on those that employ multidimensional and multivariate approaches to investigating translation features. The chapter also explores recent trends, especially in contextualizing

translation features within bilingualism-related communication. It delves into the constrained communication framework and the intersection between translation and non-native language production particularly in the Chinese-English context, and delineates potential constraints involved in translation and English as a Foreign Language.

Chapter Three is devoted to data and methodology employed in the study. It outlines the structure and design of the corpus, and the principles for data selection and preparation. It introduces a two-phase analysis comprising multidimensional analysis and feature-level analysis. Detailed explanations of key procedures of the multidimensional analysis are provided, followed by an exploration of the specific methods and statistical tools used for the feature-level univariate analysis.

Chapter Four presents the results of the multidimensional analysis. Six textual dimensions are identified, and variations of the two constrained varieties along these textual dimensions are analyzed. Given the substantial influence of register identified on constrained language use, the chapter provides an examination of how the textual variations of these constrained varieties manifest differently across sub-registers.

Chapter Five delves into the feature-level variations of constrained language varieties. This section elucidates how the distributions of features contribute to textual level variations and pinpoints distinctively distributed features in the two constrained varieties.

Chapter Six offers a comprehensive discussion of the findings, emphasizing the implications of feature-level variations for two potential constrained language "universals", i.e., simplification and explicitation. The chapter further examines textual and feature-level variations of the two constrained varieties as consequences of the interplay of various constraints, particularly interference and normalization associated with bilingual activation, their intertwining effects with register and the mediation status unique to translation.

The final chapter, Chapter Seven, synthesizes the major findings and discusses their implications. The conclusion also articulates the significance of the current study, acknowledges its limitations, and outlines corresponding directions for future research.

1.5 Terminologies

This section introduces some key terminologies used throughout the thesis, providing clarity and context for the subsequent discussions.

Firstly, in a general sense, "constrained communication" refers to "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). In the context of translation and non-native language production, the term gains a more precise definition, emphasizing the bilingual activation experienced by translators and non-native language users (Kruger & Van Rooy, 2016a). In this study, "constrained language", "constrained communication", and "bilingualism-influenced communication" are used interchangeably, suggesting that these modes of communication are primarily constrained by bilingualism or language contact. Details concerning these concepts are elaborated in Chapter 2.

This study involves constrained language varieties such as non-native English (including English as a Second Language, or ESL, and English as a Foreign Language, or EFL) and translated English. "Non-native English" applies to any English variety used by non-native speakers across all proficiency levels and contexts. This study adopts the definition in Gass and Selinker (2008: 7), using "foreign language" to denote the non-native language learned "in the environment of one's native language", and "second language" for the non-native language learned "in the environment in which that language is spoken". New Englishes and Outer Circle varieties of English are synonymous with ESL. These differ from Learner Englishes, Expanding Circle varieties of English, or EFL. They are also to be distinguished from English used by native speakers, referred to as English as a Native Language (ENL) or Inner Circle varieties of English.

Lastly, in the context of text types, "register" and "genre" offer slightly different perspectives. While the register perspective stresses "an analysis of linguistic characteristics that are common in a text variety with analysis of the situation of use of the variety", genre focuses more on "the conventional structures used to construct a complete

text variety". Despite these subtle differences, "register" and "genre" are used interchangeably in this study, as both concepts involve "the description of the purposes and situational contexts of a text variety" (Biber & Conrad, 2009, p. 2).

Chapter 2 Literature review

This corpus-based study endeavors to broaden the discourse on translation features by situating translation within the wider context of bilingualism-related communication, or constrained language in its narrower sense. As such, the literature review is methodically structured to illuminate various facets of this research. It commences with an overview of corpus-based translation studies, emphasizing those that adopt a multidimensional perspective on translation features, and subsequently highlights more recent trends in the field. Following this, an overview of the constrained language framework is provided, including an elucidation of its conceptualizations and empirical applications. This framework serves as the primary theoretical foundation for the current study, positing that the language use is shaped by a complex interplay of various constraints. Since the research is anchored at the convergence of translation and non-native language production, specifically in the context of Chinese-English, the review further encompasses studies on features of non-native English production, with a particular emphasis on English produced by native Chinese speakers, highlighting resemblances between features of translation and non-native production. Finally, drawing upon previous theoretical and empirical insights, the literature review delineates potential constraints that might influence the characteristics of translated English and English as a Foreign Language.

2.1 Corpus-based studies on translation features

Corpus-based translation studies (CBTS) have witnessed significant development since its inception nearly three decades ago, becoming a prominent field within translation studies. The use of corpora in translation studies has revolutionized research methodologies, enabling data-driven analysis and providing empirical evidence for translation phenomena. From the very beginning, scholars in CBTS have strived to uncover unique patterns that characterize translation, initially coined as “translation universals”. Over the years, CBTS has evolved and expanded to address a broader array of research interests with interdisciplinary approaches. Nonetheless, to “identify the defining features of translation

as a form of interlingual communication" (Granger & Lefer, 2022, p.1) remains one of the central themes in CBTS. Leveraging technological advancements and the growing availability of large-scale linguistic corpora, CBTS continually contributes to the understanding of intrinsic features that delineate translation, offering valuable insights into the complex nature of translation processes and the dynamics of interlingual communication.

2.1.1 From translationese to translation features

In the early stage of translation studies when prescriptive approaches dominated the newborn discipline, "translationese" was used as a pejorative term for the awkward or unnatural language typically associated with translations compared to naturally produced texts in the target language (e.g., Newmark, 1988; see Olohan, 2004, p.90).

However, the overall view on translationese has been evolving as the theoretical and methodological approaches in translation studies have shifted over time. Gellerstam (1986) is one of the first translation scholars who used the term in a purely descriptive manner, and later rephrased the influence of source texts/language in translations as "a neutral concept of *fingerprints*" (Gellerstam, 2005, p.213).

Agreeing on this non-judgmental view on the distinctive linguistic patterns spotted in translation, scholars have also used "a third code" (Frawley, 1984/2000) and "a hybrid text" (Duff, 1981) to describe translation in a neutral sense. Yet different from the emphasis on the source language influence in Gellerstam's (2005) definition, these terms imply that translation is affected by and different from both the source and target languages.

Such observation of the hybrid nature of translation was mainly obtained based on manual analysis of a small number of translated texts, and the discussion was only brought into a new level when corpus-based approaches were introduced into the translation studies, marked by the seminal works by Baker (1993, 1995, 1996). Using techniques from corpus linguistics, sizeable data stored in electric form could be analyzed with automatic or semi-

automatic methods, enabling translation scholars to "identify features of translated text which will help us understand what translation is and how it works" (Baker, 1993, p.243).

To provide a starting point for CBTS, Baker sketched a tentative list of universal features of translations, including simplification, explicitation, normalization and leveling out (Baker, 1993, p.243). These features were believed to be deeply rooted in the very activity of translation where "the need to communicate in translated utterances operates as a major constraint on translation behavior and gives rise to the patterns which are specific to translated texts" (Baker, 1993, p.242), regardless of the source or target languages involved. In other words, this proposal goes beyond the language-specific effects in the early discussions and opens up a new path for comparable analysis between translation and original utterances in the target language.

The translation universal is now more often formulated as common features or general tendencies reoccurring in translation, as the "universality" in this term has been called into question given the diversity of translation practices across different contexts (Becher, 2011). Yet Baker's proposal has a profound impact on translation studies in that it initiates a long-lasting academic interest in searching for features typifying translation. Numerous studies have been devoted to identifying translation features, with the most investigated ones on Baker's list: simplification [1] (Blum & Levenston, 1978; François & Lefer, 2022; Grabowski, 2013; Jantunen, 2001; Laviosa, 1998, 2002; Teich, 2003; Kajzer-Wietrzny et al., 2016; Liu & Afzaal, 2021), explicitation[2] (Blum-kulka, 1986; Olohan & Baker, 2000; Olohan, 2003; Becher, 2011; Puurtinen, 2004; Pápai, 2004; Kruger & De Sutter, 2018),

[1] Baker (1996: 181) defines simplification as translators' tendency to "subconsciously simplify the language or message or both". Yet as acknowledged by Ferraresi et al. (2018: 734), the term *simplicity* would be more appropriate for this phenomenon, while *simplification* would be more suitable for a parallel approach comparing translation with its source texts. These terms are used interchangeably in some literature. In this study, *simplification* is adopted given its wide currency in monolingual comparable studies, and no ambiguity will be caused as this study only deals with the relationship between translation and original texts in the target language using a comparable approach.

[2] According to Baker (1996:180), *explicitation* refers to the tendency to "spell things out rather than leave them implicit". A terminology classification has been made between *explicitation* and *explicitness* (Hansen-Schirra et al., 2007). In short, *explicitness* is a property of lexico-grammatical or cohesive structures and configurations in one text, while *explicitation* refers to the process of making the intra- or interlingual output more explicit than their counterparts in terms of their lexico-grammatical and cohesive properties. *Explicitation* is adopted in the current study which employs a comparable corpus approach.

normalization[3] (Scott,1998; Mauranen, 2007; Bernardini & Ferraresi, 2011; Williams, 2005; Hansen & Hansen-Schirra, 2012) and levelling out[4] (Grabowski, 2013; Laviosa, 2002; Williams, 2005; Redelinghuys, 2016).

These proposals of translation features have encouraged a body of working hypotheses and empirical data that have enriched the discipline enormously. Besides these four mostly discussed universals, other hypothesis about translation features have also been proposed, for example, the Asymmetry Hypothesis[5] (Klaudy & Károly, 2005), the Unique-item Hypothesis[6] (Tirkkonen-Condit, 2004), the Shining-through Hypothesis[7] (Teich, 2003), and the Gravitational Pull Hypothesis [8] (Halverson, 2017). Some of these are reformulations of the translation features under specific conditions (e.g., the Asymmetry Hypothesis dealing with the relationship with explicitation and implicitation in different translation directions), while others are more complex models that attempt to provide explanations especially when contradicting observations are made in the empirical studies (e.g., the Gravitational Pull Hypothesis).

The intense efforts made to summarize and elucidate translation tendencies demonstrate the significant research interest in this area. However, the presence of overlapping and ambiguous definitions within these concepts has been subject to criticism (Lefer & Vogeleer, 2013; Pym, 2008; Olohan, 2004; Bernardini & Zanettin, 2004). For instance, the manifestation of simplification, characterized by reduced lexical diversity and syntactic variation, may also be attributed to the tendency towards the center of the lexico-grammatical continuum in the target language (Olohan, 2004; Lefer & Vogeleer, 2013).

---

[3] In Baker's (1996: 193) definition, normalization is translators' "tendency to conform to patterns and practices that are typical of the target language".

[4] Levelling out is defined by Baker (1996: 184) as "the tendency to gravitate around the centre of any continuum rather than towards the fringes".

[5] The Asymmetry Hypothesis posits that "explicitations in the L1-L2 direction are not always counterbalanced by implicitations in the L2-L1 direction because translators – if they have a choice – prefer to use operations involving explicitation, and often fail to perform optional implicitation" (Klaudy & Károly, 2005, p.14).

[6] According to Tirkkonen-Condit (2004), translations tend to contain fewer "unique items", i.e., linguistic elements with no straightforward equivalents in the source language, than comparable non-translated texts.

[7] The Shining-through Hypothesis posits that "In a translation into a given target language (TL), the translation may be oriented more towards the source language (SL), i.e., the SL shines through" (Teich, 2003, p.207).

[8] The Gravitational Pull refers to "a cognitive force that makes it difficult for the translator to escape the cognitive pull of highly salient representational elements in the source language" (Halverson, 2017, p.14), and the revised gravitational pull model posits that the translation outcome could be over- and under-representation of linguistic features depending on three sources: target language salience, source language salience and the cross-linguistic link strength.

Similarly, the higher frequency of connectives identified in translations has been interpreted as a manifestation of explicitation, normalization, or source language interference in different contexts of language pairs and text categories (Kunilovskaya & Corpas Pastor, 2021).

Another related deficiency in the exiting research concerns the prevalence of investigating translation tendencies through the narrow lens of single-feature operationalization. Inspired by register studies and variationist linguistics, scholars increasingly recognize the importance of treating translation as a sub-language characterized by systematic variation, which could be best captured by considering a constellation of linguistic features (Evert & Neumann, 2017; Hu et al., 2016; Kunilovskaya & Corpas Pastor, 2021; Prieels et al., 2015). As highlight by Evert and Neumann (2017:2), translation properties are systematic and are rarely discernible through a singular feature alone. Just as certain linguistic phenomenon may be associated with more than one proposed universal tendencies, such tendencies are mostly likely to be realized through the combined contribution of multiple linguistic features. Studies that solely focus on individual features are unable to unveil the intricate correlations between features, potentially leading to an incomplete understanding of the unique properties of translation. Thus, a multivariate approach is strongly advocated for a systematic investigation of translation features.

2.1.2 A multivariate perspective on translation features

Despite the increasing awareness of the need for a multivariate approach in translation studies, its adoption remains limited. A recent survey conducted by Granger and Lefer (2022) analyzed translation studies published between 2012 and 2019, which reveals that a significant proportion of CBTS focused on individual linguistic phenomena either as the main research focus or as a means to evaluate the validity of specific translation features. While a wide range of linguistic phenomenon is considered, only a few studies assessed translation features "on the basis of a mixture of words, phrases, and structures (the MX category) rather than a single linguistic phenomenon" (Granger & Lefer, 2022, p.27). A few exceptions that embraced a multivariate perspective include the work of Delaere et al.

(2012), which verified the standardization hypothesis by examining the variation of a group of lexical items between translated and non-translated Dutch. Similarly, Kruger and Van Rooy (2012) operationalized three proposed translation universals using three sets of linguistic variables at the lexical level. Although not strictly adhering to a multivariate method, these studies acknowledged the importance of considering translation features from a multivariate perspective.

Empirical efforts to examine translation features using multivariate techniques have also been made. For instance, drawing on inspiration from Biber's multidimensional analysis on spoken and written language, Hu et al. (2016: 25) conducted an exploratory factor analysis to uncover "statistically significant, consistently distributed, and systematically co-occurring differences" between translations and non-translations across registers and genres. Based on the analysis on 96 lexical, syntactic, and textual features, a group of features was clustered into a translational dimension. The interpretation of the communicative functions of the grouping allows them to identify a "incomplete reflection of source-language informality" (Hu et al., 2016, p.29) as a translation typicality in addition to the well-known simplification and explicitation universals.

Another notable multivariate study was conducted by Evert and Neumann (2017) with the aim of investigating the impact of translation directionality on translation features. Two common methods for multivariate research, namely, principle component analysis and linear discriminant analysis, were adopted respecively to make the comparison and to visualize the results. The study revealed different degrees of source language shining through effect in the two directions of English-German translation. Importantly, the findings were based on patterns found in a complex combination of features rather than interpretation of individual features, as relying solely on individual features "may lead to spurious results that could be counteracted by other features not included in the study" (Evert & Neumann, 2017, p.29). By adopting a multivariate approach, the study sheds light on the intricate relationship between translation directionality and the manifestation of source language influence in translations.

While the research of translation peculiarities has significantly advanced with the introduction of concepts and methods from corpus linguistics, the integration of machine learning (ML) techniques from computational linguistics has further enriched this line of research. ML techniques have facilitated the task of translationese detection, which focuses on distinguishing translated texts from non-translated texts. In this context, "translationese" is used to refer to the statistical difference between translated and non-translated texts without any negative connotation (Volansky, 2015). One pioneering study by Baroni and Bernardini (2006) employed Support Vector Machine (SVM) to differentiate translated geopolitical texts in Italian from original Italian comparable texts. Following this research, SVM has become a commonly adopted text classifier in subsequent translationese detection studies (Ilisei et al, 2010; Volansky et al. 2015; Popescu, 2011; Lapshinova-Koltunski, 2022).

It should be noted that text classification inherently involves a multivariate approach, as it analyzes datasets with multiple variables to discover underlying structures. However, feature selection in translationese detection research using ML text classification may vary significantly depending on the research objectives. Two distinct research purposes can be identified. The first aims to test the application of text classification techniques to the task of distinguishing translations from non-translations, often referred to as building a "translation spotter" (Baroni & Bernardini, 2006). Scholars with this research aim tend to choose surface language forms for feature selection, such as characters, words, lemmas, part-of-speech (POS) tags, mixed n-grams of these units, and punctuation marks (Kurokawa et al., 2009; Popescu, 2011; Grieve, 2007). These features are selected for their ease of operation and efficiency for a computational challenge.

On the other hand, some research is more linguistically oriented, focusing on providing insights into the linguistic specificity of translations beyond a classification task. In such studies, indicators informed by translation studies are selected. Commonly tested indicators include type-token ratio, function words ratio, frequencies of conjunctions and pronouns, ratio of contractions to full forms, average sentence length, and mean word rank. More elaborately engineered features have also been adopted, such as entropy (Hu & Kübler,

2021; Liu et al., 2022) for information density and complexity, point-wise mutual information (Volansky, 2015) for normalization, and other measures based on syntactically parsed data (Ilisei et al. 2010; Ilisei & Inkpen, 2011; Kunilovskaya & Kutuzov, 2019; Kunilovskaya & Corpas Pastor, 2021).

The implementation of linguistically informed features involves two approaches. The top-down approach assigns features to established translation tendencies (Ilisei et al., 2010; Volansky et al., 2015), while the bottom-up approach empirically establishes the role of features in generating various translational effects. An example of the latter is Kunilovskaya and Corpas Pastor (2021), which extracted 45 morphosyntactic features and 11 abstract lexical features to capture distinct choices observed in Russian translations from English compared to original Russian texts across four registers. By examining deviations of translations from non-translations, they inductively concluded that these deviations reflect translation trends such as shining-through, over-normalization, and adaptation.

In summary, the investigation of translation features has greatly benefited from multivariate analysis, particularly with the integration of computationally intensive studies that employ sophisticated methods and extensive feature sets. Machine learning algorithms have achieved high accuracy in distinguishing translations from non-translations, providing compelling evidence for the presence of distinctive features of translations. However, while these studies excel in classification, many of them do not aim to identify the linguistically interpretable factors or the underlying causes that explain the specific characteristics of translated texts. Sophisticated techniques can capture the linguistic patterns that are otherwise unavailable, but they do not guarantee in-depth understanding of the patterns they capture. As highlighted by Kunilovskaya and Corpas Pastor (2021, p. 168), “the machine learning results can be convincing mathematically, but they remain a noumenon unless they are related to human perception.” In other words, while machines can address ontological questions, it is human perception that enables us to comprehend and interpret the results obtained. The “epistemological unease” (Volansky et al., 2015, p. 27) remains unresolved, necessitating the development of theoretical frameworks within the field of translation studies to achieve a comprehensive understanding.

2.1.3 New trends in corpus-based translation studies

In recent years, CBTS has undergone further evolution and expansion, driven by emerging trends that advocate for a multifactorial, multimethodological, and interdisciplinary approach (De Sutter & Lefer, 2020). As scholars have increasingly recognized the multidimensional nature of translation both as a linguistic product and an activity situated within specific socio-cultural contexts, there has been a growing emphasis on capturing the complexity of translation features and investigating the diverse factors that influence them.

The shift to a multivariate approach reflects a deeper awareness that translation, as a sub-language, possesses systematically different properties, thus the description of translation phenomena and the exploration of their underlying influences require a multidimensional perspective. As discussed above, researchers have moved beyond examinations of isolated linguistic variables and have begun to consider translation features as part of a larger constellation of factors. The multivariate studies have shed light on the intricate interplay between various linguistic elements and uncovered patterns that may have gone unnoticed in conventional single-feature analyses. Similarly, these translation features are consequences of various factors related to the translation activity, including the communicative settings (e.g., language pairs, text types and registers, and translation direction, etc.) and the entities involved (e.g., translators' language proficiency and task expertise, sponsorship, etc.). To better understand translation, it is crucial to consider various factors that shape language use in translation. Thus, studies are probing into the relationship between the manifestations of translation features and a range of conditioning factors, such as the source language (Volansky et al, 2015; Koppel & Ordan, 2011; Hu & Kübler, 2021), translation direction (Kurokawa et al, 2009; Lembersky et al., 2011), register (Kunilovskaya & Corpas Pastor, 2021), and translator proficiency (Rubino et al., 2016; Kunilovskaya & Lapshinova-Koltunski, 2022; Lapshinova-Koltunski et al., 2022).

Methodological advancements are also evident, as traditional corpus-based methods are being complemented by experimental designs that incorporate eye-tracking techniques and other state-of-art experimental tools (Neumann et al., 2022). These methods allow

researchers to triangulate findings, enhancing the validity and reliability of the results. Within the realm of CBTS, studies are also gaining new insights by combining sophisticated statistical methods and computational techniques. Alongside statistical techniques that facilitate multidimensional analysis (e.g., Principal Component Analysis, Factor Analysis, Cluster Analysis), statistical methods such as Correspondence Analysis (CA), the Multifactorial Prediction and Deviation Analysis with Regressions (MuPDAR) have been combined with supervised ML algorithms (e.g., SVM, and Linear Discriminant Analysis). This integration of different empirical methodologies and specific methods enables effective identification of translation features, analysis of the influential factors, and visualization of the complex patterns observed.

Another noteworthy trend is the growing emphasis on interdisciplinary collaboration with related fields. Interdisciplinary efforts are extending beyond mere utilization of the analytical tools borrowed from corpus linguistics and computational linguistics, or solely drawing theoretical insights from cognitive science or sociology. Instead, researchers are actively striving for a more profound integration of findings across multiple disciplines, including second language acquisition and variationist linguistics. Especially, when significant overlaps in the observed patterns in translation, second language production and other contact varieties, it is natural to promote more interdisciplinarity between these fields, grounded in the recognition that these language varieties share a common cognitive foundation stemming from some form of language contact. While the notion of bridging disciplines involving language contact situations is not a novel concept, previous proposals have often remained at the level of assumptions with limited theoretical or empirical validation (Blum-Kulka, 1986; Blum & Levenston, 1978; Jarvis & Pavlenko, 2008; Granger, 2015; Kolehmainen et al., 2014). However, there has been a growing effort to move beyond these early assumptions and establish a more rigorous framework and empirical investigations. One notable framework that has emerged is the concept of "constrained communication" (Lanstyák & Heltai, 2012). Adopting such an approach will contribute to extend the scope of translation studies by attempting to generalize translation "universals" to "constrained communication universals". More importantly, a theoretical

significance will be achieved by looking at these varieties with a unified lens, as this research agenda will cultivate a holistic understanding of the similar phenomena observed within these various fields and shed light on how language is shaped by various intra- and extra- linguistic constraining factors.

The idea of constrained communication is of particular relevance to the current study, as it allows for a comprehensive exploration of translation features within the broader context of language contact phenomena. The next section will elaborate on the constrained language framework, providing a theoretical foundation for the study and enabling a deeper understanding of the intricate relationship between translation and other language contact phenomena.

2.2 Translation features: A constrained communication perspective

2.2.1 Constrained communication framework: conceptualization and empirical investigations

Constrained language is defined by Kruger and Van Rooy (2016a: 27) as "the language produced in communicative contexts characterized by particularly conspicuous constraints". Central to the constrained language framework is the pivotal concept of "constraint", which is also a recurrent term in discussions related to factors influencing language use within translation studies. Chesterman (2004) emphasized that translation is subject to constraints, noting that these constraints may not be unique to translation. Instead, they may "be present in other kinds of constrained communication, such as communication in a non-native language or under special channel restrictions, or any form of communication that involves relaying messages, such as reporting discourse or even journalism" (Chesterman, 2004, p.10). Lanstyák and Heltai (2012) further extended this idea by conceptualizing factors that affect translation and bilingual communication as constraints. They argued that human communication is inevitably subject to a range of linguistic and extra-linguistic constraints. They employ the term "constrained

communication" in a narrower sense to denote scenarios where one or several limiting factors are more pronounced than usual.

Building on these insights, Kotze (2022) extended the conceptualization of constraint, identifying five overarching dimensions that collectively influence language production. These five dimensions are as follows:

1) Language activation: In bilingual communication where two languages are simultaneously activated, language users are thought to face an elevated level of cognitive demand. The typological distance between the two languages and the directionality of communication also affect the cognitive processing. 2) Modality and Register: This dimension encompasses written, spoken, and multimodal categories of language production, and includes considerations of genres and stylistic expectations. 3) Text production: When producing a text based on a pre-existing one (e.g., a source text in translation), there are additional restrictions compared to independent or non-mediated communication. 4) Proficiency: The language producer's proficiency level (native/proficient vs. learner) can affect language production. 5) Task expertise: Experts in tasks like academic writing or translation are expected to perform differently from novices.

Two key characteristics of these dimensions should be noted. Firstly, each dimension possesses a continuous nature rather than a binary one. Secondly, both cognitive and social aspects are relevant to each dimension. These features reflect the principles of a usage-based linguistics perspective, which posits that language emerges through repeated patterns of use which is conditioned by a dynamic interplay between cognitive and social factors. Language choices are not deterministic; they are influenced by the likelihood of certain options being more prevalent or preferred, making the frequency of usage or exposure significant. It emphasizes individual cognitive processing and broader social communicative contexts, resonating with the socio-cognitive nature of translation (Halverson & Kotze, 2021). Kotze and Van Rooy (forthcoming) also refers to Schmid's (2015) Entrenchment-and-Conventionalization model to represent the socio-cognitive

mechanism underlying linguistic processes, accounting for both cognitive and social aspects of constrained language use.

Within the context of this framework, scholars have empirically contrasted translations with various constrained language varieties including non-native indigenized varieties, learner varieties, and edited language (Gaspari & Bernardini, 2010; Granger, 2018; Ivaska & Bernardini, 2020; Kruger & De Sutter, 2018; Kruger & Van Rooy, 2016; Bisiada, 2017; Kruger, 2012) across diverse registers and language pairs. Additionally, different modes of communication have been considered, such as interpreting, which has been compared to other non-native spoken varieties (Shlesinger & Ordan, 2012; Kajzer-Wietrzny & Ivaska, 2020; Kajzer-Wietrzny, 2022). These investigations have identified shared traits among the constrained language varieties, such as more explicit lexical-grammatical encoding, lexical simplification, increased formality, and reduced personal involvement.

These characteristics manifest in varying degrees in different constrained varieties, reflecting constraints of language pairs, registers, and expertise levels of the language users. This variability resonates with the constraining dimensions identified within the framework, illustrating how constrained language is shaped by the complex interweaving of various constraints. As Kotze (2022) emphasizes, these factors, or "constraints", exert different impacts across constrained language varieties, leading to nuanced differences in linguistic distribution among them. These variations not only validate the framework's underlying principles but also enable insights into the interplay and relative significance of cognitive and social constraints in diverse contexts.

2.2.2 Studies on non-native English features

In a globalized environment where bilingualism and multilingualism are increasingly common, English has been incorporated into the linguistic repertoire of many non-native speakers as a lingua franca (De Groot & Christoffels, 2006; Kroll et al., 2014). The scholarly examination of non-native English, a field of interest to both variationist linguistics and Second Language Acquisition (SLA), has been shaped by a dichotomy in

the categorization of English types. Within the framework of World Englishes, English varieties are segmented by Kachru's Three Circles model (1982), delineating Inner, Outer, and Expanding Circle Englishes. Outer Circle Englishes, affiliated with local norms as supplementary official or semi-official languages, are denoted as L2 indigenized or New Englishes. Conversely, Expanding Circle English, devoid of official language status, is mainly learned formally and termed Learner Englishes. In parallel, the field of SLA demarcates between English as a Second Language (ESL) and English as a Foreign Language (EFL), reflecting the Outer and Expanding Circle Englishes respectively. Researchers in World Englishes concentrate on the idiosyncratic linguistic patterns of indigenized L2 varieties, while SLA research zeroes in on the structural and lexical variations of ESL and EFL, often with educational implications (Laporte, 2012; Gries & Deshors, 2015).

Despite these distinctions, both types of non-native Englishes share a common origin in language contact situations and are acquired in institutionalized contexts, albeit to varying degrees (Mukherjee & Hundt, 2011). This shared aspect, recognized in both World Englishes and SLA research, gives rise to similar characteristics in the language use across ESL and EFL, as demonstrated in empirical studies. Such resemblances encompass variations in the frequency of linguistic structures and functions, such as high-frequency lexical items (Laporte, 2012; Edwards & Laporte, 2015) and phraseology (Nesselhauf, 2009; Gilquin, 2011; Götz & Schilk, 2011).

Analogous to translation studies, investigations of non-native Englishes reveal overarching tendencies. For instance, a preference for grammatical analyticity over syntheticity was noted by Szmrecsanyi and Kortmann (2011), signifying an inclination towards explicit typological profiling. Other traits include a lack of informality or "monostylistic" approach (see Gilquin, forthcoming), and simplified language usage (see Gilquin & Granger, 2011) - characteristics parallel to those observed in translation. Cross-linguistic interference or traces of the first language in non-native production has also been identified (Jarvis & Pavlenko, 2008; Biewer, 2011).

Interestingly, studies have explicitly found resemblances between translated language and non-native language (Lefer & Vogeleer, 2013). For example, Gaspari and Bernardini (2010) observed both English written by Italian speakers and English translated from Italian manifesting a higher frequency of sentence-initial *therefore* than original English in the same genre and domain. More broadly, Koppel and Ordan (2011) discovered that features utilized to discriminate between translated and original language also proved effective in differentiating between non-native and native writing.

Recent scholarship emphasizes a shifting perspective on EFL/Learner Englishes and ESL/New Englishes, advocating for a continuum rather than a strict dichotomy between the two (Gilquin & Granger, 2011). Meanwhile, driven by the growing population of English-literate individuals and its increasing use within an intranational context, English in China is increasingly recognized as a developing variety in the midst of codification and standardization (Xu, 2010). The term "Chinese English" has been adopted to denote the English used by people with Chinese as their mother tongue, and has been positioned as an English variety in its own right (Kirkpatrick & Xu, 2002; Albrecht, 2021). The consensus generally positions Chinese English towards the EFL end of the cline, reflecting the limited status and functions of English in everyday life in the country (Bolton & Botha, 2015). Nevertheless, scholars have identified its distinct characteristics in areas such as phonology, prosody, lexical and syntactic structures, pragmatics, and discourse preferences (Liang & Li, 2017; Xu, 2020; Ren, 2017; Xu, 2008). Specific examples include higher frequencies of complex nominalization with greater reliance on premodification (Liu et al., 2017) and the increased presence of certain syntactic features, such as parallel and particularly multiple-coordination structures, and modifying-modified sequencing (Xu, 2008; Xu, 2010).

Some of the distinctiveness has been attributed to interference from the Chinese language, characterized by the transfer of linguistic and cultural norms in various areas. For example, the placement of subordinate conjunctions, such as *although*, *because*, *if*, and *when*, may conflict with the expectations of speakers of other varieties of English without being ungrammatical. This alignment with sequential patterns in Chinese represents an

intersection of language evolution and cultural context (Jiang, 2017). It illustrates the dynamic nature of Chinese English, which embodies a unique linguistic identity while maintaining functionality within the global English-speaking community. This ongoing research into the features and positioning of Chinese English underscores its growing significance and the complexity of categorizing and understanding English varieties worldwide.

2.2.3 Translation and EFL: shared and distinctive constraints

Compared to native language production, bilingual activation emerges as the most significant constraint for both translated English and EFL, marking the first constraint dimension within the constrained language framework. In the context of this constraint dimension, there are several influential factors, including the typological relationship between the languages in question, socio-cultural factors such as the relative prestige of the two languages within their respective communities, and translation directionality, etc.

A critical consequence of bilingual activation is that it induces interaction and competition between the bilingual individuals' two languages during language production (Costa & Sebastián-Gallés, 2014). This necessitates a series of cognitive operations such as selection, switching, and inhibitory control between languages, which results in elevated processing costs and may potentially diminish available cognitive resources, and consequently leads to effects in various aspects of language production such as "restricting lexical range and grammatical complexity, prompting increased syntactic explicitness, causing decreased sensitivity to factors like style or register" (Kotze, 2020, p.77).

Meanwhile, characteristics of the production may be language-specific, depending on the specific language pair in question. For instance, EFL writing often reveals traces of the writers' first language (L1), and in translation, evidence of the source language interference may similarly emerge in the target text. This phenomenon, often referred to as cross-linguistic influence (CLI) or transfer, has been extensively studied under various terminologies in SLA studies (Odlin, 1989; Jarvis & Pavlenko, 2008). Cross-linguistic

transfer can be further classified into overt and covert transfer (Mougeon & Beniak, 1991). Overt transfer involves the integration of lexical items and syntactic structures from L1/source language into L2/target language when direct equivalences do not exist. Covert transfer results in an altered distribution of pre-existing lexical or syntactic features within the L2/target language.

Various models for bilingual language representation have been proposed to account for the psycholinguistic and socio-cognitive mechanisms underlying cross-linguistic transfer (Kroll & Tokowicz, 2005). Cross-linguistic transfer denotes the convergence of two languages in the bilingual mind. On the one hand, this can stem from the elevated cognitive load during language production, leading to less control over language boundaries. Further, bilingual individuals build mental associations between linguistic elements in both languages based on their formal or functional attributes (Croft, 2000; Matras & Sakel, 2007), which augments this convergence. From a socio-cognitive standpoint, bilinguals may strategically transfer elements from one language to another, exploiting potential communicative advantages (Kranich, 2014).

Conversely, EFL writing and translation may also display a tendency to adhere to the standard patterns, or "repertoremes" (Toury, 1995), of L2/source language. Referred to as "standardization", "normalization", or "conventionalization" in EFL and translation studies, this tendency is often explained through a socio-cognitive lens. For instance, the risk aversion hypothesis proposed by Pym (2008) posits that translators lean towards the standards of the target language or culture to minimize communicative risks inherent in cross-cultural and cross-linguistic mediation. The use of frequently occurring, standard items and structures of the target language could be a safer strategy that circumvents potential misunderstandings arising from ambiguous or simply challenging aspects of translation.

In the context of translation, mediation emerges as a distinctive constraint differing from EFL. However, the status of being mediation is not exclusive to translation. Other forms of rewriting such as editing all involve "a certain degree of mediation on the part of the

writer/translator to adapt texts to the new audience" (Lefevere, 1992, p.9). This view is echoed by Ulrych and Murphy (2008) which contends that translation and "editing, copy-editing, revision or postediting" as well as ghost-writing "are processed, or rewritten, for particular audiences and are thus mediated for a purpose" (Ulrych & Murphy, 2008, p.151).

Despite the theoretical commonalities, empirical support for the concept of "mediation universals"—shared features between translation and other mediated texts—remains elusive. The overarching definition of mediated discourse makes it challenging to obtain unmediated texts for empirical investigations. Research exploring common phraseologies and universal features has been conducted (Ulrych & Murphy, 2008; Kruger, 2012; Bisiada, 2017), but no convincing evidence has emerged to substantiate the existence of these shared attributes. Instead, Kruger (2012) found that differences between the translated and edited texts may be attributed to the variations in monolingual/bilingual processing and free/constrained production circumstances. Bisiada (2017) found little evidence in favor of mediation universals shared by edited and translated texts but emphasized that "[it] does not mean that changes to the text are negligible, but rather that editors do not intervene in such a way to make the articles more like the non-translated articles." (Bisiada, 2017, p.269) One possible explanation for the discrepancies posits that translators, more so than editors, may exhibit a heightened sense of risk aversion, as either they or the original authors would bear the responsibility for any communication issues (Pym, 2008; Becher, 2010).

It must be emphasized that these constraint dimensions do not operate in isolation. Changes in one constraint dimension can affect the manifestations of another, creating a multifaceted, dynamic system where constraints' collective impact must be considered. Therefore, shared constraints between translation and EFL, along with translation's unique mediation constraint, interact with other dimensions including registers examined, and EFL writers and translators' task expertise and language proficiency. This interconnectedness underscores the importance of a holistic consideration when assessing their influence on language production.

2.3 Research gaps

Translations, as established by existing literature, exhibit certain characteristics that distinguish them from non-translated texts. A growing consensus suggests that these characteristics embody as "systematic properties of text … [which] are hardly ever observable on the basis of just a single feature" (Evert & Neumann, 2017, p.2). For this reason, earlier research focusing on individual features occasionally led to inconsistent findings. This has led to a recognition of the need for multivariate techniques in investigating the systematic properties of translation through a cluster of linguistic features (Hu et al., 2016; Evert & Neumann, 2017). However, recent literature surveys indicate that such multivariate studies, despite their potential, remain relatively underrepresented (Granger & Lefer, 2022; Zanettin et al., 2015; Van Doorslaer & Gambier, 2015). Acknowledging this research gap, the present study seeks to extend this multivariate line of research. The research foci are the shared properties of translation and EFL at the textual level, which necessitates the integration of a multidimensional perspective. By doing so, we aim to provide a more comprehensive and nuanced understanding of the complex interplay of linguistic features defining translation and EFL.

The notion of broadening translation studies to encompass a larger scope is not a recent development. Speculations have been made that translation "universals" are common characteristics of mediated or contact language in a more general context. However, empirical exploration in this area has begun to gain momentum only recently, as indicated by the growing body of work within the constrained communication framework. These research efforts have mirrored the trend in translation studies, demonstrating a shift from the examination of single variables towards a more comprehensive, textual level (Kotze & Van Rooy, forthcoming). Exemplary studies displaying this textual orientation include works by Kruger and Van Rooy (2016a), Kruger and Van Rooy (2018), and Liu et al. (2023). These pieces of research symbolize the move towards a comprehensive, holistic analysis that captures the intricate complexities of translation and mediated language use, and this study aims to contribute to this expanding research frontier.

Existing empirical research on constrained language use presents a conspicuous gap: the primary focus lies in contrasting translations with ESL, non-native indigenized varieties, or learner varieties. This study, however, aims to broaden the scope by emphasizing EFL beyond the learner level. This shift is partly motivated by the dynamic landscape of World Englishes, in which traditionally Expanding Circle Englishes are undergoing considerable changes in an increasingly globalized and interconnected environment, subsequently cultivating unique linguistic characteristics (Edward, 2011). This approach is in alignment with the recent call in SLA and World Englishes studies to "bridge the gap" between EFL/Learner Englishes and ESL/New Englishes (Mukherjee & Hundt, 2011). Particularly, "there has been a dearth of studies concerning the grammatical and morphosyntactic features of Chinese Englishes" (Bolton et al., 2020, P.506). This ties into another deficiency in most existing research on constrained communication, which is the primary focus on European/African languages while giving less attention to typologically distant languages like Chinese and English. A notable exception to this trend is the work of Liu et al. (2023), although their focus was on spoken language production. By acknowledging and addressing these research gaps, the present study intends to offer new perspectives by incorporating less investigated EFL in the context of Chinese/English to the understanding of constrained language use.

Chapter 3 Data and methodology

The exploration of typical linguistic patterns in constrained language use calls for examination of an expansive dataset that represents authentic linguistic usage. This type of in-depth, large-scale examination is made possible through the use of corpus-based methods and tools. A corpus-based linguistic approach has become indispensable in the realms of translation studies, SLA studies, and contact linguistics research due to its empirical orientation and methodological precision, which aid in identifying specific linguistic characteristics of the language varieties under investigation. Especially, "corpus-based studies in translation are clearly aligned with the descriptive perspective" (Olohan, 2004, p. 10), a viewpoint fostered by Toury's (1995) descriptive translation studies (DTS) that has maintained its prominence to this day. To comprehend how language is used in constrained communication, it is crucial to leverage authentic data and the empirical rigor offered by a corpus-based linguistic approach. This section outlines the specifics of corpus construction, including corpus design, as well as data collection and preparation procedures for corpus compilation. An overview of the data from the perspective of constrained communication is provided, illustrating the constraint matrix concerning the language varieties under examination. Subsequently, a comprehensive account of a two-phase data analysis is presented, including the multidimensional analysis and a follow-up examination zooming into the individual linguistic features.

3.1 A Corpus-based approach

3.1.1 Corpus design

This study employs a corpus-based approach to juxtapose translated English and EFL writing, using native English writing as a baseline. The corpus representing native English writing is the Press sub-corpus of CLOB, a balanced contemporary written English corpus developed by the Beijing Foreign Studies University (Xu & Liang, 2013). The EFL and TE sub-corpora, compiled to represent EFL writing and translation respectively, are

derived from online English media archives in China: translations from *Global Times* and *Sixth Tone*, and EFL writings from *China Daily*.

Considerable care was taken to ensure the comparability of the sub-corpora representing NE, EFL, and TE. For a study aiming to bridge the gap between studies on different language varieties, one of the main challenges is obtaining data that is comparable within an integrated paradigm. Given the primary interest of this study is the effects of bilingual activation, a common constraint in translated English and EFL, it is essential to control other constraint dimensions.

To address this issue, the study specifically examines published written news articles produced by professional writers and translators. The Press register is considered a suitable genre for this study, as newspaper articles serve as a convenient source for studying language variation and their representativeness of language use in specific time periods and regions (Leech et al., 2009). News writing is also regarded as a form of formal writing (Biber et al., 1988). However, it is important to note that the language of newspaper "is not a single, and homogeneous object of study" (Semino, 2009, p.533). It is influenced by factors such as newspaper styles, subject domains, and cultural considerations. A crucial factor determining newspaper language use is its sub-registers including categories such as news reports, editorials and feature articles or review articles, to name a few (Biber & Conrad, 2009).

News reports, regarded as the "staple of newspaper writing, or the core content of the newspaper industry" (Ngai, 2022, p. 56), aim to provide objective and neutral coverage of events. They often include a significant amount of quoted materials or attribution to convey a wide range of perspectives. On the other hand, editorials are unsigned statements crafted by a newspaper's editorial board to advocate a particular viewpoint on a current issue (Ngai, 2022, p. 85). Editorials, as an "inherently argumentative genre" (Virtanen, 2005, p. 172), are laden with comments and evaluation that aim to assess events and persuade readers. Given the distinct discourse functions of news reports and editorials, it is expected that there will be observable differences in language use between the two sub-registers. While

previous studies (e.g., Hu et al., 2016) drawing on Press data sometimes treated these sub-registers as an integrated whole due to the difficulties in collecting and classifying texts, the author finds it important to make a distinction between news reports and editorials to give justice to their heterogenicity in terms of their functions and language use.

3.1.2 Data collection and preparation

To tap into existing corpus resources, the Press sub-corpus of the established CLOB corpus is chosen to represent native English news writing (NE). The CLOB corpus, following a similar sampling frame of FLOB, is built to reflect more recent linguistic data for contrastive studies on language use in terms of diachronic change and regional variation (Xu & Liang, 2013). Press is one of the four genres covered. Within the Press sub-corpus, newspaper articles are further categorized into three sub-registers, namely, reportage, editorials, and reviews. A total of 255 texts are available, with 65 editorials and 136 reportage articles. The remaining review articles are excluded from the current study due to challenges in identifying this specific register in the counterparts of EFL and translation. In total, 201 text files are included in the current study, with each text file containing one piece of news article. Metadata, such as article names, publisher names, publication years, and writers' names, are recorded when available. The majority of the texts in CLOB was published in 2009 or one year before or after 2009.

EFL writing samples were selected from *China Daily*, the first national English-language daily newspaper in China. The 2009 archive was chosen to align with the period covered by CLOB. The texts were randomly selected from each topic classified by Factiva, with the number of texts in each topic reflecting the proportion of each category. These news articles cover a wide range of topics, including international relations, arts and entertainment, economic growth, education, and health, to name a few. Importantly, only articles specifying reporters with Chinese names were selected, presuming these English reports are EFL writing produced by native Chinese speakers. In total, 191 pieces of news articles were selected, including 42 editorials and 149 news reports. Each piece of news

article was saved as a single text file. Metadata including article names, writer names, and publication dates were recorded.

The translation corpus (TE) includes data from *Global Times* and *Sixth Tone*, with each contributing to about half of the size of the translation corpus. *Global Times*, a renowned state-run news tabloid, launched its English edition in 2009. The English translations can be traced back to their Chinese source in the original edition (Y. Liu & Li, 2022). *Sixth Tone* is an emerging online magazine established in 2016 by Shanghai United Media Group. Despite its later establishment, it has gained international influence, with the mainstream outlets such as the BBC often citing *Sixth Tone* as their source when reporting on Chinese social stories (Ni, 2018). This online magazine devotes a section called *Sixth Tone X* to translations from respected Chinese and international media outlets, and its other sections also include translations of timely reports and contributions from experts and commentators. As prominent English press outlets in China, *Global Times* and *Sixth Tone* provide a broad range of discussion of current issues and personal viewpoints. The translation corpus covers the period from 2017 to 2022, slightly different from the other two sub-corpora in the study. A total of 140 translated news articles were selected, including 88 editorials and 52 news reports. Each piece was saved as an individual text file. Metadata recorded include article names, publication dates, and translators' names when applicable. An overview of the corpus composition is provided in Table 3.1.

In preparation for the multidimensional analysis, individual texts were merged to form longer text units, each comprising between 2,000 to 2,800 running words. This practice aims to mitigate the analytical issues often encountered with shorter texts, which aligns with the common practice in MDA research (Kruger & Van Rooy, 2016a; Hu et al., 2016). This process resulted in a total of 168 consolidated text samples: 52 text samples for EFL, 65 for NE, and 51 for TE. Each text sample encompasses the work of multiple writers or translators, which help alleviate the potential impacts of individual idiosyncrasy.

Table 3.1 Descriptive information about the corpus.

	EFL		NE		TE	
Sub-register	editorials	reports	editorials	reports	editorials	reports
No. of texts	42	149	65	136	88	52
Sub-register size (tokens)	32,606	89,715	54,643	89,342	54,441	66,141
Mean size	776	602	841	657	618	1272
Standard deviation	215.3	235.6	482.2	385.3	165.2	662.1
Total size	122,321		143,985		120,582	

3.1.3 Corpus constraint matrix

Applying Kotze's constraint model (Kotze, 2022) to the corpus data representing translated English, EFL and the reference native English, the comparative constraint matrix is made to indicate how each variety is positioned along the five constraint dimensions (Table 3.2).

Table 3.2 Constraint matrix for EFL, NE, and TE.

	EFL	*NE*	*TE*
Language activation	Bilingual activation (Mandarin Chinese/English)	Monolingual activation	Bilingual activation (Mandarin Chinese/English)
Modality and Register	Published news articles	Published news articles	Published news articles
Text production	Independent text production	Independent text production	Mediated text production
Language Proficiency	Proficient users	Proficient users	Proficient users
Task expertise	Professional	Professional	Professional

The primary distinguishing constraint between the two constrained language varieties (TE and EFL) and non-constrained native English is language activation. EFL writing and translation both involve texts produced by bilinguals, and their L1/source language (in this case, Mandarin Chinese) is activated even when monolingual English texts are produced.

Conversely, language activation predominantly remains monolingual in native English production, although individual variability can occur. The major constraint dimension that sets translation apart from the other two is the mode of text production. Different from EFL and native English writing, translation is a typical form of mediated text production, depending on a pre-exiting source text. In terms of modality and register, all three sub-corpora consist of published news articles, encompassing two sub-registers, news reports and editorials. It should be acknowledged that all published writing undergoes an editing process, which likely varies across different news agencies and cultures. Importantly, even though a specific register may carry a singular name in different cultures, standards and expectations for the text type can differ significantly based on the cultural context and the news agency involved. Lastly, writers and translators employed by these news agencies are generally considered professionals who are proficient English users, meaning their professional writings and translations are expected to contain few grammatical errors or outright mistakes, though individual variability may still exist.

3.2 A Multidimensional approach on textual variation

Multidimensional Analysis (MDA) is an analytical method that enables the exploration and interpretation of data across multiple variables (Sardinha & Pinto, 2019). It involves dimension-reduction statistical techniques such as factor analysis, cluster analysis, and discriminant analysis, which allow the identification of co-occurring patterns among variables. Since its foundational work by Biber (1988), MDA has been extensively employed in language variation studies across a range of disciplines, including sociolinguistics, discourse analysis, translation studies, and second language acquisition (see Goulart & Wood, 2021). Researchers have used this method either additively, building on existing dimensions such as those outlined by Biber (1988), or by conducting a novel MDA to extract dimensions that best suit their data. In the current study, a novel MDA is employed to examine the constrained language use in translated English and EFL, and this choice is primarily based on the following considerations.

Firstly, MDA inherently acknowledges the multi-faceted nature of language. It appreciates that language dimensions are not isolated features but rather function in synchrony, manifesting themselves in simultaneous patterns. This perspective aligns with the study's aim to explore the complexity of translated English and EFL where numerous factors are at play to shape the final linguistic output.

Methodologically, MDA enables an effective combination of quantitative analysis with functional interpretation. The statistical techniques offer a robust methodology to expose underlying structures within the linguistic data, illustrating how the linguistic features group and interact with each other. The statistical patterns are then interpreted in light of the communicative functions of these features, providing a substantive and nuanced understanding of language use.

Lastly, MDA serves as a systematically rigorous approach to examine language variation, facilitating direct comparison across diverse studies. This enables researchers to contrast results and insights, which help foster a comprehensive understanding of the examined linguistic phenomena. Given the intricate nature of constrained communication, a single study can only offer a specific viewpoint, restricted by the constraint dimensions it can consider. Therefore, employing a stringent method like MDA is critical, not just for the reliability of the current research, but also for the comparability and integration of future investigations. This systematic approach enhances the potential for meaningful synthesis of findings across studies, thus contributing to a more holistic understanding of the field.

Overall, given its holistic perspective on language, the combination of quantitative methodology with functional interpretation, and rigorous systematic approach, MDA stands as a pertinent choice for exploring the intricate panorama of constrained language use in translated English and EFL.

3.2.1 Major procedures of MDA

Multidimensional analysis comprises several interconnected steps, which can be delineated into two major phases (Biber, 1988). The initial phase commences with the selection of pertinent linguistic features that align with the research objectives. Subsequently, the data is annotated or tagged using appropriate computational tools, laying the groundwork for the analysis. Upon completing the preparatory phase, the analysis enters the second stage of the actual execution of statistical methods to extract meaningful patterns from the data. The following section first introduces the procedures of selecting, retrieving, and normalizing of linguistic features, and then outlines subsequent steps for the statistical implementation of MDA.

3.2.2 Feature selection, retrieval, and normalisation

3.2.2.1 Linguistic feature selection

The integrity and accuracy of multidimensional analysis in corpus-based studies are contingent on the meticulous selection of linguistic variables and the proficient extraction of these variables through automated inquiries, considering the vast volumes of data involved. As underscored by Dayter (2018: 257), to circumvent an excessive interpretation of superficial statistical data, it is imperative to conceptualize investigations 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".

Following the Biberian approach, the selection of linguistic features in this study is driven by theoretical and functional linguistics considerations. This selection draws on prior research (Biber, 1988, p.223-245; Lu, 2010, p.479; Hu et al., 2016, p.34-35) and aims for comprehensiveness to encompass potentially salient aspects of linguistic variation pertinent to constrained communication. The analysis comprises a total of 69 carefully chosen linguistic features across lexical, syntactic, and textual levels (Table 3.3). These

include part-of-speech classes, verb tense and aspect, semantic categories of verbs, noun modification structures, most frequent lexico-grammatical constructions, and general textual properties such as lexical density and lexical diversity, among others.

Most of the selected linguistic features stem from Biber's seminal work (1988) originally intended to differentiate between spoken and written English. Subsequent studies have widely adopted this feature list to explore variations in other types of English registers, such as translated English, non-native indigenized varieties of English, L2 student written registers, and web discourse (Hu et al., 2016; Kruger & Van Rooy, 2016a; Berber Sardinha, 2018). The broad applicability of these features underscores their relevance in analyzing specialized domain discourses in English. Other features are informed by relevant studies on translation features such as Hu et al. (2016). Structures related to verbs and nouns are prominent features incorporated in the analysis, and this choice is informed by previous research on constrained language use. For instance, Ivaska et al. (2022:152) claimed that the tendency for phrasal versus clausal elaboration identified in different constrained language varieties "could be further investigated focusing on nouns and verbs, and structures around them." Particularly, noun modification structures such as appositive modification and clausal modification are relevant in the context of Chinese-English language pair as reviewed in Chapter Two.

Table 3.3 69 selected linguistic features for MDA.

(A)	**General Text Properties**		(H)	**Prepositions**	
1	AWL	Average word length	37	IN	Prepositions
2	TTR	Lexical diversity	(I)	**Adjectives**	
3	LDE	Lexical density	38	JJPR	Predicative adjectives
(B)	**Verb Semantics**		(J)	**Pronouns**	
4	ACT	Activity verbs	39	PIT	*It* pronouns
5	ASPECT	Aspectual verbs	40	QUPR	Quantifying pronouns
6	CAUSE	Facilitation and causative verbs	41	PP1	First person pronouns
7	COMM	Communication verbs	42	PP2	Second person pronouns
8	EXIST	Existential or relationship verbs	43	PP3	Third person pronouns
9	MENTAL	Mental verbs	(K)	**Adverbials**	

10	OCCUR	Occurrence verbs
11	DOAUX	*Do* auxiliary
(C)	**Verb Features**	
12	CONT	Verbal contractions
13	PEAS	Perfect aspect
14	PROG	Progressive aspect
15	RP	Particles
16	VBD	Past tense
17	VBG	Non-finite *-ing* verb forms
18	VBN	Non-finite *-ed* verb forms
19	VPRT	Present tense
20	PASS	All *be* and *get* passives
(D)	**Modal Verbs**	
21	MDCA	Modal *can*
22	MDCO	Modal *could*
23	MDMM	Modals *may* and *might*
24	MDNE	Modals *ought*, *should,* and *must*
25	MDWO	Modal *would*
26	MDWS	Modals *will* and *shall*
(E)	**Stative Forms**	
27	BEMA	*Be* as main verb
28	EX	Existential *there*
(F)	**Coordinators and Conjuncts**	
29	CONC	Concessive conjunctions
30	COND	Conditional conjunctions
31	CUZ	Causal conjunctions
32	ELAB	Elaborating conjunctions
33	CC	Coordinating conjunctions
(G)	**Determinatives**	
34	DEMO	Demonstrative pronouns and articles
35	DT	Determiners
36	QUAN	Quantifiers
44	PLACE	Place adverbials
45	TIME	Time adverbials
46	ADVMOD	Non-clausal adverbs or adverbial phrases
(L)	**Negation**	
47	XX0	Negation
(M)	**Noun Semantics**	
48	NNP	Proper nouns
49	NOMZ	Nominalizations
50	NCOMP	Noun compounds
51	NN	Total nouns
(N)	**Noun Modification**	
52	AMOD	Adjectival modifiers
53	POSS	Possessive modifiers
54	PPMOD	Preposition phrase postmodifiers
55	APPOS	Appositional modifiers
56	NUM	Numeric modifiers
57	RC	Finite relative clauses
58	NFRC	Non-finite relative clauses
(O)	**Lexis**	
59	COMPAR	Comparatives
60	SUPER	Superlatives
61	AMP	Amplifiers
62	DWNT	Downtoners
63	EMPH	Emphatics
64	HDG	Hedges
(P)	**Syntax**	
65	SPLIT	Split auxiliaries and infinitives
66	THATD	Subordinator *that* omission
67	CSUBJ	Clausal subjects
68	ADVCL	Adverbial clauses
69	CCOMP	Complement clauses

3.2.2.2 Linguistic feature retrieval and normalisation

To retrieve the selected linguistic features and normalise their frequencies, the Multi-Feature Tagger of English (MFTE) is employed in the current study. It is a novel automatic tagger built for multi-feature analysis of linguistic variation in English. This tool is readily accessible on GitHub (Le Foll, 2021).

MFTE is designed to comprehensively capture a wide array of grammatical, lexical, and semantic features essential for conducting multivariable analysis of linguistic variation in English. Inspired by simplified Hallidayian system networks, it encompasses three sets of features: the Simple Tagset, comprising 74 core features; the Extended Tagset, comprising 64 semantics-related lexical and syntactic features; and the Extended Composite Tagset, containing 23 composite categories derived from the two prior tagsets. In total, the MFTE is able to tag and retrieve 161 linguistic features. In the final MDA analysis, 58 features were retained. The remaining 11 features required for the analysis were retrieved from texts tagged by Stanford Parser and its R implementation provided in Lu (2010). Since the different ranges and distributions of the selected features will affect the multivariate analysis, the obtained normalised frequencies were standardized to a mean of 0.0 and a standard deviation of 1.0 as z-scores following the procedures expounded in detail in Biber (1988: 93–94) in preparation for the multidimensional analysis.

It is acknowledged that the Biber Tagger is the most commonly utilized tagger in MDA studies due to its capacity to identify Part-of-Speech tags alongside semantic information (Goulart & Wood, 2021). However, a broader application of MDA in the research community is hindered by its inaccessibility to researchers unaffiliated with Biber. When the Multidimensional Analysis Tagger (MAT), a replication of the 1988 version of the Biber Tagger, was developed and made accessible by Nini (2019), there is an increase in MDA analyses conducted outside Biber's home institution. Developed on the basis of MAT, MFTE is preferred and chosen for the current analysis for two main reasons.

Firstly, a significant distinction between MFTE and previous taggers lies in their approaches to feature identification. While MFTE shares many similarities with the core feature portfolio of the Biber Tagger, it stands out by not requiring (semi-)manual annotation for certain features as seen in Biber's work (1988, 2006). This strategic design aims to strike a balance between incorporating an exhaustive and principled set of linguistic features while ensuring the tagger's ability to automatically retrieve these features reliably. The improvement in MFTE's feature identification can be attributed not only to the more sophisticated tagging facilitated by the Stanford Parser which serves as the first layer of automatic tagging since the advent of MDA studies in the 1980s, but also to its distinct operationalization of problematic elements (e.g., highly multifunctional items like *just*, *most*, and *really*) and unsatisfactory categorizations (e.g., modal verbs with diverse meanings and contextual uses) present in previous taggers. An extensive evaluation process involving human annotators manually checking over 30,000 tags and addressing problematic instances concludes that MFTE achieves an impressive overall accuracy of 96.17% (Le Foll, 2021, p.34).

Another key aspect that sets MFTE apart is that it adopts different normalisation baselines for linguistic features. Traditionally, corpus linguistics follows a word-based normalisation approach, dividing raw counts by the total number of tokens or words in the text. The Biberian MDA approach also follows this tradition using word-based normalised frequencies. However, this approach may inadvertently "conflate frequency of use and opportunity of use" (Le Foll, 2021, p.20), as it does not consider the limited choices language users have once they choose a particular word. For instance, Biber (1988) observed a high positive correlation among verbal contractions, negation, and present tense - three structures whose frequencies are influenced by the use of verbs. Normalising frequencies based on words may yield results that primarily reflect the number of verbs per 1,000 words, introducing a risk of bias. To address this concern, MFTE takes a different approach and calculates feature frequencies based on distinct normalisation baselines to "model the actual choices that language users make when producing language" (Le Foll, 2021, p.20). This innovative method ensures the tagger's ability to accurately capture

language users' actual choices, providing a more precise and meaningful analysis of linguistic variation. The normalisation baselines for the linguistic features adopted in the current study could be found in Appendix 1.

3.2.3 Implementation of MDA

Once feature selection, retrieval, and normalisation are completed, the implementation of a novel MDA first requires the decision on the specific statistical technique for the analysis. One of the most commonly adopted techniques is factor analysis, which allows for the unveiling of the latent patterns within variables by deriving a concise set of related variables known as factors. As the present study aims to investigate the textual characteristics of constrained language use through an extensive array of linguistic variables across diverse language levels, factor analysis proves to be the most suitable choice.

There are two common types of factor analysis, exploratory factor analysis (EFA) and principal components analysis (PCA). There is a conceptual difference between the two. PCA analyzes all variance, whereas EFA only focuses on covariance, that is the shared variance among variables (Tabachnick et al., 2013). For this reason, EFA is often adopted when there are no specific expectations concerning the number and nature of underlying factors in the data, while PCA is employed when there are preconceived hypotheses regarding the presence and characteristics of underlying factors in the data. EFA is "considered a high-quality decision" when the study aims to comprehend the underlying structure of a set of variables, otherwise PCA is more suitable when the purpose is "pure reduction of variables" (Conway & Huffcutt, 2003, p.150). Given the exploratory nature and the major objective of the study, EFA is the preferred method.

An EFA in an MDA study generally consists of the following steps (Egbert & Staples, 2019): (1) validating statistical assumptions for EFA; (2) executing EFA; (3) interpreting the extracted factors; and (4) computing factor scores for each text and getting the mean score of each text type. This process entails several subjective decisions by the researcher,

which include: (a) the method used for factor extraction; (b) the number of factors to be extracted; (c) the rotation method; (d) the method to calculate factor scores; and (e) the interpretation of dimensions. These decisions significantly influence the outcomes of the analysis. Detailed justification and elaboration of these decisions within the context of this study are provided in this section.

Before initiating an EFA, the first step is to assess the factorability of the data. Primarily, variables should exhibit a linear relationship and moderate correlations. The sample size is also a key consideration as correlations can be sensitive to the sample size. Various estimates suggest that the number of subjects or items should range between 3 to 20 times to the number of variables examined (Tabachnick et al., 2013; Thompson, 2004). In the current study, 69 linguistic variables are included across a sample of 168 texts. Despite being less than ideal, the sample size does not invariably compromise the accuracy of factor solutions or correlations, as a large sample size is not always necessary (MacCallum et al., 1999). An alternative approach to verify the suitability of a sample is to conduct post hoc tests such as the Kaiser-Meyer-Olkin (KMO) measure of sampling adequacy. Spanning from 0 to 1, higher KMO values suggest superior sampling adequacy, and a value less than 0.50 generally implies that the matrix is unfit for factoring (Kaiser, 1974). As shown in Table 3.4, the current data yielded a mediocre KMO score of 0.634, signifying its suitability for EFA.

In parallel with determining an appropriate sample size, it is critical to examine the correlations and communalities amongst the variables being considered for EFA. Bartlett's Test of Sphericity can be employed to identify any undesirable low correlations. This test assesses whether correlations between variables significantly deviate from 0 (Field, 2009). A significant result with $p < 0.05$ suggests that the variables are sufficiently correlated for EFA. As shown in Table 3.4, the current data yielded a significant Bartlett's Test of Sphericity result ($p = .000$), confirming that the variables are correlated and suitable for EFA.

Table 3.4 Suitability test for factor analysis.

KMO and Bartlett's Test	
Kaiser-Meyer-Olkin Measure of Sampling Adequacy.	0.634
Approx. Chi-Square	6975.424
Bartlett's Test of Sphericity df	2346
Sig.	0.000

Next, the statistical method for factor extraction needs to be determined. Various methods are available, including principal components analysis, unweighted least-squares method, maximum-likelihood method, principal axis factoring, and image factoring, each with its unique advantages and limitations. While some methods are suitable for normally distributed data, such as maximum-likelihood method, others cater to non-normally distributed data, such as principal axis factoring. Biber (1988:82) advocated principal axis factoring (PAF), since this method maximizes shared variance among variables while minimizing the number of factors. Furthermore, the solutions produced are found to be more accurate, making them preferable in contemporary social science research. Thus, this study employs principal axis factoring for factor extraction.

Determining the number of factors to be extracted is one of the primary challenges in EFA, since this decision considerably affects the interpretation of the results. Various techniques could guide the decision-making process to identify the optimal number of factors. For instance, Kaiser's rule recommends retaining factors with eigenvalues greater than 1.0 (Kaiser, 1960). These eigenvalues denote the proportion of variance accounted for by each factor, hence a higher eigenvalue signifies a factor accounting for more variance. Eigenvalues could also be visualized in a descending order by a scree plot. The "elbow" or inflection point on this plot often indicates the cutoff point for selecting factors (Cattell, 1966). However, identifying this inflection point can be subjective, making the interpretation of scree plots potentially challenging. Due to such ambiguity, it is recommended to consider the scree plot alongside other factor retention criteria, such as the parallel analysis (Donavan et al., 2007) or setting a threshold for cumulative variance explained (Field, 2009; Plonsky & Gonulal, 2015).

Factor rotation is an essential step in EFA, aiming to simplify and clarify the interpretation of data by optimizing the correlations between variables and factors. Two primary methodologies for rotation are often employed: orthogonal rotation exemplified by Varimax; and oblique rotation represented by Promax. The major difference between the two is that the former minimizes the number of variables that have high loadings on each factor, while the latter allows for correlations among the factors. For this reason, Biber (1988:85) opts for the Promax method, recognizing the likelihood that underlying dimensions may correlate with each other. In alignment with this stance, the current study implements the Promax approach for factor rotation.

Following the rotation process, a rotated factor matrix is generated, illustrating the weights or factor loadings of each linguistic feature on the extracted factors. Factor loadings are "regression-like weights used to estimate the unique contribution of each factor to the variance in a variable" (Tabachnick & Fidell, 2007, p.616), ranging from -1 to 1. Variables can load positively or negatively on a dimension, reflecting its positive or negative relationship to the overall factor makeup. A high absolute value of a factor loading signifies that the linguistic feature exerts a strong influence on the factor, thereby significant in interpreting the corresponding factor. Based on this factorial structure matrix, interpretation of the factors or dimensions will be made in terms of the functions shared by the linguistic features on each factor, and subsequently mean factor scores of the language varieties across sub-registers under examination could be calculated and compared.

3.3 Univariate analysis on feature-level variation

To augment the multidimensional analysis with a detailed exploration at the feature level, this study proceeds with a micro-level examination of the distribution of 69 linguistic features across EFL, NE, and TE. Since the multidimensional analysis explores textual variations of the two constrained varieties, a focused examination could reveal additional insights into how these variations are manifested through individual features. This is done through contrasting constrained and non-constrained varieties based on frequency data of the linguistic features, similar to the prevailing unidimensional analysis in translation and

EFL studies. However, key differences exist. Unlike approaches where only one or a few linguistic features are selected to represent certain aspects of translation or EFL production, this study refrains from making pre-assumptions. Instead, it adopts a bottom-up approach, allowing the distinguishing features to emerge from the analysis. This strategy not only offers a nuanced view of EFL and TE, but also facilitates a robust comparison between the current research and preceding studies. In turn, this comparison enriches our understanding of the characteristics of constrained language use, providing new layers of insight.

Specifically, the study would examine the normalised frequencies of the selected 69 linguistic features through statistical tests to detect any statistically significant differences in the distributional patterns that could set the two constrained varieties apart from NE.

Based on these results, two primary observations can be made. First, within each dimension, it is possible to identify which linguistic features are overused or underused by TE and EFL as compared to NE. Special attention will be paid to those differences that are statistically significant. Second, these features would be further analyzed beyond the frame of the dimensions identified by the multidimensional analysis. In other words, features that are statistically overused and underused by TE and EFL in comparison to NE will be examined, which enables the emergence of further implications regarding the characteristics of constrained language use.

Prior to conducting the statistical analysis, the distribution of the 69 features must be examined to determine the appropriate statistical methods. Two specific tests of normality, the Kolmogorov-Smirnov test with Lilliefors significance correction and the Shapiro-Wilk test, were applied to the dataset. The null hypothesis for these two tests asserts that the population is normally distributed, and a p-value less than 0.05 would suggest a rejection of the null hypothesis, indicating a non-normal distribution.

According to the results of these normality tests (see Appendix 2), non-normal distributions were found for 35 linguistic features in the EFL sub-corpus, 37 in the TE sub-corpus, and 41 in the NE sub-corpus. Given over half of the linguistic features across all three sub-

corpora deviating from a normal distribution, the non-parametric Kruskal-Wallis test was chosen as the appropriate statistical method for comparing the distribution of the 69 linguistic features among these varieties. The Kruskal-Wallis test is advantageous in this context, as it makes no assumptions about the distribution of the data, thus allows for a robust comparison even when the assumption of normality is violated.

When the Kruskal-Wallis test indicates statistically significant differences ($p<0.05$) across the three language varieties, post-hoc tests are subsequently employed for pairwise comparisons to discern which specific groups differ from one another. The Bonferroni correction method is employed to control for Type I error across multiple comparisons, ensuring the conclusions drawn from the pairwise comparisons are both statistically valid and reliable. This two-step process would help identify the linguistic features that exhibit shared distributional patterns in TE and EFL in contrast to NE, as well as those that differ between TE and EFL. All normality tests, Kruskal-Wallis tests, and subsequent post-hoc tests were performed using SPSS. Detailed results of these analyses are reported in Chapter 4.

Chapter 4 Textual characteristics of constrained language: A multidimensional analysis

In this chapter, the results of the factor analysis, including the scree plot, the total variance explained table, and most importantly, the rotated factorial structure matrix, are reported, revealing the co-occurring patterns of the selected 69 linguistic features in EFL, NE, and TE. Based on the results, the interpretation of the dimensions is elaborated, and the relations of the language varieties under examination along these dimensions are analyzed in detail. The effect of sub-registers is also considered which plays an interactive role with the common constraint shared by EFL and TE. The shared and distinctive textual properties of the two constrained English varieties are summarized.

4.1 Results of exploratory factor analysis

The EFA implementation extracted 20 factors that explained 70.782% of the total variance. However, retaining all 20 factors is unrealistic and potentially less meaningful because of a lack of theoretical clarity (Biber, 1988). Meanwhile, the first two factors, as shown in Appendix 3, account for the two most substantial variance portions, 13.539% and 11.621% respectively. The third factor accounts for only 6.125%, and the sixth factor accounts for approximately 3% of the total variance, while the remaining factors contribute less. The scree plot (Figure 4.1) exhibits noticeable inflections after factor 4 and factor 6, suggesting a 4- to 6-factor solution. After considering both the cumulative variance explained and the scree plot, and exploring solutions with 4, 5, and 6 factors, the 6-factor solution was chosen to balance the need to explain as much variance as possible and to keep the number of factors manageable.

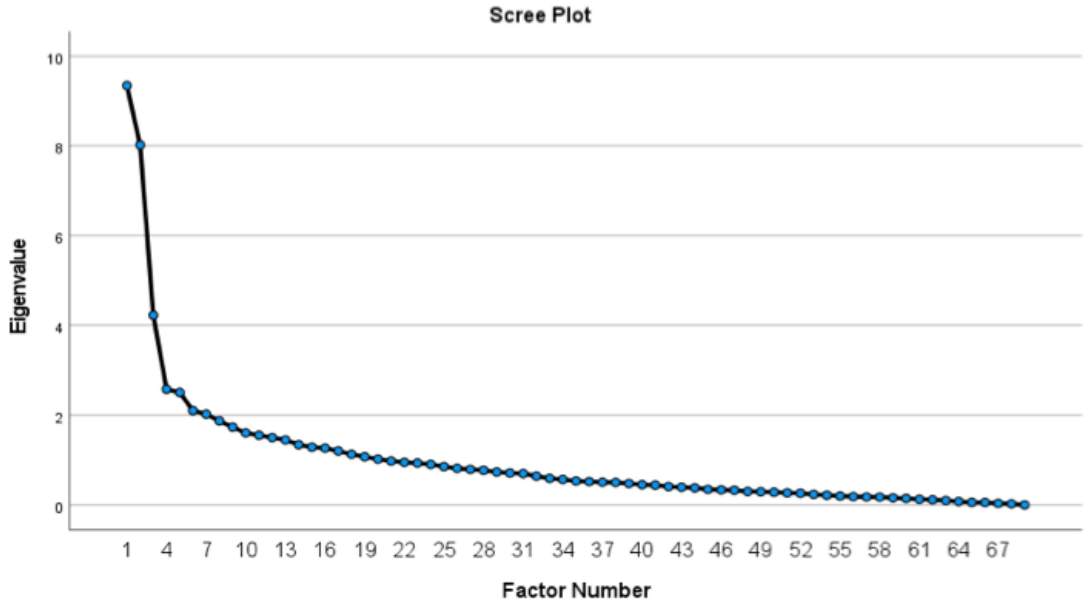

Figure 4.1 Scree plot of the eigenvalues of the 69 linguistic features.

Following this decision, the author ran a second time factor extraction, aiming to reveal the total variance explained by the six factors. Table 4.1 shows that the first six factors account for about 37% of the total variance among the three English varieties. Table 4.2 exhibits the factorial structure after Promax rotation. Six factors are listed, each associated with linguistic features that have larger-than-0.30 factor loadings.

Table 4.1 Total variance explained by the first six factors.

Total Variance Explained							
Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	9.342	13.539	13.539	8.825	12.790	12.790	7.708
2	8.019	11.621	25.161	7.557	10.952	23.742	7.196
3	4.226	6.125	31.285	3.639	5.274	29.016	5.917
4	2.579	3.738	35.023	2.006	2.908	31.924	5.373
5	2.504	3.630	38.653	1.855	2.689	34.613	2.314
6	2.103	3.048	41.701	1.451	2.103	36.716	2.293

Extraction Method: Principal Axis Factoring.

Table 4.2 Rotated factorial structure matrix.

Six factors extracted (no. of linguistic features)	Positive features	Loadings	Negative features	Loadings
Factor 1 (21)	Non-clausal adverbs or adverbial phrases	0.751	Total nouns	-0.929
	Finite relative clauses	0.712	Lexical density	-0.779
	Quantifiers	0.680	Proper nouns	-0.606
	First person pronouns	0.545	Average word length	-0.567
	Mental verbs	0.533	Modals *will* and *shall*	-0.417
	Possessive modifiers	0.529	All *be* and *get* passives	-0.339
	Verbal contractions	0.516	Occurrence verbs	-0.327
	Third person pronouns	0.505	Progressive aspect	-0.302
	Quantifying pronouns	0.491		
	Non-finite relative clauses	0.437		
	Downtoners	0.410		
	Do auxiliary	0.372		
	Second person pronouns	0.327		
Factor 2 (22)	*Be* as main verb	0.695	Past tense	-0.807
	Present tense	0.673	Communication verbs	-0.557
	It pronouns	0.650	Third person pronouns	-0.544
	Modals *ought*, *should*, and *must*	0.491	Complement clauses	-0.483
	Modal *can*	0.481	Subordinator *that* omission	-0.318
	Predicative adjectives	0.464	Prepositions	-0.317
	Conditional conjunctions	0.434		
	Negation	0.415		
	Modals *will* and *shall*	0.381		
	Split auxiliaries and infinitives	0.369		
	Demonstrative pronouns and articles	0.365		
	Emphatics	0.355		
	Superlatives	0.354		
	Nominalizations	0.308		
	Existential *there*	0.302		
	Adjectival modifiers	0.300		
Factor 3 (15)	Adverbial clauses	0.703	Proper nouns	-0.383
	Non-finite *-ed* verb forms	0.692		
	Preposition phrase as postmodifiers	0.590		

	Non-finite *-ing* verb forms	0.555		
	Prepositions	0.551		
	Concessive conjunctions	0.515		
	Adjectival modifiers	0.491		
	Existential or relationship verbs	0.454		
	Average word length	0.439		
	Elaborating conjunctions	0.408		
	Non-finite relative clauses	0.382		
	Predicative adjectives	0.325		
	Particles	0.325		
	Causal conjunctions	0.314		
Factor 4 (9)	Time adverbials	0.621	Nominalizations	-0.368
	Place adverbials	0.556		
	Hedges	0.537		
	Numeric modifiers	0.523		
	Appositional modifiers	0.408		
	Second person pronouns	0.373		
	Concessive conjunctions	0.345		
	Verbal contractions	0.332		
Factor 5 (7)	Activity verbs	0.403	Proper nouns	-0.438
	Modal *could*	0.331	Coordinating Conjunctions	-0.428
	Noun compounds	0.311	*Be* as main verb	-0.382
			Preposition phrase noun postmodifiers	-0.373
Factor 6 (7)	Lexical diversity	0.488	Determiners	-0.422
	Lexical density	0.412	Modal *would*	-0.346
	Noun compounds	0.353	Existential *there*	-0.344
	Mental verbs	0.319		

4.2 Interpretation of textual dimensions

As illustrated above, the rotated factorial structure reflects the quantitative co-occurrence of the linguistic features. To understand the functional underpinnings of these patterns, however, qualitative interpretation is required based on "the situational, social, and cognitive functions most widely shared by the linguistic features" (Conrad & Biber, 2001, p.6). It should be noted that as emphasized by Biber (1988: 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". This implies that even though the interpretation stems from a careful analysis of the shared functions of the linguistic features within each dimension, and is supported by prior research on linguistic variation, the interpretation does entail subjective decisions by the researcher. Consequently, these interpretations remain provisional and invite further scholarly verification.

When interpreting the dimensions, it is common to find that features may load on multiple factors, and different treatments could be done for features that load on more than one factor. Some studies, such as Biber (1988), retained features exclusively in the factor where they had the highest loading, while others considered all the features regardless their relative weights on different dimensions (e.g., Hu et al., 2016). The current study takes the latter approach, that is, retains all features in their respective factors as long as their loading values exceed the threshold. This decision is mainly motivated by the following reason. Theoretically speaking, it is reasonable to expect that a feature could load on multiple dimensions as it contributes to multiple textual functions (Neumann, 2014; Hu et al., 2016). Given that the factor loadings only indicate but "may not be an accurate representation of the differences among factors" (DiStefano et al., 2019, p3), it is unwise to exclude certain features based on their relative weights on different factors. Disregarding features in dimensions where they exhibit salient loadings may result in a neglect of their contributions to those particular dimensions, potentially leading to biased interpretation of the textual dimension.

During interpretation, it is a common practice to exclude linguistic variables with low loadings from the extracted factors. For the current study, a factor loading cut-off of 0.3 is adopted, meaning only features with an absolute loading value greater than 0.3 are retained. This cut-off, although slightly more lenient than the 0.35 threshold used by Biber (1988), has been frequently employed in recent MDA studies (Goulart & Wood, 2021). The decision to employ this lower threshold is justified by the relatively smaller number of text samples in our study compared to Biber (1988) and by our objective of maximizing the utility of the extracted features.

Therefore, for the following factor interpretaiton, all features are retained if their weights exceed the specificied 0.3 cut-off point, regardless of the number of factors they load on.

4.2.1 Interpretation of Factor 1

Factor 1 captures 21 features, among which 13 have positive loadings and 8 have negative loadings. On the positive pole, 8 features have loading values greater than 0.5, suggesting their high representativeness of the underlying construct. The most salient two are adverbial modifiers and relative clauses as noun modifiers. Adverbial modifiers are adverbs and adverbial phrases that provide additional information about time, place, frequency, degree, manner, reason of an action or state expressed in a sentence (Biber et al., 2002, p. 193-194). Relative clauses are finite subordinate clauses that modify a noun or noun phrase in a sentence, which are used for "explicit and elaborated identification of nominal referents" (Biber, 1988, p. 144). Such structures do not constitute core complements that are obligatory in sentence construction, but add depth and precision by conveying additional details about the actions or states being described (Nivre et al., 2020). Mental verbs convey information about "individuals' mental states, perceptions, and emotional experiences" (Biber et al., 2002, p. 107). Examples of mental verbs include *think*, *know*, *believe*, and *feel*, among others. Other positive features share a specificity focus, emphasizing specific persons, objects or quantities being referred to. Such features include quantifiers, quantifying pronouns, possessive modifiers, and first/second/third person pronouns. A shared orientation towards informality is confirmed by the reduced surface forms marked by verbal contractions and *do* as auxiliary verbs (Biber, 1988, p.106). In general, the co-occurrence of the positive features suggests a less-densely but more elaborated manner of information presentation with an involvement and interpersonal orientation.

Along the negative side, five out of eight features have loading values greater than 0.5. Representative negative features include total nouns, proper nouns, general text features (lexical density and average word length) and passive structures, pointing to informational density, formality, and objectivity. The number of total nouns reflects the "overall nominal assessment of a text" (Biber, 1988, p.228). Proper nouns are used to refer to names of

persons, places, and institutions. Both features imply that texts marked by the negative side rely heavily on nouns to convey information, potentially resulting in a denser informational presentation. Passive structures "have been taken as one of the most important surface markers of the decontextualized or detached style" and marks an "abstract presentation of information" (Biber, 1988, p.228). The presence of passive voice indicates an abstract presentation of information, with the focus on the actions or processes rather than the subjects performing them. The negative pole is also associated with more diversified vocabulary and the use of longer words. Such linguistic choices contribute to a more sophisticated and formal writing style.

Both positive and negative features exhibit similarities to Biber's (1988) Dimension 1 'Involved versus Informational Production'. In terms of the positive pole, the shared features exhibit an inclination towards involvement, as evidenced by the presence of mental verbs and first/second person pronouns. Conversely, the overlapping negative features consist of nouns, average word length, and passive structures, indicative of a formal and informative discourse style. The features also show resemblance to those on Dimension 2 identified by Kruger and Van Rooy (2018), which is represented by adverbial modifiers, private verbs, and predicative adjectives on the positive side and total nouns on the negative side. This dimension is described as 'Elaborated-involved versus Integrated-informational Presentation' in Kruger and Van Rooy (2018). Given such similarities, Dimension 1 is characterized as 'Elaborated-involved versus Integrated-formal Production'.

4.2.2 Interpretation of Factor 2

Factor 2 identifies 22 features, among which 16 load on the positive pole and 6 on the negative one. The positive side is characterized by an evaluative focus. The frequent co-occurrence of copula *be*, present verb tense, and pronoun *it* implies a discourse that focuses on topics of immediate relevance and underscores the information being presented by removing focus from any temporal sequencing (Biber, 1988, p.224). When a copula *be* is followed by predicative adjectives, the most common function is to express stance and evaluation (Biber et al., 2002, p.142 & p.188). The evaluation and stance-taking focus is

evident by other typical features on the positive side: necessity modal verbs (i.e., *ought*, *should* and *must*), modal verb *can*, predicative modal verbs (*will* and *shall*), and negations. It is not surprising to see the co-occurrence of the split auxiliaries, copula *be*, and modal verbs, as split auxiliaries are identified when the infinitive marker *to* or an auxiliary verb (modals/*do*/*be*/*have*) is followed by one or two adverbs and a verb base form. The co-occurring emphatics and superlatives also align with the pattern. Emphatics adds emphasis and intensity to the statement, and superlatives indicates the highest degree or superiority of a quality or attribute, both could function as evaluative devices that reflect the writer's assessment of their own or other's propositions (Alamri, 2023). To describe the evaluation, assessment, and stance-taking orientation, 'Evaluative discourse' is used for the positive end of Dimension 2.

On the negative pole, the co-occurrence of the high-value features (past tense of verbs, complement clauses, and communication verbs) indicates a typical discourse of reporting or retelling. Communication verbs refer to a specific subclass of activity verbs that are commonly used to describe speech and writing, such as *describe*, *tell*, *ask*, and *claim* (Biber et al., 2002, p.107). Third-person pronouns serve to mark relatively imprecise references to individuals outside the immediate interaction (Biber, 1988, p. 225). Previous research has shown that they often co-occur with past tense, functioning as markers of reported narrative (Xu, 2021, p. 120). An inclination towards informality and reduced surface form is also suggested by the negative loading of *that*-deletion in subordinating clauses. The negative side shows resemblance to Dimension 4 'Reported communication' in Biber and Egbert (2016) and Dimension 4 'Speech reporting or retelling in written registers' in Kruger and Van Rooy (2018), both characterized by high-value features such as complement clauses, communication verbs, and subordinator *that* deletion. Thus, the negative pole of Dimension 2 is described as 'Reporting/retelling discourse'.

4.3.3 Interpretation of Factor 3

Factor 3 captures 15 features, while only one feature loads on the negative side. This distribution of features suggests that the factor mainly identifies one discourse style

characterized by the positive features. The grouping of prominent positive features can be related to detailed narration and depiction. For instance, the highest-value feature is the adverbial clause, which provides information on the context where the main action or event occurs, such as condition, concession, purpose, manner, cause, and effect, among others (Biber et al., 2002, p.257). Compared to adverbs or adverbial phrases, they serve similar functions but in a more elaborated manner. Preposition phrases as postmodifiers also add extra details to the noun heads. The semantic category of existential and relationship verbs (e.g., *appear*, *exist*, *represent*) reports a state of existence or a logical relationship between entities (Biber et al., 2002, p. 109). The non-finite *-ed* verb forms and *-ing* verb forms correspond to the past and present participial forms in Biber (1988, p.232), and both could be identified in past (present) participial clauses and past (present) participial WHIZ deletion relatives. When used as participial clauses, such structures are detached in their syntactic form, and present participial clauses are often used to "create vivid images in depictive discourse" (Biber, 1988, p.109). When used in participial WHIZ deletion relatives (as captured by another positive feature: non-finite relative clauses), they offer additional information to the nouns they modify. For example, the co-occurrence of such verb forms and particles could specify the manner in which the activity is carried out or identify the location and direction of an action (e.g., *Walking down the street*, he suddenly heard a loud noise.). Elaborators are elaborating conjunctions that introduce additional information and enhance the overall coherence of a text. Such detailed and elaborated narration is echoed by other positive features including adjective modifiers for nouns, predicative adjectives, and average word length. Consequently, Dimension 3 is believed to reflect a 'Depictive and detailed narration'.

4.3.4 Interpretation of Factor 4

Factor 4 groups 9 features together, among which 8 have positive loadings and only one has negative loading, a factor structure similar to Factor 3. The most salient positive features include adverbials for time and place references, which mark a situated rather than abstract textual content (Biber, 1988, p. 224). This shows partial similarities to the positive

end of Dimension 3 ‘Oral Narration’ in Biber and Egbert (2016). Numerical modifiers and appositional modifiers for nouns also have relatively high weights. Numerical words or phrases add numerical or quantitative information to the modified nouns, and appositional structures expand information by providing extra description, identification, or explanation. Concessive conjunctions indexing a subclass of adverbial subordination are common for framing purposes or introducing background information (Biber, 1988, p.236). Hedges (e.g., *kind of*, and *maybe*) are often associated with approximation, adding nuances or caution in the expression. They are considered “informal, less specific markers of probability or uncertainty” (Biber, 1988, p.240) and occur more frequently in interactive and involved discourse. Informality and involvement are also hinted by the reduced surface form of *that*-deletion in subordinating structures and second person pronouns. The only negative feature is nominalization, which often occurs in texts that have an informational and abstract focus (Biber, 1988, p.227). Given that it is the only negative feature, Factor 4 is characterized by an absence of informational and abstract narration. Overall, Dimension 4 represents a narrative discourse that focuses on the spatial-temporal context, thus is labelled as ‘Descriptive narration with a spatial-temporal focus’.

Before moving into the interpretation of the remaining two factors, it should be noted that there are fewer linguistic features grouping on both factor 5 and factor 6 compared to previous ones, and each only has three features whose loading values are greater than 0.4. These two factors are retained in the hope to obtain as much as textual information as possible from the factor analysis, and this decision is justified by the smaller sample size in the current study compared to Biber (1988), but it is well acknowledged that any interpretation of these underlying dimensions is largely tentative.

4.3.5 Interpretation of Factor 5

Factor 5 consists of seven features, with three on the positive side and four on the negative one. The highest-value positive feature is activity verbs, which often refer to a volitional action performed intentionally by an agent (Biber et al., 2002, p.106). Examples of activity verbs include *work*, *bring*, and *come*. Following activity verbs are modal verb *could* and

noun compounds. The main function of modal verbs is associated with stance. Especially, the modal verb *could* is a versatile modal that is able to mark permission, ability and possibility (Biber et al., 2002, p.176). Noun compounds are combinations of two or more nouns appearing consecutively without the inclusion of any function word in between (Biber et al., 2002, p.273). Such noun plus noun sequences are found to be especially common in news writing "where they help to pack a lot of information into a small space" (Biber et al., 2002, p.92). Features that cluster on the negative pole include proper nouns, coordination conjunctions, copula *be*, and prepositional phrases as post-modifiers for nouns. The most common coordinating conjunctions include *and*, *or*, and *but*, etc. (Biber et al., 2002, p.227). Such coordinators could link words, phrases, and independent clauses, creating a cohesive and unified structure in a text. Preposition phrases are common noun postmodifiers. Compared to other postmodifiers such as relative clauses, such structures "occur in extremely dense, embedded sequences" (Biber et al., 2002, p.269) and are frequent in academic writing and news prose. Dimension 5 is decided to potentially mark 'Activity focus versus Referential precision'.

4.3.6 Interpretation of Factor 6

Factor 6 groups four positive features and three negative features. The positive features include general text characteristics such as lexical density and type-token ratio. Lexical density is calculated as the ratio of content words to the total number of words in a text. Higher lexical density suggests a greater level of specificity and information density. The type-token ratio measures the proportion of unique word forms. A higher TTR suggests more diversified use of vocabulary. Information density is also exemplified by the co-occurring noun compounds, or noun plus noun sequences (Biber et al., 2002, p.273). This structure condenses only content words and omits explicit markers that specify the logical relationship between the constituents. Thus, discourses characterized by rich compound nouns are often associated with increased information density and abstraction. The negative pole encompasses determiners, modal verb *would*, and existential *there*. Modal verb *would* often marks volition and prediction (Biber et al., 2002, p. 181), associated with certain

situation or action that is not known to have happened. Existential *there* is often used to introduce new topics that are going to be the focus of the following discourse (Biber et al., 2002, p. 418). Dimension 6 is thus tentatively labelled as 'Information density versus Irrealis'.

Table 4.3 summarizes the interpretation of the six factors extracted. The first two factors are the most salient ones, capturing 21 and 22 linguistic features respectively. Dimension 1 showcases a contrast between elaborated-involved and integrated-formal production styles, while Dimension 2 presents a contrast between evaluative discourse and reporting/retelling discourse. The subsequent two factors exhibit a similar factor structure, with the majority of features loading on the positive pole. Dimension 3 is characterized by a depictive and detailed narration, and Dimension 4 emphasizes a descriptive narration with a spatial-temporal focus. The final two factors are less prominent, featuring fewer loaded features on each factor. Dimension 5 is tentatively associated with activity focus versus referential precision, while Dimension 6 highlights a contrast between information density and the use of irrealis. These dimensions provide valuable insights into the diverse textual styles and narrative orientations found in the analyzed language varieties.

Table 4.3 Summary of textual dimensions.

Dimension 1	Elaborated-involved versus Integrated-formal production
Dimension 2	Evaluative discourse versus Reporting/retelling discourse
Dimension 3	Depictive and detailed narration
Dimension 4	Descriptive narration with a spatial-temporal focus
Dimension 5	Activity focus versus Referential precision
Dimension 6	Information density versus Irrealis

4.3 Textual variations of constrained language

To observe the textual variations of the two constrained varieties of English, i.e., EFL and TE, in contrast to non-constrained native English along the identified dimensions, factor scores need to be calculated for each text in the three sub-corpora. Then the mean scores of the three sub-corpora can be calculated and compared, to identify the similarities and

differences among the three varieties. In this section, the method for factor score calculation is first provided and then based on this method, the results of the calculation are provided including the descriptive statistics of the factor scores for the three language varieties across the two sub-registers.

4.3.1 Factor score calculation

For factor score calculation, the current study retains all features whose values are greater than the threshold. As elaborated in the dimension interpretation procedure, it is meaningful to retain features in all the dimensions they appear, since they may provide valuable information for understanding the latent functional motivations driving each dimension. On top of this reason, this decision for factor score calculation is also justified by the calculation method adopted in the study. Unlike many studies which calculated factor scores by simply summing up the z-scores of the features loading on a factor (Biber, 1988; Goulart & Wood, 2021), this study obtains factor scores by using the Bartlett's approach, a refined statistical method for dimension score calculation (DiStefano et al., 2019). The Bartlett's approach for factor score estimation uses a regression model that minimizes the difference between the predicted and unique factor scores (Hershberger, 2005). This approach offers several advantages, as it considers both the loadings of the linguistic features and the relations among features, leading to more accurate estimations of the underlying latent dimensions (DiStefano et al., 2019; Hershberger, 2005). In other words, this method recognizes the relative importance of individual features in relation to specific factors, rather than treating all features equally.

Hence, all features are kept if their weights surpass the predetermined 0.3 threshold in factor score calculation, irrespective of the number of factors they are associated with. Factor scores are calculated using the Bartlett's approach, and the computation is automatically performed using SPSS.

4.3.2 Factor score comparison

Table 4.4 lists the descriptive statistics of the factor scores for EFL, NE, and TE along six dimensions. Complementing to the descriptive statistics table, boxplots are presented to visualize the distribution of the factor scores of the three varieties (Figure 5.1, and dimension statistics across the two sub-registers is presented in Appendix 4).

Table 4.4 Descriptive dimension statistics for EFL, NE, and TE along six dimensions.

	D1			D2			D3		
	EFL	NE	TE	EFL	NE	TE	EFL	NE	TE
Median	-0.01	-0.03	-0.39	-0.12	-0.09	0.11	-0.47	0.00	-0.27
MIN	-2.44	-2.68	-1.26	-1.47	-1.99	-2.31	-2.18	-2.39	-2.09
MAX	2.67	3.60	3.49	2.81	2.20	1.69	3.02	2.52	2.50
Q3	0.61	0.48	0.49	0.36	0.89	0.94	0.72	0.63	0.60
Q1	-0.67	-0.60	-0.69	-0.66	-0.82	-0.79	-0.74	-0.59	-0.53
IQR	1.28	1.08	1.18	1.02	1.70	1.73	1.46	1.21	1.13
	D4			D5			D6		
	EFL	NE	TE	EFL	NE	TE	EFL	NE	TE
Median	0.21	-0.21	-0.04	0.02	-0.24	-0.05	0.04	0.17	-0.30
MIN	-2.50	-1.61	-1.56	-1.96	-2.12	-1.92	-2.33	-2.68	-1.88
MAX	2.69	5.81	2.09	2.37	5.44	2.21	3.15	2.31	2.64
Q3	0.84	0.25	0.53	0.65	0.28	0.54	0.78	0.70	0.67
Q1	-0.95	-0.54	-0.72	-0.54	-0.72	-0.60	-0.93	-0.66	-0.81
IQR	1.79	0.79	1.25	1.19	1.00	1.15	1.71	1.36	1.48

Key descriptive features could be observed from the boxplots, such as the median (the median line in each box), the lower/upper quartile (lower/upper ends of the box), range of the data beyond the IQR (the vertical whisker extending from the box), and the potential outliers (dots). The height of the box reflets the distance between Q1 and Q3, also called interquartile range (IQR), which captures the middle 50% of the data. The vertical extreme line indicates the highest and lowest value excluding outliers (Q3+1.5*IQR to Q1-1.5*IQR). Jitters beyond the extreme line demonstrate potential outliers. Given the presence of outliers, the IQR seems to be a more meaningful measure for the variability or the spread of the data. Compared to standard deviation and range, they are less sensitive to extreme values or outliers. A larger IQR indicates a greater spread in the data. Asymmetric IQR with one end longer than the other may indicate skewness.

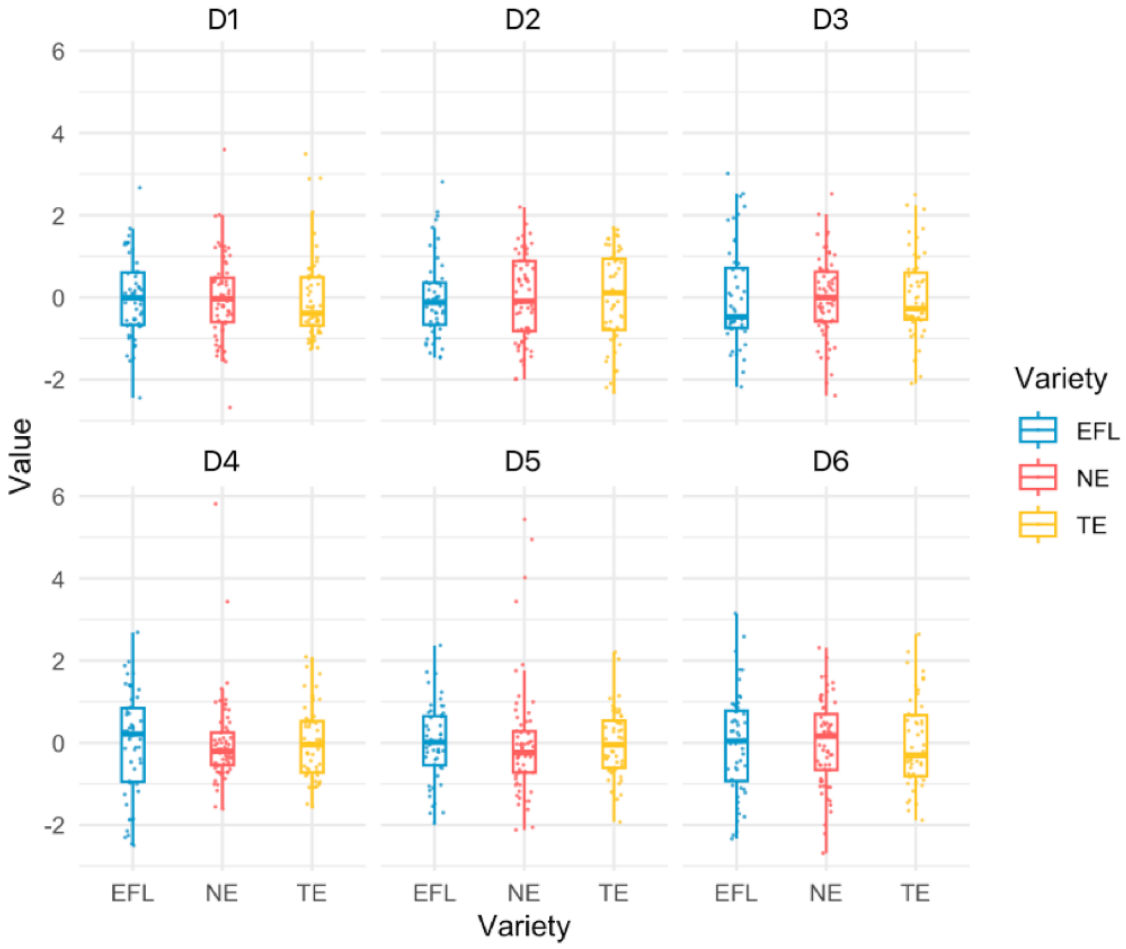

Figure 4.2 Boxplots of factor score distribution for EFL, NE, and TE along six dimensions.

To our surprise, the mean values of factor scores for the three varieties are all equal to zero in the six dimensions. There are several potential reasons for this outcome. One is that the data has undergone a standardization procedure that results in mean scores being centered around zero, which is a common practice for a factor analysis. However, a trial analysis using non-standardized normalised frequencies yielded similar results. Another possibility is that the sub-corpora share an identical sample structure, meaning that variations between the two sub-registers neutralize each other in every dimension. To investigate this possibility, regression analysis is conducted (Section 4.3.3) to study the impact of register. Additionally, it should also be noted that the mean value, as a measure of central tendency, is sensitive to outliers. As indicated by the boxplots (Figure 4.2), outliers (dots falling out

of the vertical extreme line of boxplots) could be observed in all three varieties across six dimensions. In this case, the median score is a more robust statistic for the central tendency, which is reported in Table 4.4. Therefore, the author proceeds with the analysis using the median scores of the three sub-corpora.

A few observations could be drawn from the descriptive statistics. Regarding Dimension 1, EFL and NE exhibit nearly identical median scores, both higher than that of TE. This suggests the TE tends to lean towards the negative end of the dimension, characterized by an integrated-formal discourse. Additionally, the box heights of the three varieties suggests that text samples cluster around the median to a similar degree. Moving to Dimension 2, there is a noticeable difference in terms of the within-group variability: both NE and TE display greater variability compared to EFL. While the median of TE is slightly above zero, the median values of other two are negative. This indicates that TE shows a slight inclination towards the positive end of Dimension 2, which is associated with evaluative discourse. Regarding Dimension 3, both TE and EFL exhibit median scores lower than NE, marking a lack of depictive and detailed narration in constrained varieties. As for Dimension 4, both EFL and TE demonstrate greater within-group variability than NE. EFL has a higher median value, which indicates that it is more marked by a descriptive narration with a spatial-temporal focus. Turning to Dimension 5, NE exhibits a slightly lower median score compared to TE and EFL, suggesting its mild inclination towards the negative end of the dimension marked by more emphasis on referential precision. Lastly, Dimension 6 shows TE has the lowest median score, which is lower than 0, indicating its inclination towards more use of irrealis or decreased level of information density.

However, viewed from a holistic perspective, no consistent patterns could be identified in terms of the contrast between constrained and non-constrained varieties. In other words, the textual dimensions identified seem to be unable to effectively distinguish the constrained language varieties from the non-constrained one.

Before concluding that constraints shared by TE and EFL production have a negligible effect that leads to no observable or consistent textual-level differences between

constrained and non-constrained varieties, it is crucial to consider the possibility that the various constraints involved may interplay and at times have counterbalancing effects. Especially as explained in Chapter 3, the Press register consists of two sub-registers of news reports and editorials, which serve different discourse functions and thus may vary significantly in the language use. Therefore, a regression analysis was conducted to gain further insights into the impact of register and language variety on the textual variations of constrained varieties of English represented by TE and EFL.

4.3.3 Textual variations and register effect

4.3.3.1 Regression analysis

To explore the degree to which the textual features of the three language varieties can be attributed to differences in register and variety, a Generalized Linear Model (GLM) is fitted for each dimension, with the factor scores of individual text samples as the dependent variable. Each model features two categorial predictors: VARIETY and REGISTER, the former comprising three levels (EFL, NE, and TE) and the latter consisting of two levels ("ed" for editorials and "report" for news reports).

To account for the possibility of an interaction between the two predictors, two models were initially considered: one with an interaction term, and an alternative without the interaction. To evaluate the goodness of fit, an ANOVA test was performed to compare the two models for each dimension. This test involves calculating the Residuals Sums of Squares (RSS) and corresponding F-statistic. The RSS quantifies the sum of the squared differences between the observed values and the predicted values. A lower RSS indicates a better fit of the model, while the F-statistic measures whether the RSS reduction is statistically significant. The results of the model comparison can be found in Appendix 5.

Among all six dimensions, the models incorporating an interaction term exhibited a lower RSS compared to the alternative models. The differences in RSS were statistically significant ($p<0.05$) for all dimensions except Dimension 2 and Dimension 6. Therefore, for all dimensions, models featuring an interaction term were retained, while it is

recognized that the improvement of fitness resulted from the interaction term is limited for Dimension 2 and Dimension 6. The regression modeling was conducted using the *lmer* package in R. The detailed summary of the resulting models for each dimension is in Appendix 6.

Table 4.5 presents the results of the regression analysis, including the F statistic (ANOVA), the p-value associated with the F statistic, and the adjusted R-squared values. The F-statistic assesses the overall significance of the regression model, and a larger value indicates a stronger relationship between the predictors (VARIETY and REGISTER) and the dependent variable (factor scores). The associated p-value indicates whether the relationship is statistically significant. The adjusted R-squared value indicates the proportion of the variation in the dependent variable explained by the predictors. These indicators collectively indicate the amount of variation in the factor scores explained by the predictors, and reflect the predicative power of each dimension in distinguishing the three language varieties across the two sub-registers in the current study.

Table 4.5 Results of regression analysis.

	F-statistic	p-value	Adjusted R*R (% variance explained)
D1	7.528	<0.001	16.35%
D2	38.59	<0.001	52.95%
D3	15.08	<0.001	29.66%
D4	15.53	<0.001	30.31%
D5	2.073	0.071	3.11%
D6	3.235	<0.05	6.27%

As shown in Table 4.5, along the six dimensions, five show a significant relationship between dimensional variation and the predictors (VARIETY and REGISTER as a whole) at a significance level of $p<0.001$. However, the strength of the relationship varies across dimensions. Notably, Dimension 2 exhibits a robust relationship, with an R-squared value exceeding 50%. Dimension 3 and Dimension 4 follow closely, with R-squared values at approximately 30%. Dimension 1 displays a moderate relationship with an R-squared value of 16.35%. For the first four dimensions, the predictors explain a substantial portion of dimension variation. On the other hand, Dimension 6 only exhibits weak discriminative power, as indicated by its R-squared value below 10%. Dimension 5 has the lowest R-

squared value of below 5%, and its p-value suggests an absence of statistically significant relationship between predictors and factor scores.

The statistics discussed above provide insights into the explanatory power of the overall model, considering the predictors collectively rather than individually. To better illustrate the relative strength of the two predictors, effect-size plots (Figure 4.3 – Figure 4.8) are presented for each dimension model. These plots visually present the estimated impact of one independent variable on the dependent variable, while holding the other independent variable constant. By presenting the information graphically, it becomes easier to interpret the direction and magnitude of the predictors' impact. The effect size plots were generated using the *Effects* package in R. In the following section, the results are briefly discussed with a focus on the relative strength of the two predicators along the six dimensions.

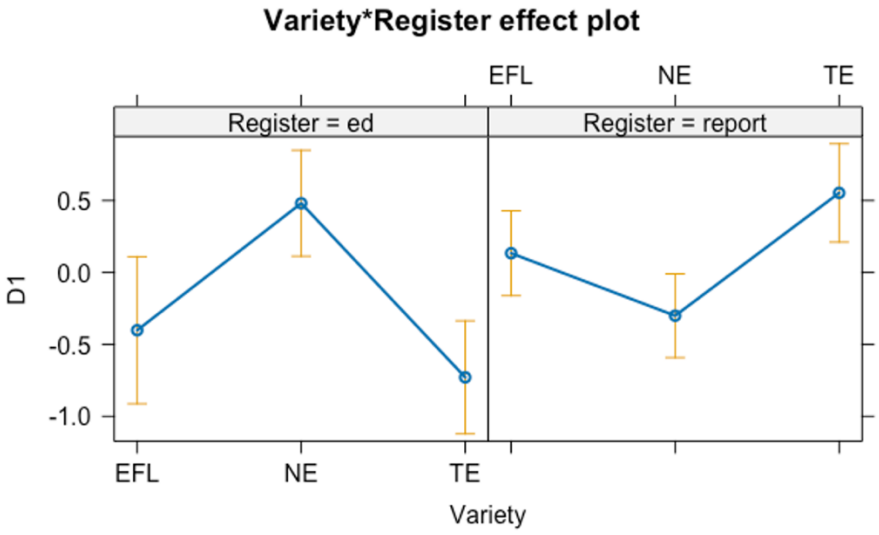

Figure 4.3 Effects plot for Dimension 1.
(Notes: ed = sub-register "editorials", report = sub-register "news reports")

Dimension 1 reveals a significant interactive relation between the predictors. The effects of REGISER vary for each variety. In the case of constrained varieties, dimension scores are higher in reports compared to editorials. However, for native production, the dimension score is higher in editorials than reports. Within each sub-register, the two constrained varieties exhibit similar patterns in comparison to NE. Specifically, translated and EFL editorials show negative scores, whereas NE editorials is positioned on the positive side.

In contrast, translated and EFL reports have positive scores, while NE reports show a negative score.

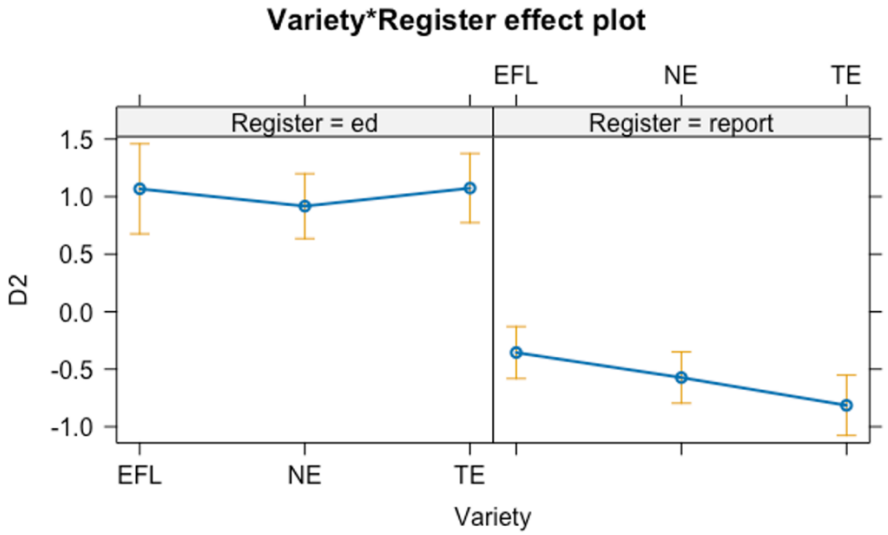

Figure 4.4 Effects plot for Dimension 2.
(Notes: ed = sub-register "editorials", report = sub-register "news reports")

In the case of Dimension 2, a noticeable REGISTER effect is observed: editorials consistently exhibit positive scores across the three varieties, whereas reports have lower scores on the negative pole. This result aligns with the nature of Dimension 2, which opposes two discourse styles: evaluative and reporting/retelling discourse. The former typifies editorials, and the latter is characteristic of news reports. Compared to the REGISTER effect, the VARIETY effect appears relatively minor, as the three varieties are closely positioned within each register. Based on the relative effects illustrated in the plot, it could be inferred that the strong discriminative power of Dimension 2 mainly stems from its ability to differentiate between the two registers.

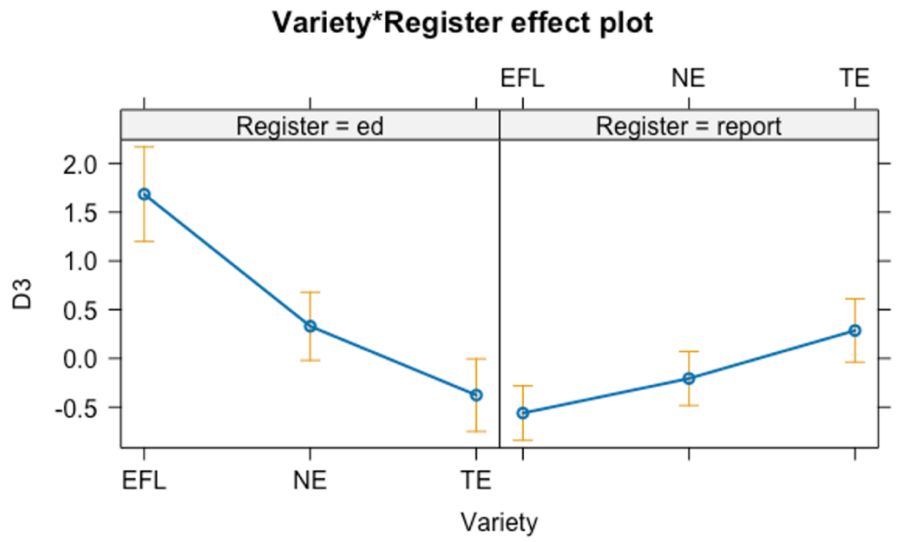

Figure 4.5 Effects plot for Dimension 3.
(Notes: ed = sub-register "editorials", report = sub-register "news reports")

Dimension 3 demonstrates a clear interplay between the effects of REGISTER and VARIETY: the effect of VARIETY varies for each register, and vice versa. Notably, the variation is more pronounced in editorials, as evidenced by the greater discrepancies across the three varieties in the left part of the plot. On the other hand, the differences are relatively flattened in reports. Another noticeable observation is that EFL exhibits greater variation, while TE and NE are positioned closer together, particularly in editorials.

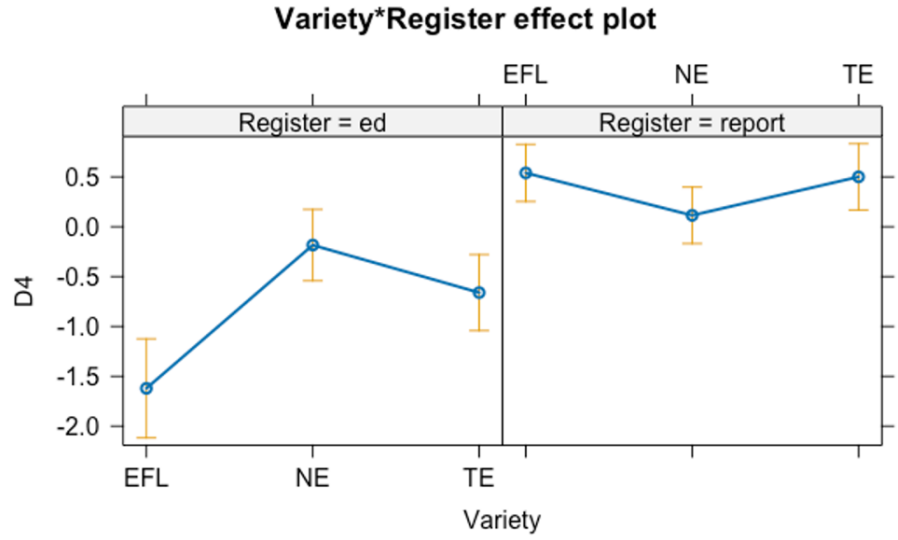

Figure 4.6 Effects plot for Dimension 4.
(Notes: ed = sub-register "editorials", report = sub-register "news reports")

Dimension 4 demonstrates a consistent REGISTER effect across varieties, with reports consistently displaying higher scores compared to editorials. Similar to Dimension 3, greater variation across varieties is observed in editorials, with EFL standing out from the other two varieties. Additionally, similar to Dimension 1, TE and EFL diverge from NE in the same direction for both editorials and reports, with EFL exhibiting a more pronounced divergence.

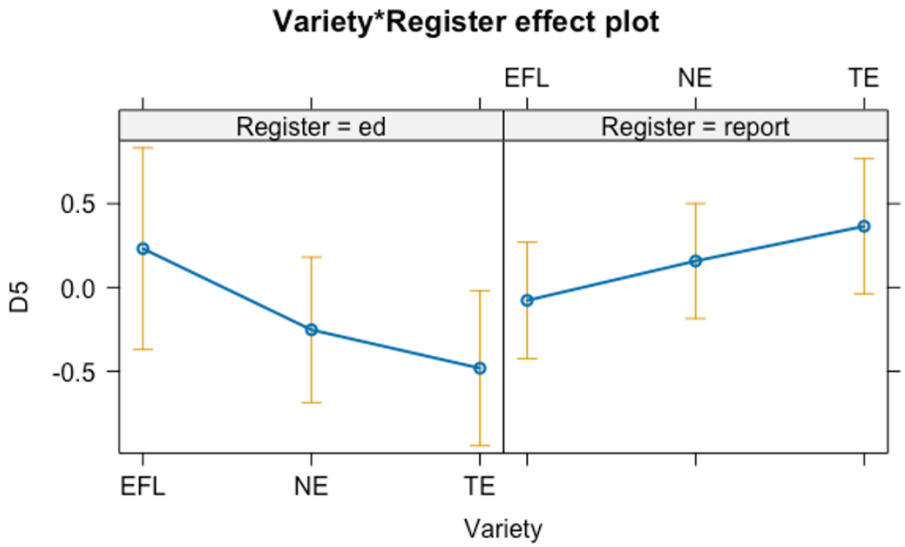

Figure 4.7 Effects plot for Dimension 5.
(Notes: ed = sub-register "editorials", report = sub-register "news reports")

Dimension 5 shows an evident interactive effect of VARIETY and REGISTER. Similar to Dimension 3, EFL shows greater variation compared to the other two varieties, locating on the opposite side of zero in comparison to NE and TE in both registers. The discrepancy is particularly noticeable in editorials. In terms of the REGISTER effect, EFL also stands out: while reports have higher scores than editorials for both NE and TE, the opposite is true for EFL, where editorials have higher scores than reports.

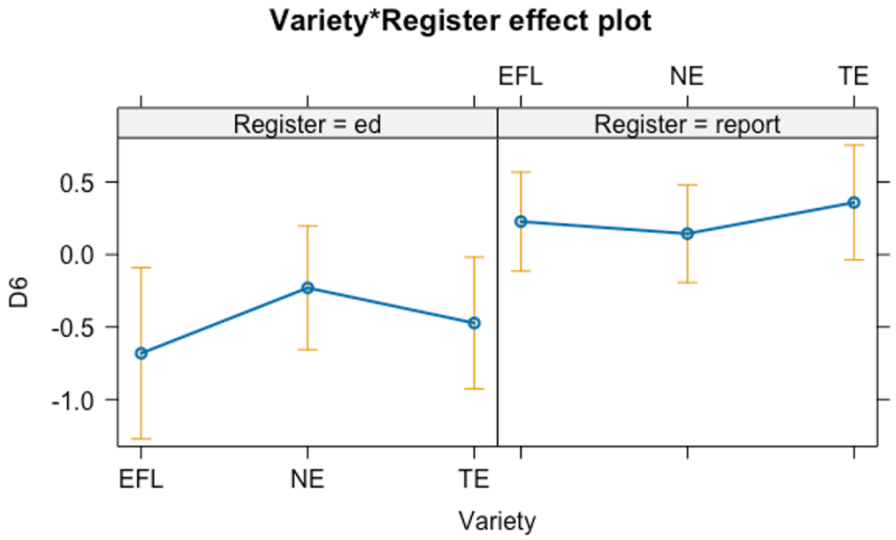

Figure 4.8 Effects plot for Dimension 6.
(Notes: ed = sub-register "editorials", report = sub-register "news reports")

Dimension 6 showcases more independent effects of REGISTER and VARIETY, similar to Dimension 2. Regardless of the varieties, reports consistently exhibit higher scores. However, similar to Dimension 1 and Dimension 4, both constrained varieties diverge from NE in the same direction, which holds true for both registers, although the levels of divergence vary between editorials and reports. Once again, EFL shows greater dissimilarity from NE than TE, particularly in editorials, as observed in the previous three dimensions.

In summary, the regression analysis and the corresponding effect size visualization yielded three main observations that contribute to our understanding of the relationship between register, language variety, and textual features of EFL, NE, and TE. Firstly, the inclusion of the interaction term in the models highlights the intertwined effects of register and variety on textual features. This indicates that the effect of register differs for each variety, and vice versa. However, this interactive effect is less evident in Dimension 2 and 6 where the register effect takes precedence. The relative strengths of register effect and variety effect vary across other dimensions.

Secondly, a comparison between the two sub-registers reveals that there are more noticeable inter-variety differences in editorials, whereas the differences among the three

language varieties are less prominent in reports. This finding suggests that in editorials, the textual features are more susceptible to the influence of language variety, as indicated by the higher degree of variation among the three varieties.

Lastly, a comparison among the three varieties reveals interesting insights into the effects of shared constraints in TE and EFL. In seven out of twelve cases (across the six dimensions and two sub-registers), both EFL and TE exhibit divergence from NE in the same direction, indicating similar patterning in the textual dimensions. However, in the remaining cases, they diverge in different directions, suggesting that there are also differences in the way these varieties deviate from NE. Notably, it is EFL that often stands out in terms of the degree of divergence, indicating that the constraining effects may be more pronounced in EFL compared to TE.

Overall, the regression analysis yielded valuable insights into the combined effects of VARIETY and REGISTER, suggesting the complexity of constrained language production.

4.3.3.2 Textual variations across sub-registers

The previous section has indicated the intricate effects of register and variety on textual variations of EFL and TE. To visualize how exactly the constrained varieties behave for each sub-register along the six dimensions, plots of median scores of the language varieties are presented across the two sub-registers in this section. It demonstrates the textual variations of EFL and TE for editorials and reports in contrast to NE.

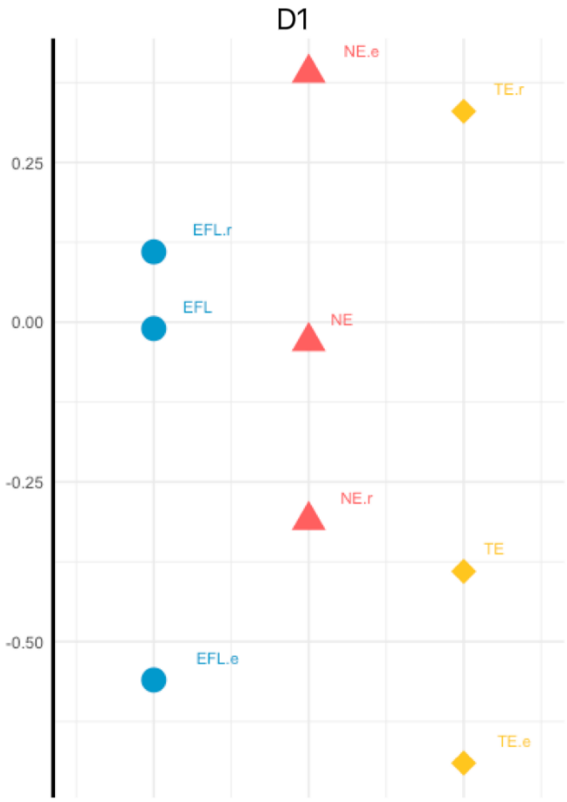

Figure 4.9 Median scores of Dimension 1 for EFL, NE, and TE across sub-registers.
(Notes: EFL.e = EFL editorials; EFL.r = EFL news reports; NE.e = NE editorials; NE.r = NE news reports; TE.e = TE editorials; TE.r = TE news reports)

Figure 4.9 plots the median scores for EFL, NE, and TE on D1 ('Elaborated-involved versus Integrated-formal production'). In both sub-registers under investigation, EFL and TE exhibit similar tendencies, grouping together when contrasted with NE. Interestingly, however, the two constrained varieties shift their positions in contrast to NE in the two sub-registers. For editorials, both TE and EFL show negative scores, while NE shows a positive score. For reports, TE and EFL both position on the positive pole, while NE locates on the negative end. The results show that both TE and EFL editorials are marked by an integrated and formal writing, while TE and EFL reports are marked by an elaborated and involved production. Such shared tendencies set them apart from the non-mediated, native NE.

An additional observation pertains to the variation within each variety. As illustrated in Figure 4.9, the distance separating TE editorials and reports is more pronounced than that in both EFL and NE, especially due to lower scores for TE editorials. This divergence may

lead to TE's lower positioning compared to EFL and NE when the two sub-registers are considered collectively.

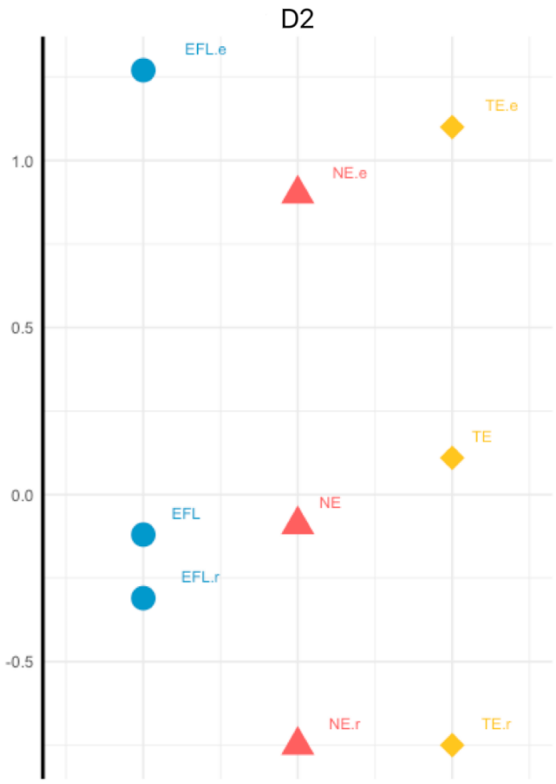

Figure 4.10 Median scores of Dimension 2 for EFL, NE, and TE across sub-registers.
(Notes: EFL.e = EFL editorials; EFL.r = EFL news reports; NE.e = NE editorials; NE.r = NE news reports; TE.e = TE editorials; TE.r = TE news reports)

The influence of register on D2 ('Evaluative discourse versus Reporting/retelling discourse') is clearly illustrated in Figure 4.10, where all three language varieties exhibit positive median values for editorials, juxtaposed by negative scores for reports. This dichotomy is indicative of the nature of D2, namely, contrasting styles of evaluative discourse and reporting or retelling narratives.

While the within-variety differences between editorials and reports for TE and NE are identical, EFL presents a pronounced divergence: both EFL editorials and reports show the highest scores among the three varieties. This emphasizes that both EFL editorials and reports are marked by a stronger tendency towards evaluative discourse, thereby setting EFL apart from TE and NE.

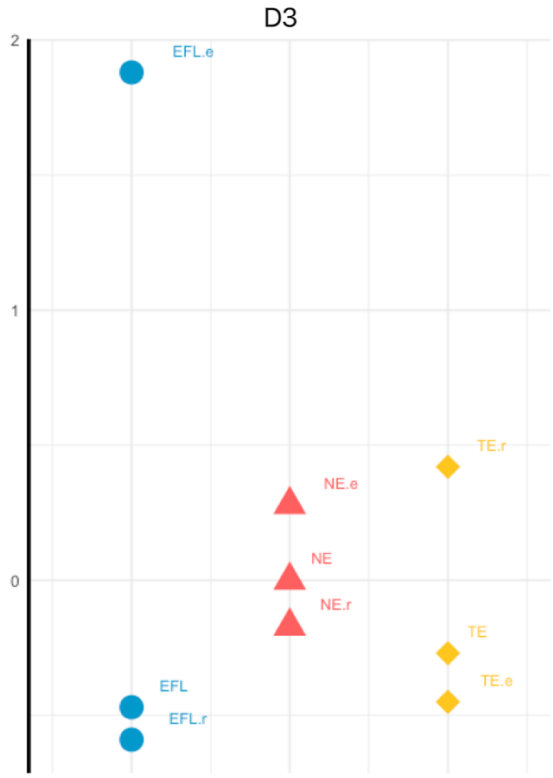

Figure 4.11 Median scores of Dimension 3 for EFL, NE, and TE across sub-registers.
(Notes: EFL.e = EFL editorials; EFL.r = EFL news reports; NE.e = NE editorials; NE.r = NE news reports; TE.e = TE editorials; TE.r = TE news reports)

Figure 4.11 demonstrates the median values for the three varieties across sub-registers on D3 ('Depictive and detailed narration'). The positive scores exhibited by both EFL and NE editorials are markedly distinct from TE's negative scores. Interestingly, this difference takes an inverse turn in news reports, where EFL and NE score negatively while TE attains positive scores. In other words, TE stands out by being more inclined towards a depictive and detailed narration in reports, while avoiding such a tendency in editorials.

Adding complexity to these observations is the pronounced within-variety difference exhibited by EFL. The highest median value for EFL editorials underscores its strong marking of a depictive and detailed narration. This pattern accentuates the distinctive nature of EFL in D3, further differentiating it from both NE and TE.

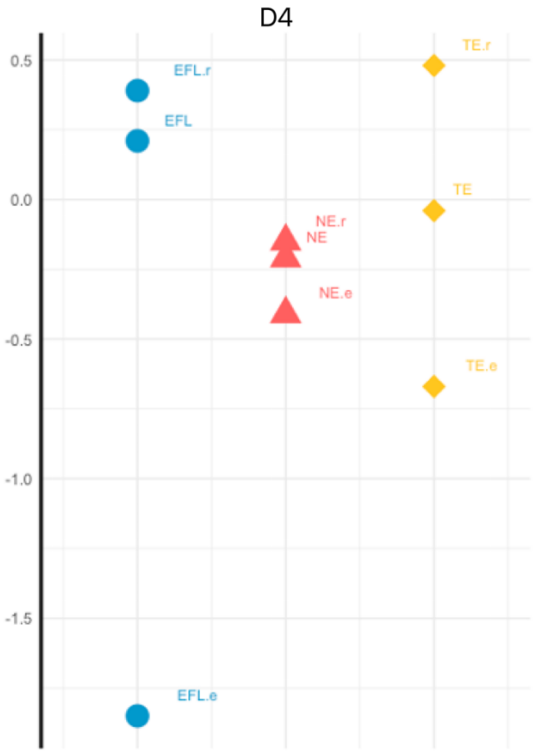

Figure 4.12 Median scores of Dimension 4 for EFL, NE, and TE across sub-registers.
(Notes: EFL.e = EFL editorials; EFL.r = EFL news reports; NE.e = NE editorials; NE.r = NE news reports; TE.e = TE editorials; TE.r = TE news reports)

Figure 4.12 illustrates a clear register effect on the median scores on D4 ('Descriptive narration with a spatial-temporal focus'), evident by a sharp contrast between the positive median values for reports and negative values for editorials. A shared tendency in TE and EFL could also be observed on D4, though the tendency is register-sensitive. Specifically, in the editorial context, both EFL and TE demonstrate more negative scores relative to NE. This underscores a commonality in their absence of descriptive narration with a spatial-temporal focus, creating a unified trait distinct from NE. Conversely, in the report context, EFL and TE both score positively, reflecting a shared concentration on the spatial-temporal elements in their descriptive narrations.

In terms of the within-variety difference, NE reports and editorials are more similar on D4, demonstrated by the closeness of their positions on the plot. This is significantly different from TE and especially EFL, which shows a notably negative score for editorials,

indicating a more pronounced avoidance of descriptive narration with a spatial-temporal focus.

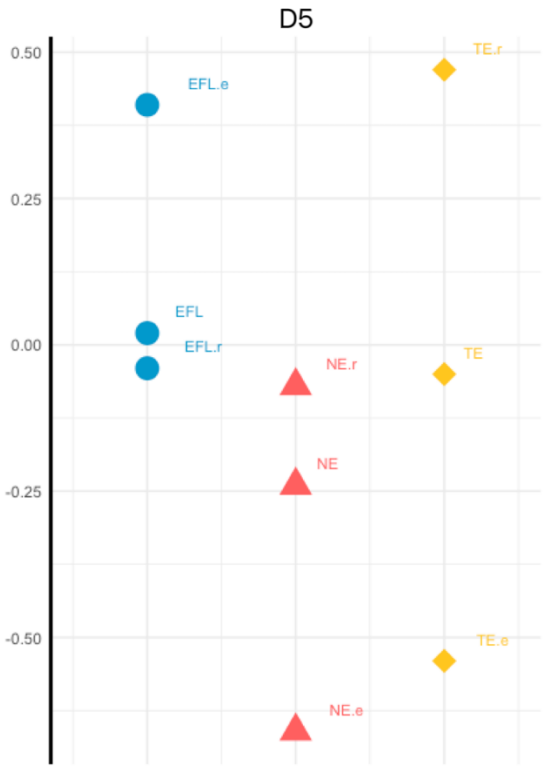

Figure 4.13 Median scores of Dimension 5 for EFL, NE, and TE across sub-registers.
(Notes: EFL.e = EFL editorials; EFL.r = EFL news reports; NE.e = NE editorials; NE.r = NE news reports; TE.e = TE editorials; TE.r = TE news reports)

The intertwining effects of register and variety are confirmed on D5 ('Activity focus vs Referential precision'), as depicted in Figure 4.13. Specifically, in the context of editorials, EFL stands out with a positive score, contrasting the negative scores observed in both NE and TE. Conversely, in reports, TE stands out by displaying a positive score, while both NE and TE show slightly negative scores. These findings highlight that EFL editorials are inclined to showcase an activity focus, whereas TE reports manifest a preference for referential precision. Moreover, TE seems to show the largest within-variety difference compared to the other two, indicating a stronger disparity between TE editorials and TE reports.

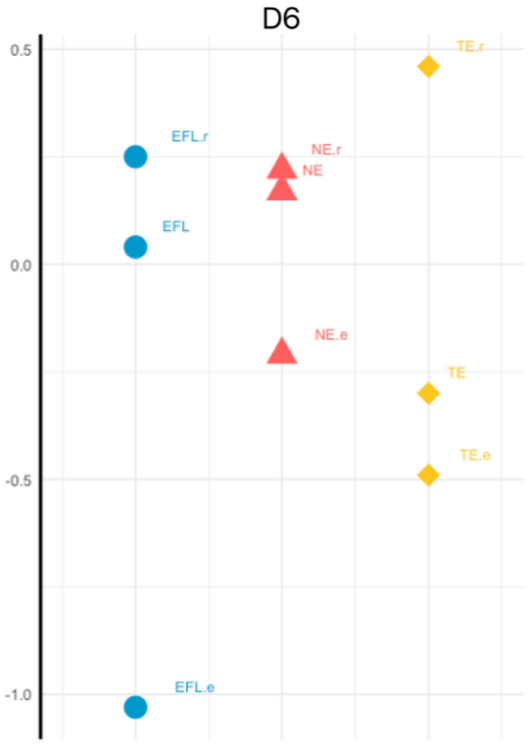

Figure 4.14 Median scores of Dimension 6 for EFL, NE, and TE across sub-registers.
(Notes: EFL.e = EFL editorials; EFL.r = EFL news reports; NE.e = NE editorials; NE.r = NE news reports; TE.e = TE editorials; TE.r = TE news reports)

Figure 4.14 illustrates the median values of the three language varieties across two sub-registers on D6 ('Information density vs Irrealis'). All three varieties exhibit negative median scores for editorials, with EFL editorials demonstrating the most negative values. Conversely, positive scores are observed for reports across all varieties, with TE reports standing out with the highest score. Tentatively, this pattern suggests that editorials in all three varieties are characterized by the use of irrealis, a tendency that is especially strong in EFL. Reports are marked by information density, with TE reports exemplifying the highest level of this trait. Such distinctions in TE and EFL underscore their unique characteristics, leading to greater within-variety differences in these two compared to NE.

In summary, the analysis across six dimensions reveals intricate patterns of similarities and differences among EFL, NE, and TE. In some contexts, EFL and TE display aligned tendencies, distinguishing themselves from NE, such as the integrated and formal writing of editorials and the elaborated and involved production of reports (D1). Similarities between the two can also be observed on D4, where EFL and TE exhibit a pronounced use

of descriptive narration with a spatial-temporal emphasis in reports writing, and conversely, a reduced tendency in editorials when compared to NE. Likewise, on D6, TE and EFL display characteristics of heightened information density in reports and increased use of irrealis in editorials, contrasting with NE. In contrast, distinctions emerge on other dimensions such as D3 where EFL editorials display inclination towards depictive and detailed narration. Additionally, unique characteristics are unveiled in each variety, such as TE's marked disparity between editorials and reports on D5, and EFL editorials' stronger tendency towards evaluative discourse (D2). These multifaceted observations delineate a complex portrait of the three language varieties across the two sub-registers, shedding light on their shared attributes and divergent characteristics which are elaborated in the following section.

4.4 Shared and distinctive patterns of EFL and TE

When discussing the similarities of and differences between TE and EFL, the notion of register emerges as a pivotal constraint in shaping their textual characteristics. As observed from the above results, the textual attributes are sensitive to register across all the identified six dimensions. This finding suggests that in-depth exploration of the complex interplay of various constraints is needed to understand the shared and unique characteristics of these two constrained varieties.

4.4.1 Shared patterns of EFL and TE

Overall, the multidimensional analysis above is unable to identify any textual characteristics that remain consistent across the two sub-registers under examination, at least not for the data used in this study. In other words, no conclusive generalization could be made stating that the constrained language varieties examined consistently manifest characteristics such as being more elaborated and involved or adopting a less depictive and detailed narration compared to the non-mediated, native production.

However, the analysis hints at similarities between the two constrained varieties: on certain dimensions, they both exaggerate the typicality of a certain sub-register. For example, NE editorials are marked by an evaluative discourse as demonstrated by the positive D2 score. In contrast, TE and EFL are pushing such characteristics to a more pronounced level, as evidenced by their higher scores on D2 compared to NE. Similar patterns emerge on D4 ('Depictive narration with a temporal-spatial focus') and D6 ('Information density versus Irrealis'). Specifically, NE reports show higher scores than NE editorials on D4, and constrained language varieties are pushing these tendencies further by showing higher scores for reports and lower scores for editorials. Likewise, NE reports are characterized by information density, the positive end of D6, while NE editorials are marked by the use of irrealis. Again, the two constrained varieties exhibit higher positive scores for reports and lower negative scores for editorials, pushing the tendency further.

In some cases, however, TE and EFL both exhibit an opposite tendency compared to NE. A clear illustration of this is found on D1 ('Elaborated-involved versus Integrated-formal production'), where NE shows a positive score for editorials and a negative one for reports. Conversely, constrained varieties both reveal negative scores for editorials and positive scores for reports, a direct inversion from NE. This discovery implies that unlike NE, TE and EFL editorials tend towards integrated and formal writing, and the reports lean towards elaboration and involvement. Notably, EFL and TE are more closely aligned with each other on this dimension, underscoring their similarity in terms of textual characteristics represented by this dimension.

4.4.2 Distinctive patterns of EFL and TE

Contrasts between TE and EFL mainly appear how far they deviate from NE, which could be observed by comparing the absolute differences between their median scores. In most instances, it is EFL that "goes more extreme". For example, on D4, although both TE and EFL editorials show similar tendencies towards the negative end, indicating an absence of depictive narration with a temporal-spatial focus, EFL editorials score even lower. Similarly, on D6, while both TE and EFL editorials exhibit aligned tendency towards the

negative end, suggesting marked use of irrealis, the tendency is more pronounced in EFL, as evidenced by its significantly lower score. Meanwhile, both TE and EFL reports exhibit a similar inclination towards the positive side, suggesting an increased level of information density, but the tendency is more pronounced in TE, demonstrated by its notably higher score.

Additional disparities are found on other dimensions, such as D3 ('Depictive and detailed narration'), where NE occupies an intermediate position while TE and EFL diverge from NE in different directions. This pattern implies that TE and EFL are marked in different ways on D3 compared to NE, and their marked language use varies with the register. For instance, in editorials, EFL emphasizes depictive and detailed narration, whereas the positive features on D3 are less frequent in TE. Similar observations occur on D5 ('Activity focus vs Referential precision'), where EFL editorials reveal a more activity-oriented approach, while this approach is more pronounced in TE reports.

In summary, the exploration of the textual characteristics of TE and EFL in contrast to NE across six dimensions provides a multifaceted picture, revealing both parallels and disparities between the two. On one hand, shared tendencies are identified, reflecting an inherent commonality in how TE and EFL tend to intensify typical textual characteristics to a more pronounced extent. Conversely, the analysis also unearths their distinctive features. The synthesis of these convergent and divergent traits highlights the nuanced complexities of constrained language varieties.

4.5 Summary

This chapter aims to explore the textual characteristics of constrained varieties of English through a multidimensional analysis. It begins by presenting the results of an exploratory factor analysis, by which six underlying factors are identified. These are subsequently interpreted as six distinct textual dimensions. Along these dimensions, three language varieties are compared based on their respective dimension scores. A critical observation is that the relationship between dimension scores and language variety is modulated by the

constraint of specific sub-registers, namely news reports and editorials. This complex interaction is verified through regression analysis, and the relative influences of register and variety are further elucidated using effect-size plots. The analysis reveals mixed findings concerning the textual properties of the constrained varieties: their features vary across sub-registers, yet within each sub-register, certain commonalities are evident. On some dimensions, both constrained varieties amplify typical features associated with the native variety; conversely, in other circumstances, they exhibit opposing tendencies relative to the native variety. The analysis lends partial support to the hypothesis that both translated English and EFL share textual commonalities and meanwhile highlights the interactive effects of constraint dimensions on shaping the textual characteristics, which offers a nuanced understanding of the intricacies of how language evolves and adapts in response to different constraints.

Chapter 5 Textual characteristics of constrained language: Feature level analysis

This chapter focuses on the linguistic variations of constrained language varieties by looking closely at individual linguistic features. The primary objective is twofold. First, the analysis aims to see how the textual variations of constrained language are achieved through feature-level variation. Given that the textual properties of constrained language have already been identified by the multidimensional analysis, probing into how feature-level variations contribute to textual-level disparities will elucidate the mechanisms that drive the textual differences. Second, the chapter tries to find out whether the feature-level patterns can offer supplementary insights into the nature of the constrained language varieties under review. Rather than coming in with any preconceptions, this exploration remains open to any emerging patterns. Moreover, this approach facilitates a comparison between the findings of the present research and those of previous works that predominantly centered on unidimensional or univariate analyses.

Specifically, feature-level variation is identified based on results of the non-parametric Kruskal-Wallis tests for the 69 linguistic features comparing EFL, NE, and TE. The statistical tests help identify linguistic features that exhibit distributions shared by EFL and TE as opposed to NE and those that display unique distributions in EFL and TE respectively. The analysis was conducted for the two sub-registers separately, due to the significant interactive effects between register and variety as observed in the multidimensional analysis illustrated in the previous chapter.

This chapter will first present the feature-level variations of TE and EFL in relation to their textual variations within each sub-register. Following this, it will focus on the linguistic features that are statistically overused or underused in TE and EFL, probing for the potential implications of these findings on the characteristics of constrained language use.

5.1 Feature-level variations across sub-registers

The aim of this section is to investigate if and how the textual features of the constrained varieties manifest through feature-level variations across the six dimensions. Given the prominent register effect revealed in the proceeding chapter, the analysis was conducted separately for the two sub-registers. To achieve this goal, non-parametric Kruskal-Wallis tests and corresponding post-hoc tests were performed on the 69 linguistic features comparing EFL, NE, and TE (see full results in Appendix 7). For each dimension, boxplots are employed to illustrate the distributions of the linguistic features, enabling a visualization of how feature-level variations contribute to dimension-level fluctuations.

5.1.1 Editorials

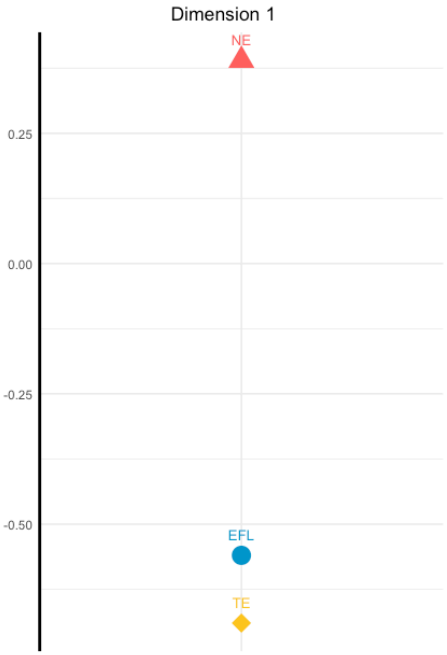

Figure 5.1 Median scores of Dimension 1 for EFL, NE, and TE editorials.

Dimension 1 'Elaborated-involved versus Integrated-formal production'

Figure 5.1 presents the median scores of D1 for EFL, NE, and TE editorials. Notably, NE exhibits a positive median score (0.39) which is higher than both TE (-0.69) and EFL (-0.56). Dimension 1 contrasts elaborated-involved production against integrated-formal production, which groups 13 positive features and 8 negative features. The results indicate

that this dimension effectively differentiates the constrained varieties from the non-constrained variety. TE and EFL editorials share a similar inclination on Dimension 1, as both varieties feature negative scores, indicating a preference for integrated and formal production. Conversely, NE editorials exhibit a higher score, suggesting a more elaborated and involved writing style.

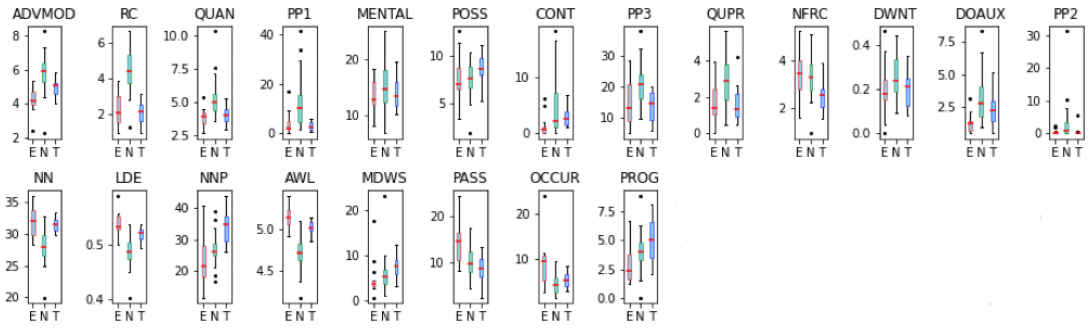

Figure 5.2 Feature distribution on Dimension 1 for EFL, NE, and TE editorials.
(Notes: E = EFL, N=NE, T= TE)

This significant contrast between constrained and non-constrained editorials stems from the distributional differences of individual features loading on this dimension (Figure 5.2). Both TE and EFL editorials demonstrate significant fewer positive features (e.g., quantifiers, quantifying pronouns, first person pronouns, relative clauses, and adverbial modifiers) and more negative features (e.g., nouns and occurrence verbs, and higher level of average word length and lexical density). Meanwhile, the lowest D1 score in TE editorials may be the result of significantly fewer positive features such as second person pronouns and non-finite relative clauses and significantly more negative features such as proper nouns and modal verbs *will* and *shall*. While for EFL editorials, the negative score may be the result of additionally fewer positive features such as verbal contractions and auxiliary *do* and more negative features such as occurrence verbs and passive structures.

In summary, NE editorials are more elaborated, evident by the use of more relative clauses for noun modification to expand sentences. Additionally, NE editorials employ more adverbial modifications and quantifiers to provide additional information and elaboration on specific contexts such as quantity, time, place, and manner. Pronouns and quantifiers are prevalent in NE editorials, making them more engaging for readers. On the other hand,

both TE and EFL editorials are more compact and less involving as evidenced by the absence of these features. Both varieties rely more on nouns, a broader range of lexical items and longer words, which contribute to a more informational and formal production. The two varieties also demonstrate distinctive features that contribute to the contrast between the constrained and non-constrained varieties.

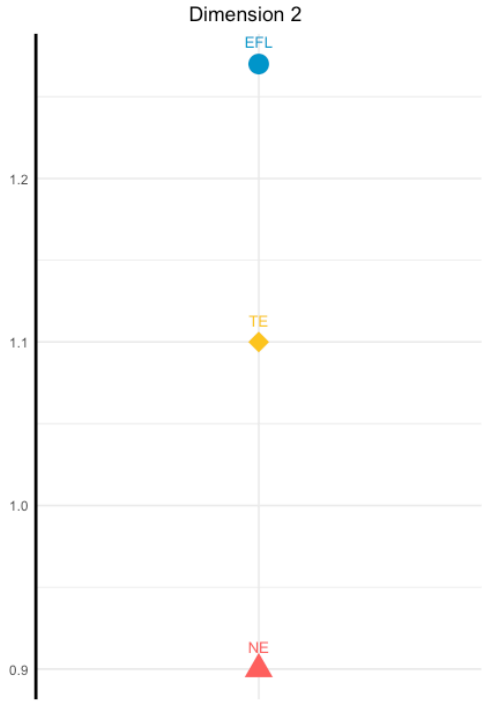

Figure 5.3 Median scores of Dimension 2 for EFL, NE, and TE editorials.

Dimension 2 'Evaluative discourse versus reporting/retelling discourse'

Figure 5.3 depicts the D2 median scores for EFL, NE, and TE editorials. All three language varieties exhibit positive scores, suggesting a prevalence of evaluative elements across the board. Among the three varieties, EFL editorials and TE editorials exhibit similar D2 score at 1.27 and 1.10 respectively, both higher than that of NE editorials (0.90).

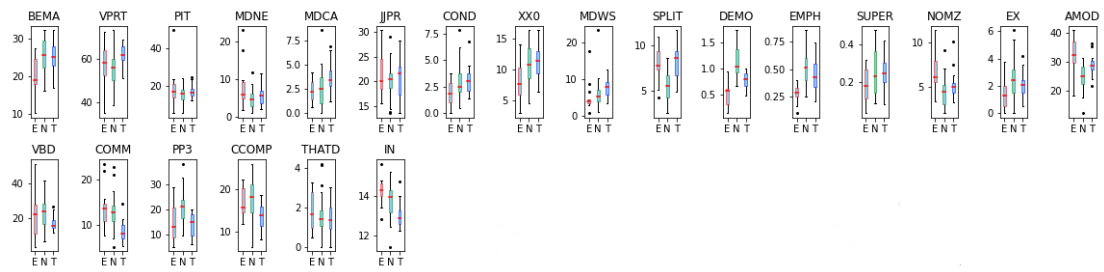

Figure 5.4 Feature distribution on Dimension 2 for EFL, NE, and TE editorials.

(Notes: E = EFL, N=NE, T= TE)

Overall, it is expected that editorials demonstrate positive D2 scores, indicating the argumentative and evaluative nature of editorial writing. This evaluative characteristic seems to be more pronounced in constrained varieties, evident by higher scores for TE and EFL editorials. As illustrated in Figure 5.4, both constrained varieties demonstrate higher frequencies of positive features (e.g., necessity modal verbs, pronoun *it*, present aspect, nominalization, split structures, and adjective modifiers) and lower frequencies of complement clauses, third person pronouns and past tense). The combination of such distributions contributes to a more evaluative tone in both TE and EFL editorials.

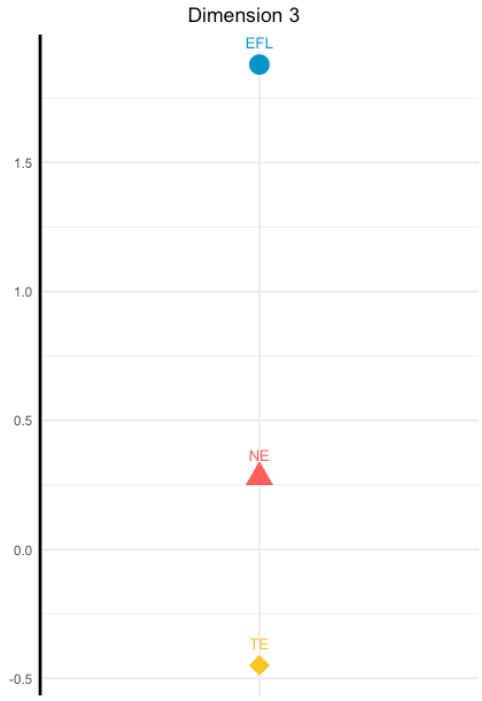

Figure 5.5 Median scores of Dimension 3 for EFL, NE, and TE editorials.

Dimension 3 'Depictive and detailed narration'

Figure 5.5 illustrates the D3 median scores across the three varieties. EFL editorials display the highest score (1.88), followed by NE editorials (0.28). Meanwhile, TE editorials stand out by displaying a negative D3 score (-0.45).

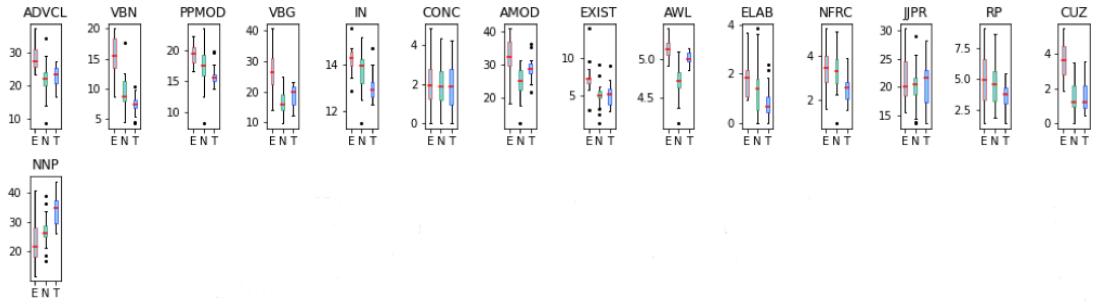

Figure 5.6 Feature distribution on Dimension 3 for EFL, NE, and TE editorials.
(Notes: E = EFL, N=NE, T= TE)

Dimension 3 highlights a depictive and detailed narration, as reflected in its grouping of 14 positive features with only one negative feature. The two constrained varieties display distinctive patterns on D3, which becomes apparent through the distributional patterns of these individual features (Figure 5.6). Compared to NE editorials, EFL editorials display a stronger positive tendency on this dimension, while TE editorials show a reverse trend. EFL's higher score may be the result of consistently higher frequencies in the majority of the positive features. These features include adverbial clauses, non-finite *-ed* verb forms, non-finite *-ing* verb forms, and non-finite relative clauses, all of which contribute to the addition of vivid descriptions and detailed narrations. Furthermore, EFL editorials demonstrate a significantly higher frequency of causal conjunctions, indicating a preference for explicitly expressing causal relationships. Conversely, TE editorials notably lack these positive features. On top of this absence, TE editorials exhibit a significantly higher frequency of proper nouns, which is the only negative feature associated with this dimension.

In summary, the siginificant differences between the two constrained varieties on D3 highlights the distinct linguistic choices employed in each variety, with EFL editorials examplifying the most pronounced characteristics related to depictive and detailed narration, while TE editorials potentially preferring a more concise and focused writing.

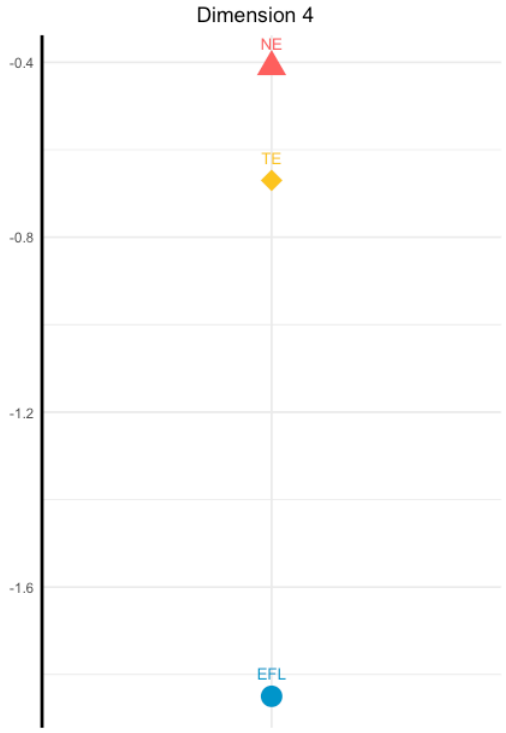

Figure 5.7 Median scores of Dimension 4 for EFL, NE, and TE editorials.

Dimension 4 'Descriptive narration with a spatial-temporal focus'

Figure 5.7 illustrates the median scores of D4 across the three language varieties. Notably, all three varieties exhibit negative scores, indicating a shared tendency towards less explicit description. Compared to NE with a score of -0.41, both EFL and TE reveal a more pronounced tendency in this direction. This is particularly evident in EFL, as demonstrated by its lower score (-1.85) compared to TE (-0.67).

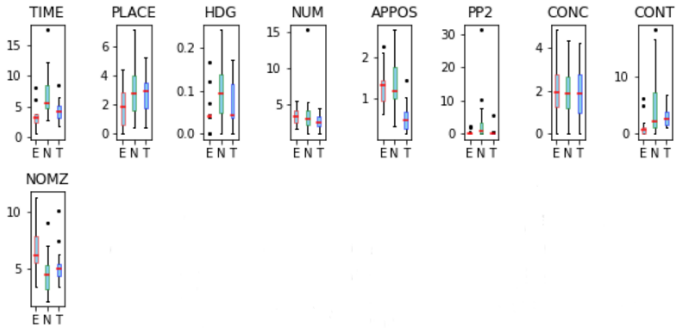

Figure 5.8 Feature distribution on Dimension 4 for EFL, NE, and TE editorials.
(Notes: E = EFL, N=NE, T= TE)

Dimension 4 underscores descriptive narration with a spatial-temporal focus. The nagative scores of all three language varieties align with the results of the regression analysis in the

previous chapter, which reveals an effect of register on this dimension. This result indicates that the descriptive aspects related to spatial and temporal references are less prominent in editorials compared to reports. The differences among varieties is supported by the distribution of individual features that contribute to this dimension (Figure 5.8). Both TE and EFL display lower frequencies of positive features such as time adverbials, second person pronouns, and hedges, and higher frequency of the negative feature, i.e., nominalizations. EFL stands out by ranking the lowest among the three in terms of median score. This is evident through the reduced frequency of positive features such as verbal contractions and place adverbials.

In general, all three language varieties display a predominantly low frequencies of positive features associated with Dimension 4, and this tendency is particulalry prominent in constrained varieties, espeicially EFL editorials. Both TE and EFL editorials exhibit a lower frequency of time referencing adverbials, indicating a reduced emphasis on providing specific descriptions related to temporal aspects. Moreover, the use of fewer hedges reflects a more assertive and direct tone, while the reduced usage of verbal contractions signifies a formal writing style. Meanwhile, TE and EFL editorials also utilize more nominalizations, the only negative feature on D4, which further contributes to their negative D4 scores. This result reveals the distinctive characteristics of constrained varieties in lacking descriptive narration with a spatial-temporal focus.

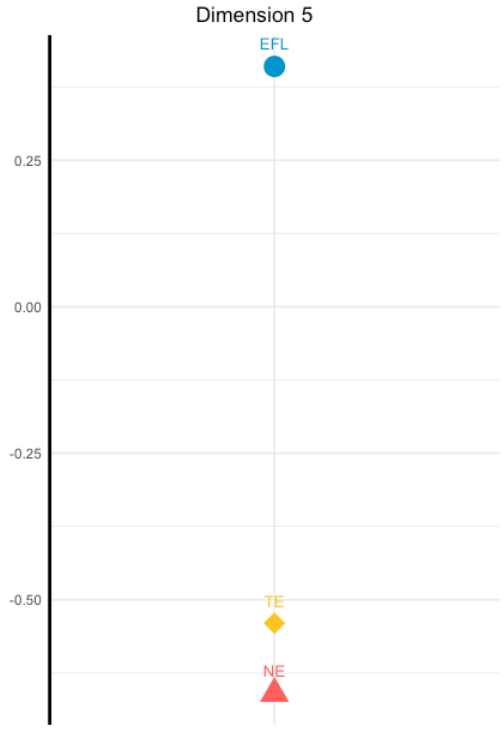

Figure 5.9 Median scores of Dimension 5 for EFL, NE, and TE editorials.
Dimension 5 'Activity focus versus referential precision'

Figure 5.9 presents the median scores of D5 for the three language varieties. The grouping of three positive features and four negative features on D5 is tentatively decided to reflect a contrast between activity focus and referencial precision. EFL editorials stand out with the highest and only positive median score (0.41), while NE and TE editorials both exhibit negative scores, recorded at -0.66 and -0.54, respectively.

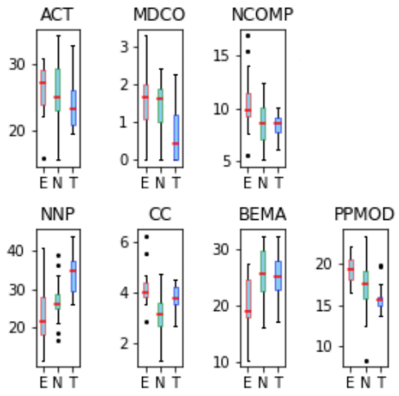

Figure 5.10 Feature distribution on Dimension 5 for EFL, NE, and TE editorials.
(Notes: E = EFL, N=NE, T= TE)

Examining the distribution of individual features associated with D5 (Figure 5.10) sheds light on the features contributing to the higher score observed in EFL editorials. EFL editorials demonstrate higher frequencies of positive features including noun compounds, activity verbs and modal *could*, and simultaneously lower frequencies of negative features such as proper nouns and copula verb *be*. This suggest a greater emphasis on action-oriented verbs and a reduced focus on specific individuals or institutions, as captured by the category of proper nouns. On the other hand, TE editorials show a reverse tendency, being less activity-focused but more precise on references. This is demonstrated by the fewer positive features such as modal *could* and activity verbs, and significantly more negative features of proper nouns.

Overall, the results suggest a inclination of EFL editorials towards the positive end of D5, emphasizing an activity-focused narrative, while TE and EFL editorials are marked by referential precision.

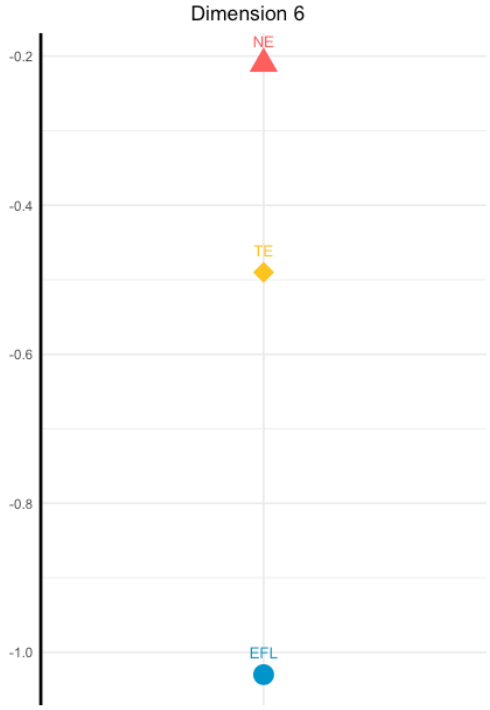

Figure 5.11 Median scores of Dimension 6 for EFL, NE, and TE editorials.
Dimension 6 'Informational density versus irrealis'

Figure 5.11 provides an overview of the median scores of D6 across the three language varieties. Notably, all three varieties display negative median scores, suggesting a shared inclination towards the negative end of D6. This tendency is more pronounced in both constrained varieties, as evident by their lower scores compared to NE. NE editorials display the highest score at -0.21, followed by TE editorials at -0.49, and EFL editorials exhibit the lowest score recorded at -1.03.

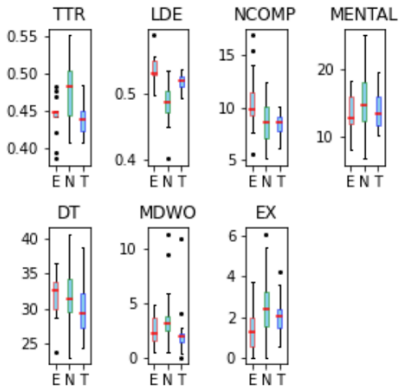

Figure 5.12 Feature distribution on Dimension 6 for EFL, NE, and TE editorials.
(Notes: E = EFL, N=NE, T= TE)

A closer examination of the individual features contributing to D6 (Figure 5.12) indicates that both TE and EFL, despite having higher level of lexical density compared to NE, show fewer positive features such as mental verbs and a lower level of type-token ratio, leading to more negative scores. Meanwhile, EFL editorials demonstrate a higher frequency of negative features such as determiners.

Overall, EFL and TE editorials lean more towards the negative end of D6 due to a reduced level of type-token ratio and fewer uses of mental verbs. EFL editorials, in particular, demonstrate an even more negative inclination due to the lower proportion of determiners. However, these characteristics should be interpreted with caution, given the limited number of features associated with D6.

5.1.2 Reports

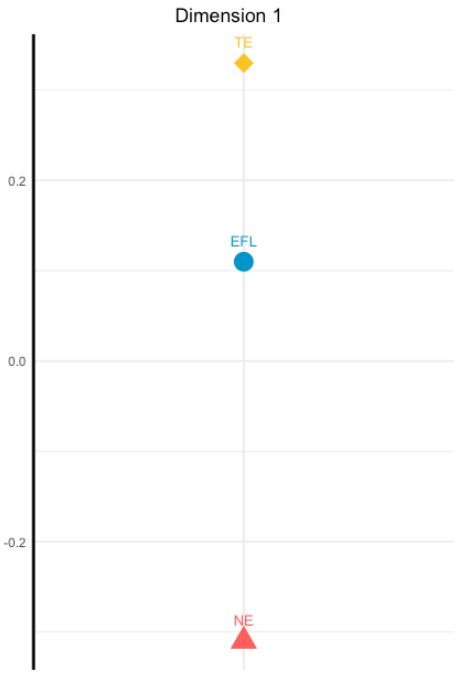

Figure 5.13 Median scores of Dimension 1 for EFL, NE, and TE reports.
Dimension 1 'Elaborated-involved versus Integrated-formal production'

Figure 5.13 provides an overview of the median scores of D1 for EFL, NE, and TE reports. Compared to NE, both EFL and TE reports display a similar pattern on D1, with positive scores recorded at 0.11 and 0.33 respectively. In contrast, NE reports are the only one to display a negative score recorded at -0.31.

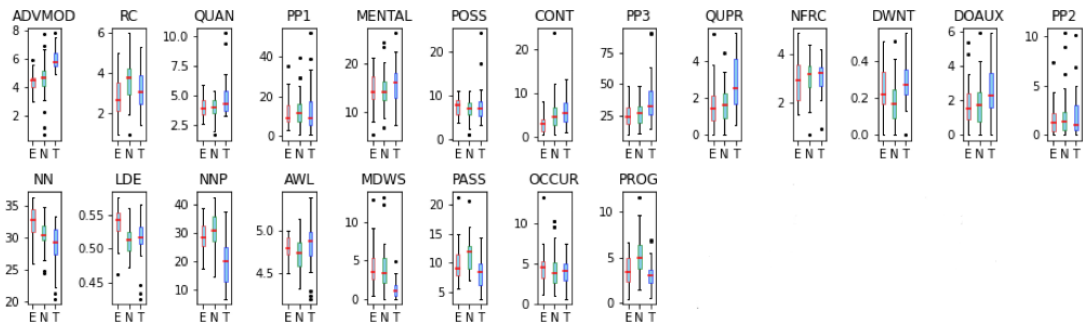

Figure 5.14 Feature distribution on Dimension 1 for EFL, NE, and TE reports.
(Notes: E = EFL, N=NE, T= TE)

The analysis reveals contrasting tendencies between constrained and non-constrained varieties on D1. Specifically, the constrained varieties exhibit an inclination towards elaboration and involvement in report writing, an inverted trend compared to NE reports. This divergence in median scores is elucidated by the distribution of linguistic features on the dimension, as depicted in Figure 5.14. The inclinations in TE and EFL reports are manifested through higher frequencies of positive features such as downtoners, possessive modifiers, and mental verbs, and lower frequencies of negative features including progressive aspect, proper nouns, and passive structures. The tendency is even more pronounced in TE reports, potentially attributable to significantly higher frequencies of positive features such as adverbial modifiers, quantifiers, quantifying pronouns, and the auxiliary verb *do*.

Overall, the analysis of Dimension 1 reveals a notable distinction between constrained and non-constrained varieties in reports writing. While NE reports lean towards integration, formality, and precision, TE and EFL reports demonstrate elaboration, informality, and a focus on expanded content and a less formal tone.

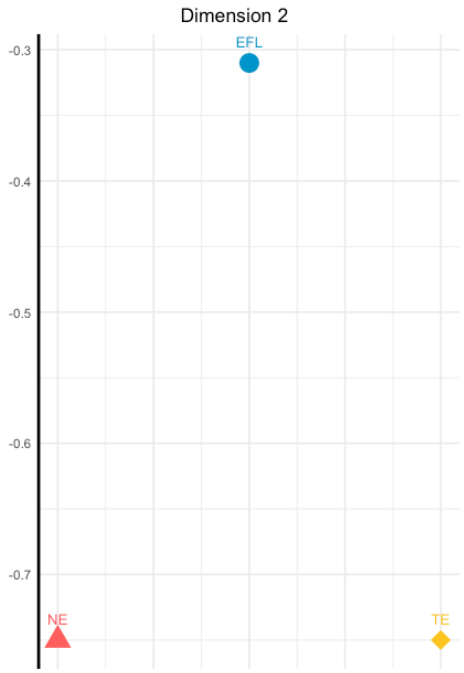

Figure 5.15 Median scores of Dimension 2 for EFL, NE, and TE reports.
Dimension 2 ‘Evaluative discourse versus reporting/retelling discourse’

All three varieties exhibit negative median scores on D2, a result consistent with expectations. Since the reporting and retelling discourse represents the negative pole of D2, such a trend is anticipated in report writing. Interestingly, EFL reports display more evaluative elements with the highest score recorded at -0.31, higher than the identical scores for TE and NE at -0.75.

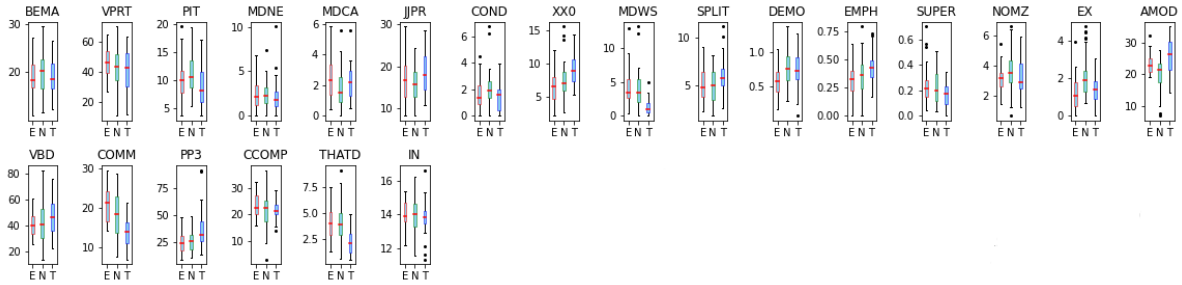

Figure 5.16 Feature distribution on Dimension 2 for EFL, NE, and TE reports.
(Notes: E = EFL, N=NE, T= TE)

The distribution of linguistic features on D2 is depicted in Figure 5.16. It seems that the higher median score of EFL reports could be attributed to more occurrences of positive features. These include the use of superlatives, present tense, and modal verbs such as *will*, *shall*, and *can*, coupled with fewer negative features such as past tense. Especially, the prevalence of these modal verbs could convey a sense of possibilities, intentions, and obligations, allowing EFL reporters to incorporate a broader evaluative dimension. Conversely, TE reports show a pronounced inclination towards the negative end of D2, characterized by significantly fewer positive features mentioned above and an increased co-occurrence of negative features such as past tense and third person pronouns.

In summary, the distribution of D2 scores indicates that EFL, NE, and TE reports are all typical reporting discourse. However, the EFL reports stand apart by incorporating additional evaluative elements, most notably through the more frequent use of modal verbs. This suggests that EFL reports exhibit a unique combination of reporting and evaluative features, distinguishing EFL from the other two varieties.

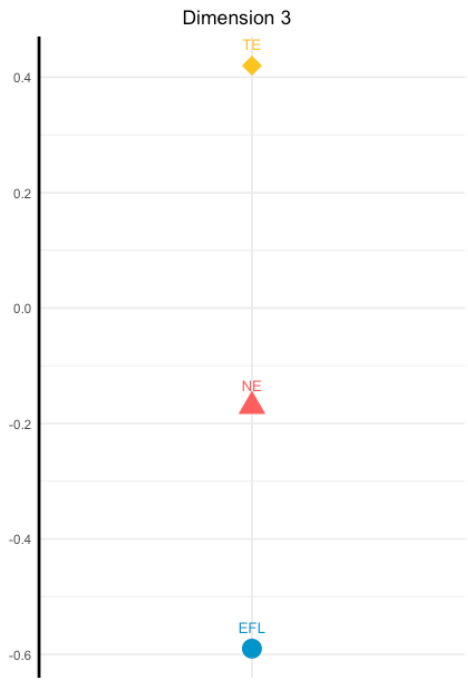

Figure 5.17 Median scores of Dimension 3 for EFL, NE, and TE reports.
Dimension 3 'Depictive and detailed narration'

The median scores of D3 for the three language varieties are depicted in Figure 5.17. It is noteworthy that distinct differences emerge among TE and EFL reports on D3. Specifically, both EFL and NE reports are characterized by negative scores, with recorded values of -0.59 and -0.17 respectively. In contrast, TE reports are marked by a positive score of 0.42, indicating an inclination towards depictive and detailed narration.

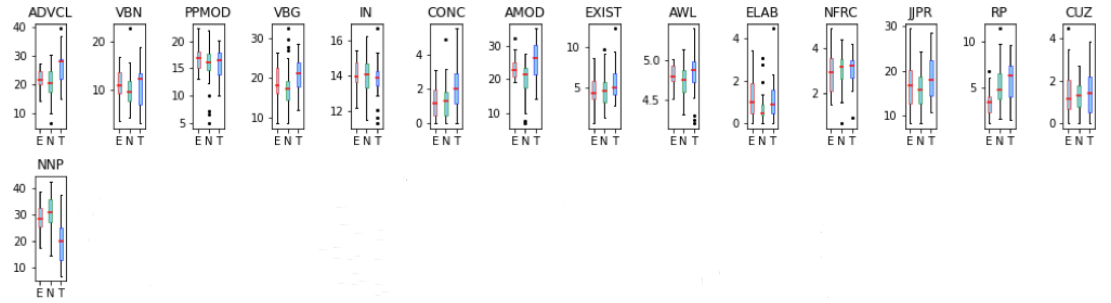

Figure 5.18 Feature distribution on Dimension 3 for EFL, NE, and TE reports.
(Notes: E = EFL, N=NE, T= TE)

A closer examination of the distribution of individual features on D3 (Figure 5.18) provides more insights into this contrast. Specifically, the more detailed narration in TE reports is achieved by more frequent use of a series of positive features including adverbial clauses, non-finite *-ed* verb forms and *-ing* verb forms. These features contribute to the inclusion of

additional information and add depth to the narrative. Both predictive and attributive ajectives are more frequent in TE reports, indicating a greater emphasis on descriptive writing. Other positive features that are more frequent in TE reports include prepositional modifiers, causal conjunctions, and exitential verbs, etc. Meanwhile, there is significant fewer proper nouns in TE reports compared to the other two, indicating a more general and less entity-focused style in TE reports. EFL reports exhibit a similar but more amplified tendency as NE reports, manifested as decreased frequencies of positive features such as non-finite relative clauses, particles, causal conjunctions, concessive conjunctions, and existential verbs.

Overall, TE reports stand out on Dimension 3 by displaying a more depictive and detailed narration compared to EFL and NE reports. This inclination is collectively realized by more frequent use of features such as adverbial clauses, non-finite verb forms, and adjectives, as well as a reduced reliance on proper nouns.

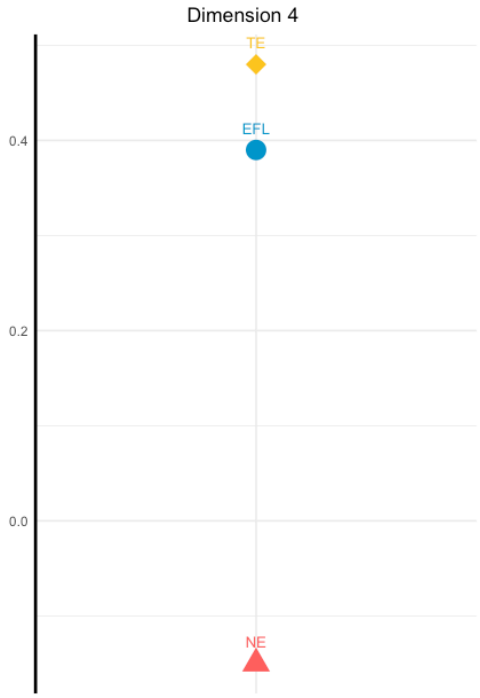

Figure 5.19 Median scores of Dimension 4 for EFL, NE, and TE reports.

Dimension 4 'Descriptive narration with a spatial-temporal focus'

The median scores of D4 across three language varieties are depicted in Figure 5.19. TE and EFL reports are found to pattern together on D4, with closely located median scores of

0.48 and 0.39, respectively. These scores are noticeably higher than that of NE reports which recorded at -0.15.

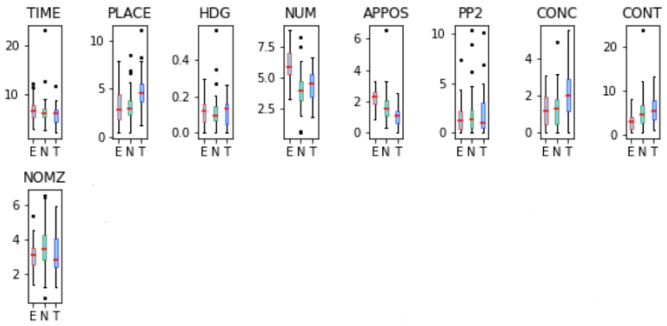

Figure 5.20 Feature distribution on Dimension 4 for EFL, NE, and TE reports.
(Notes: E = EFL, N=NE, T= TE)

Higher D4 scores correspond a tendency towards the positive end of the spectrum, characterized by descriptive narration with a spatial-temporal focus and a reduced presence of nominalizations. Upon closer examination of the distribution of individual features along this dimension (Figure 5.20), it becomes apparent that EFL and TE reports make greater use of positive features such as hedges and numeric modifiers for nouns. Additionally, TE reports employ more place adverbials to elaborate on spatial and locational references. More usage of verbal contractions and concessive conjunctions in TE reports also contribute to their higher scores. On the other hand, EFL reports rely more on appositional modifiers to provide precise information on specific details. Conversely, both EFL and TE reports exhibit a lower frequency of nominalizations compared to NE reports, indicating a preference for more concrete and direct language usage.

In summary, EFL and TE reports exhibit a collective inclination towards descriptive narration and a reduction in abstractness, marking a reverse trend when contrasted with NE.

Figure 5.21 Median scores of Dimension 5 for EFL, NE, and TE reports.
Dimension 5 'Activity focus versus referential precision'

Figure 5.21 presents D5 median scores across the three language varieites. TE reports display the highest median scores (0.47), while EFL and NE reports display scores slightly below zero, at -0.04 and -0.07 respectively. Compared to EFL and NE, TE reports follow a divergent pattern on this dimension.

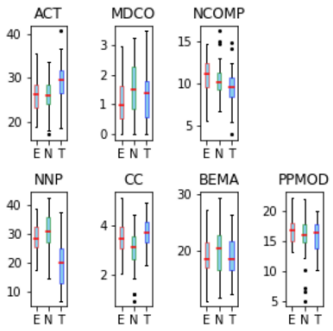

Figure 5.22 Feature distribution on Dimension 5 for EFL, NE, and TE reports.
(Notes: E = EFL, N=NE, T= TE)

Only TE reports are positioned on the positive end of D5. The higher D5 score in TE reports signifies that they are more activity-focused and demonstrate less referential precision, as evidenced in the distribution of individual features (Figure 5.22). The presence of activity

verbs in TE reports signifies an emphasis on describing actions and processes, contributing to a more dynamic and active tone in the text. Conversely, the decreased occurrence of proper nouns suggests a diminished focus on specific and identifiable entities, potentially resulting in a lower level of referential precision. Contrarily, EFL reports demonstrate a reverse tendency with fewer positive features such as the modal verb *could*, and more negative features such as propositional phrases as noun modifiers.

Overall, the finding suggests that TE reports emphasize a strong activity focus with reduced referential precision compared to the other two.

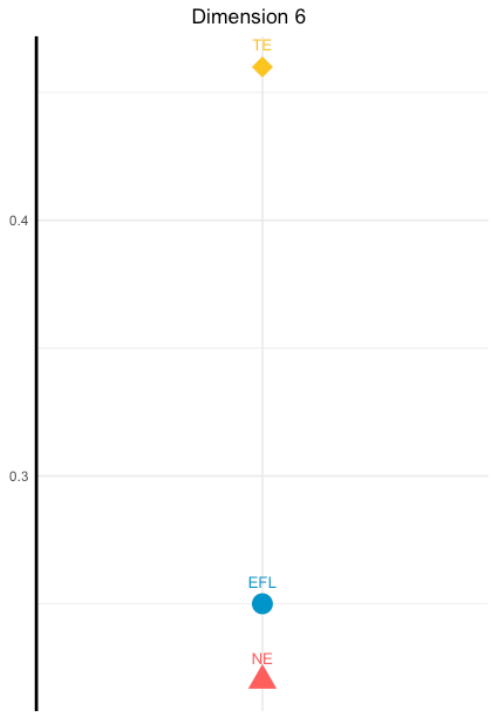

Figure 5.23 Median scores of Dimension 6 for EFL, NE, and TE reports.

Dimension 6 'Informational density versus irrealis'

All three language varieties display positive median scores on D6, as shown in Figure 5.23. TE reports exhibit the highest median score (0.46), followed by EFL reports (0.25) and NE reports (0.22). Both TE and EFL reports exhibit higher scores compared to NE, which points to a shared tendency between TE and EFL towards a higher level of informational density, though the tendency is much more pronounced in TE reports.

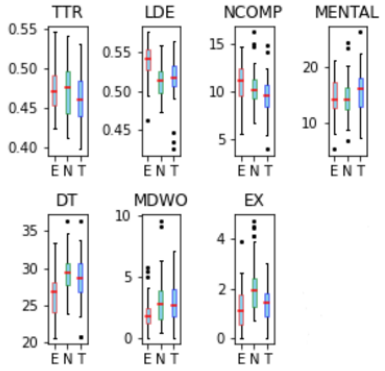

Figure 5.24 Feature distribution on Dimension 6 for EFL, NE, and TE reports.
(Notes: E = EFL, N=NE, T= TE)

This tendency is manifested in the distribution of individual features (Figure 5.24). Both TE and EFL reports appear more informational, marked by an increased level of lexical density and a greater usage of mental verbs, as well as a reduced presence of existential *there* and fewer determiners, compared to NE. Additionally, EFL reports feature a higher level of lexical density which contributes to their rich expression, and TE reports exhibit more usage of mental verbs which convey cognitive processes and thoughts, and thereby enrich the informational content of the texts. Besides, EFL reports also exhibit a higher frequency of noun compounds which enable the combination of multiple concepts into a single unit.

Overall, all three language varieties are positioned on the positive end of D6, reflecting a focus on information density. However, this tendency is more pronounced in the two constrained varieties. Meanwhile, the individual strategies for achieving informational density vary: TE reports rely on mental verbs, EFL reports make more use of noun compounds and exhibit higher lexical density, while NE reports exhibit an increased frequency of negative features on the dimension.

5.2 Shared and distinctive patterns of EFL and TE

This section focuses on the linguistic features that are statistically overused or underused in both TE and EFL compared to NE, aiming to explore the implications of these patterns

on the characteristics of constrained language use. The analysis also identifies features uniquely distributed in TE and EFL respectively, shedding light on their distinctive properties.

5.2.1 Shared patterns of EFL and TE

5.1.1.1 Editorials

Table 5.1 Distinctive features of constrained varieties for editorials.

Over-represented features	Under-represented features
AWL	QUAN
LDE	QUPR
NN	DEMO
SPLIT	TIME
AMOD	PP1
CONJ	PP3
	RC
	ADVMOD

The results of the statistical tests show that in 14 out of the 69 features, TE and EFL editorials show similar distributional patterns that set them apart from the non-constrained editorials, and the differences reach a statistically significant level ($p<0.05$). These features are listed in Table 5.1.

Altogether, six linguistic features are found to be overrepresented in both TE and EFL editorials. These features include two related to general text properties (average word length and lexical density), two lexical level features (nouns and adjectival noun modifiers, one functional lexis (coordinating conjunctions) and one syntactic feature (split auxiliary). On the other hand, the two constrained varieties are characterized by the underrepresentation of eight linguistic features, including two determiners (quantifiers and demonstrative pronouns), three categories of pronouns (first person pronouns, third person pronouns, and quantifying pronouns), adverbial modifiers and particularly time adverbials, and one syntactic feature (relative clauses).

Notably, the overrepresented features in TE and EFL editorials appear to be complexity related. Firstly, the higher level of average word length and lexical density indicates an inclination towards information density at the overall textual level. It is commonly acknowledged that longer words tend to convey more specific and specialized meanings, while shorter words are often linked to a higher frequency of usage and more general meanings (Biber, 1988, p.238). Similarly, lexical density reflects the proportion of lexical items that carry specific semantic meaning in contrast to function words which primarily fulfill grammatical or structural functions.

Besides, TE and EFL editorials also stand out by exhibiting higher frequencies of two lexis, namely, nouns and adjectival modifiers for nouns. The calculation of total nouns includes all nouns in singular and plural forms, as well as proper nouns. Adjectival modifiers encompass any adjectival phrase that serve to modify a noun or pronoun. Interestingly, it should be noted that the adjectives are normalised based on the number of nouns rather than the total word count in a given text. In other words, it measures the proportion of nouns that are pre-modified by adjectives. The higher frequency of adjectives is not directly linked to the higher frequency of nouns, and thus should be interpreted independently from the number of nouns. This result indicates that not only nouns are more frequent in TE and EFL editorials, so do the "adjective plus noun" sequences, which have been seen as "the most common building block of complex noun phrase" (Leech et al., 2009, p. 216). Both the use of nouns and "adjective plus noun" sequences contribute to a more compact packaging of information in the constrained language use in editorials.

Both TE and EFL editorials demonstrate more frequent use of the coordinating conjunctions[9], represented by *and*, *but*, and *or*. The use of coordinating conjunctions presents multiple ideas in parallel or in contrast, creating a sense of parallelism and connection, suggests a higher level of coordination complexity in the two constrained varieties.

[9] In the current study, coordinating conjunctions are tagged by the Stanford Tagger, including *and*, *but*, *nor*, *or*, *yet*, as well as the mathematical operators *plus*, *minus*, *less*, *times* (in the sense of "multiplied by") and *over* (in the sense of "divided by") when they are spelled out (Le Foll, 2021).

One syntactical feature, SPLIT[10], exhibits a higher frequency in the two constrained varieties. SPLIT refers to the structure where an auxiliary verb is separated or split by another word or phrase. This construction is generally considered non-standard or informal in standard English, yet Biber (1988: 244) reveals its prevalence in certain written genres compared to conversational language. Upon closer examination of the TE and EFL editorials, it becomes apparent that the structure is most frequently used when an evaluative adverb is involved to express a viewpoint or stance.

Excerpt (1):

It reminds us that high salary alone *cannot fundamentally curb* corruption, because greed is infinite. (EFL-ed-010)

Excerpt (2):

Most analysts believe it's unthinkable that the new Brazilian government *would significantly replace* Brazil-China trade with Brazil-US trade. (TE-ed-018)

In these examples, adverbs are placed between the auxiliary verb and the main verb. For instance, in Excerpt (1) the extent and nature of the corruption issue is emphasized by putting *fundamentally* in the verb phrase. Similarly, *significantly* in Excerpt (2) was inserted to intensify the writer's assertion about the consequence. In general, the repeated use of SPLIT structures in TE and EFL editorials has an effect of emphasizing and intensifying the intended perspective.

One notable observation regarding the underrepresented features in TE and EFL editorials is that many of them fall into the category of function words (Biber et al., 2002, p.26), such as determiners and pronouns. Determiners, particularly quantifiers and demonstrative pronouns, serve essential roles in indicating quantity, specificity, or proximity in relation to nouns. Demonstrative pronouns (such as *this*, *that*, *these*, *those*) refer to specific things within the context, either in the immediate textual surroundings or the external situation

[10] In the current study, SPLIT identify the following structures: TO/auxiliary (BE/DO/HAVE) + adverb + (adverb +) verb (Le Foll, 2021).

(Biber et al., 2002, p. 98). Quantifiers (e.g., *plenty of* organizations, *a few* months) are used to indicate the quantity or scale of subjects under discussion. Pronouns function as substitutes for nouns or noun phrases, and their references are typically discernible from the surrounding context. First person pronouns are employed to refer to the writer and are commonly found in interpersonal and involved discourse. On the other hand, third person pronouns are utilized to refer to third parties or entities, neither including the writer nor the addressee. Similar to quantifiers, quantifying pronouns are used to indicate a general or indefinite sense of quantity. These pronouns are instrumental in establishing referential links within a text.

Viewed from another perspective, these features serve a deictic function that requires interpretation based on the context of the utterance. The scarcity of these elements in TE and EFL editorials may indicate a preference for a less context-specific or more objective writing style focusing on information presentation rather than detailed quantification or personal involvement.

Besides function words, both TE and EFL editorials are also found to use fewer adverbial modifiers, especially time adverbials. Adverbial modifiers are able to fulfill a wide range of semantic roles by providing information about factual contexts, such as time (e.g., *ago*, *afterwards*), location (e.g., *in the park*), frequency (e.g., *sometimes*, *rarely*), manner (e.g., *firmly*, *softly*), and degree (e.g., *very*, *certainly*). They can convey attitudes and add comments (e.g., *truly*, *inevitably*), and serve a structural function by acting as linking elements (e.g., *overall*, *so*). A contrastive examination of EFL/TE and NE editorial texts reveal notable differences in the types and functions of adverbials employed. In the former case, many adverbials are those that modify verbs (e.g., Excerpt 3), indicating a more limited range of forms or functions compared to NE editorials. Additionally, EFL and TE editorials tend to rely more on formulaic or multiword adverbials (e.g., Excerpt 4).

Excerpt (3):

China has *closely* followed the development of Africa and *sincerely* wishes to make its contribution to the African people in developing their nations and creating a better life. (EFL-ed-003)

Excerpt (4):

The Chinese government has *not only* maintained the dominant status of public ownership, *but also* created room for the private ownership *so* that they can play their respective parts in the economy and distribution of resources. (EFL-ed-013)

The only syntactic feature that is less frequent in TE and EFL editorials is the relative clauses. Relative clauses provide additional information about the nouns and function as adjectival modifiers. However, in contrast to adjectives and prepositional phrases that serve as modifiers, they permit more intricate and comprehensive descriptions, as they may encompass subjects, verbs, and other corresponding components within the clause. On the other hand, "they are the most fully explicit form of noun modification" (Leech et al., 2009, p. 226). This deficiency of relative clauses aligns with the overrepresentation of adjectival modifiers for nouns, indicating a preference for more compact phrasal modification for nouns in constrained language varieties.

5.1.1.2 Reports

Table 5.2 Distinctive features of constrained varieties for reports.

Over-represented features	Under-represented features
AMOD	EX
	PROG

Table 5.2 presents the features exhibiting distinctive distribution shared by TE and EFL reports in contrast to NE reports. Compared to editorials, the distinction between reports produced by constrained and non-constrained language users is less prominent, with only three features displaying statistically significant differences. Specifically, TE and EFL

reports demonstrate an increased use of adjectival modifiers for nouns, while a reduced use of existential *there* and progressive aspect.

This result highlights a significant homogeneity between constrained and non-constrained language varieties concerning reports, which echoes with the result of the multidimensional analysis. Only three out of 69 features demonstrating distinctive distributional patterns in constrained language varieties. Notably, adjective modification for nouns is overrepresented, and this trend is also observed in editorials, indicating a consistent prevalence of the "adjective plus noun" structures in both constrained language varieties across sub-registers.

On the other hand, there is no overlapping in terms of the underrepresented features across editorials and reports. The underrepresentation of two structures in reports may be attributed to the lack of priming in EFL and translation production. Firstly, the use of existential *there* is commonly found in informational writing to introduce new topics or entities with minimal additional information (Biber, 1988). In contrast, Chinese employs "have (*you*)" to fulfill the same function, a structure resembling that of a possessive sentence, with the locative expression acting as the subject of the sentence (Chan, 2004, p.59). Moreover, the progressive aspect is frequently utilized to emphasize the ongoingness of current or past events. Combining the progressive aspect with the meaning of the past introduces "an additional level of complexity for L2 learners" (Hinkel, 1992, p. 566). As a consequence, non-native English reporters and translators, despite being professionals with an advanced level of language proficiency, may exhibit deviations in their usage of this structure from its native production use.

The underrepresentation of the aforementioned features may also be influenced by the relatively fewer direct quotations in EFL and TE reports than in NE reports where these structures are recurrently found, as illustrated in Excerpt (5).

Excerpt (5):

"*There has never been* a better time to be on the Paralympic circuit," he told PA Sport. "The awareness is phenomenal and the number of races we are being offered is great. *There is* a huge need for races like the Paralympic World Cup -- especially following on from Beijing which was such an amazing experience." (NE-report-061)

5.2.2 Distinctive patterns of EFL

Table 5.3 Distinctive features of EFL editorials.

Over-represented features	Under-represented features
CAUSE	BEMA
CUZ	CONT
EXIST	DOAUX
PEAS	EMPH
VBG	
VBN	
PASS	
ADVCL	

The results of statistical tests reveal distinct patterns for 12 features in EFL editorials compared to TE and NE, including an overrepresentation of eight features and an underrepresentation of four features. These features are detailed in Table 5.3.

Specifically, eight features are significantly more prevalent in EFL editorials. These include two categories of verb semantics (existential or relationship verbs[11], and facilitation or causative verbs[12]), one conjunction (causal conjunctions[13]), four verb-related features (perfect aspect, passive structures, non-finite *-ing* verb forms, and non-finite *-ed* verb

[11] Following Biber (2006: 247, based on the LGSWE, pp. 364, 369, 370–371), existential or relationship verbs are assigned to all forms of the following verbs: *seem, stand, stay, live, appear, include, involve, contain, exist, indicate, concern, constitute, define, derive, illustrate, imply, lack, owe, own, possess, suit, vary, deserve, fit, matter, reflect, relate, remain, reveal, sound, tend* and *represent*. This variable does not include the copular *be*. *Look* was removed from Biber's original list because it frequently acts as an activity verb, too, e.g., *I was looking for my glasses* (Le Foll, 2021).

[12] Following Biber (2006: 247, based on the LGSWE, pp. 363, 369, 370), facilitation or causative verbs are assigned to all forms of the following verbs: *help, let, allow, affect, cause, enable, ensure, force, prevent, assist, guarantee, influence, permit* and *require* (Le Foll, 2021).

[13] These include *as a result*, *on account of*, f*or that/this purpose*, *thanks to*, *to that/this end*, consequently, *in consequence*, *hence*, *so that*, *therefore, thus* (Le Foll, 2021).

forms), and one syntactic feature (adverbial clauses[14]). Meanwhile, four features are significantly less frequent in EFL editorials. Among these, three are verb-related: the static verb form *be*, the auxiliary verb *do*, and verbal contractions. The remaining feature is emphatics, a type of adverbial modifiers used for emphasizing the tone.

A general glimpse of these features signifies a verb-related distinctiveness in EFL. Firstly, the presence of the two categories of verb semantics in EFL editorials reveals insights regarding their emphasis. Existential or relationship verbs are those that report a state of existence or a logical relationship (Biber et al., 2002, p.461), and facilitation or causative verbs are used to express actions that enable or cause something to happen. The overrepresentation of these features indicates that EFL editorials predominantly focus on providing information on existence, possession, states, and relationships, especially, the cause-and-effect relationship between actions and consequences. The emphasis on logical and causal relationships is further reflected in the overrepresentation of causal conjunctions in EFL editorials. Meanwhile, the use of the perfect aspect may be more prevalent when authors aim to emphasize the ongoing relevance or consequences of past events or actions, as it marks actions in past time with "current relevance" (Quirk et al., 1985, p.189 ff).

Meanwhile, the most common function of non-finite *-ed* and *-ing* verb forms is to modify nouns. They can serve as stand-alone words modifying nouns directly (e.g., "*developing* countries", "a *marked* rise"). They can also be regarded as part of a participial clause or non-finite relative clause (e.g., Excerpt 6 and 7).

Excerpt (6):

Possessing the richest carbon emission resources, China has also become the largest carbon emission cutter under the CDM mechanism. (EFL-ed-003)

[14] An adverbial clause modifier is a clause which modifies a verb or other predicate (adjective, etc.), as a modifier not as a core complement. This includes things such as a temporal clause, consequence, conditional clause, purpose clause, etc. (Le Foll, 2021).

Excerpt (7):

The underlying reason is largely decided by the nature of the economic crisis *caused* by overproduction. (EFL-ed-008)

The sentences in the excerpts illustrate how EXIST verb (*possess*) and CAUSE verb (*cause*) in *-ing* and *-ed* forms are used to express the logical relations and provide important contextual information. For example, in Excerpt (6), the participial clause "possessing the richest carbon emission resources" highlights the logical relation between China's possession of these resources and its status as the largest carbon emission cutter under the CDM mechanism. Similarly, in Excerpt (7), the participial clause "caused by overproduction" modifies the noun phrase "the nature of the economic crisis", indicating the logical relation between the underlying reason and the economic crisis caused by overproduction.

Passive structures are associated with a more detached style. The use of passive structures helps create a sense of objectivity and neutrality, and maintains an impersonal and authoritative tone. As illustrated by the following excerpt, "is said to" distances the writers from making direct claims, shifting the focus away from specific individuals or entities to emphasize the information itself. Similarly, the passive structure in the second sentence detaches the subject from the action, emphasizing "taking measures" itself rather than responsible individuals or entities. This maintains objectivity and presents the idea as a general recommendation rather than a personal opinion.

Excerpt (8):

The US *is said to* hold 17 percent of the IMF vote share and the European Union 32 percent. ⋯Simultaneously, measures should *be taken* to push for the development of the SDR as a super-sovereign international reserve currency. (EFL-ed-009)

Adverbial clauses serve various functions, providing additional information like time, place, manner, frequency, and degree. Both adverbial modifiers and adverbial clauses

fulfill similar roles in sentences, but they differ in form. Adverbial modifiers are typically single words or phrases, while adverbial clauses contain a finite verb and often start with conjunctions like *when*, *because*, *although*, and *in order to*. The prevalence of such adverbial clauses in EFL editorials may suggest that the content being discussed is more complex, requiring detailed explanations or in-depth analysis.

EFL editorials show significantly fewer occurrences of four linguistic features. Three of these are verb-related: the static verb form *be*, the auxiliary verb *do*, and verbal contractions. The deficiency of these verb features may suggest a preference for stronger and more assertive verbs to convey messages directly.

Specifically, the main verb *be* is often used to associate an attribute with the subject or to locate the subject's position in space or time (Biber et al., 2002, p.140ff). The auxiliary verb *do* is commonly employed to facilitate negation or interrogatives and can act as a pro-verb. These verbs are less preferred in comparison to more specific and meaningful verbs, especially in full verb forms instead of contracted forms. The avoidance of contracted verb forms also indicates a tendency to maintain a more explicit and formal tone, emphasizing arguments effectively through the use of full verb forms.

On the other hand, emphatics is a type of adverbial modifier used for emphasizing tone. Emphatics add emphasis to statements and are often associated with informal and colloquial discourse. EFL editorials' avoidance for emphatics may suggest a deliberate choice to prioritize more precise and authoritative language.

Table 5.4 Distinctive features of EFL reports.

Over-represented features	Under-represented features
LDE	CONT
NN	DT
APPOS	MDWO
NUM	RP

Table 5.4 lists the eight linguistic features that exhibit distinctive distributions in EFL reports compared to NE and TE reports, with statistically significant differences observed.

Specifically, EFL reports see an overrepresentation of four features, including one general text property (lexical density), total nouns, and two noun modifiers (appositive modifiers and numeric modifiers). Meanwhile, EFL reports are characterized by an underrepresentation of four features, including two verb-related features (particles and verbal contractions), determiners, and modal verb *would*.

The overrepresented features appear to be closely related to noun phrase complexity. In EFL reports, there is a notable increase in the complexity of noun phrases, represented by higher frequency of appositive and numeric modifier for nouns, which may lead to denser packaging of information.

Excerpt (9):

Consequently, Mengniu posted a net loss of *948.6 million yuan* in 2008, its first loss since listing in Hong Kong in 2004. (EFL-report-021)

Excerpt (10):

"The fragrance of winter's plum blossom comes from the bitter cold." - Ancient Chinese proverb. Li Xianzhang, *a member of the China Arts Association and famed master of rooster and peony paintings*, must have a deep understanding of that insight. (EFL-report-021)

From an alternative perspective, EFL reports tend to adopt a more "report-like" style, presenting detailed information on quantity (e.g., Excerpt 9) and the identity or status of the referents (e.g., Excerpt 10) using numeric and appositive modifiers. These frequently used features are commonly found in report writing, as shown in the above excerpts. Additionally, EFL reports show a greater proportion of content words, contributing to the higher lexical density. Similarly, the underused features in EFL reports also align with this tendency. There is a reduced usage of function words such as determiners (e.g., *a*, *the*, *both*, *either*, *another*, *each*) and particles, as well as fewer occurrences of verbal contractions, contributing to an elevated level of formality. Additionally, there are fewer modal verb

would, which is often adopted to express predictions or personal volition in hypothetical situations.

5.2.3 Distinctive patterns of TE

Table 5.5 Distinctive features of TE editorials.

Over-represented features	Under-represented features
MDWS	COMM
NNP	IN
	MDCO
	PPMOD
	APPOS
	NFRC

TE editorials exhibit distinctive usage patterns in eight linguist features compared to EFL and NE editorials, and these differences hold statistical significance ($p<0.05$). These features are outlined in Table 5.5.

Notably, TE editorials display a higher frequency of two features, namely, predicative modals *will* and *shall*, and proper nouns. Conversely, TE editorials demonstrate a reduced usage of six features. Among these features, one is related to verb semantics (communication verbs), and three pertain to noun modifying structures (prepositional post-modifiers, appositive modifiers, and non-finite relative clauses). The remaining two are prepositions, and the modal verb *could*.

TE editorials demonstrate prevalence of proper nouns, which are used to refer to specific countries (e.g., “Australia”), individuals (e.g., “Queen Elizabeth”), organizations (e.g., “the Indian Ministry of Defense”), time (e.g., “Monday”), locations (e.g., “Bay of Bengal”), and events (e.g., “Exercise Malabar 2020”). TE editorials also use modal verbs *will* and *shall* more frequently than NE and EFL editorials. These modal verbs share functions of making predictions and indicating intentions, primarily in the context of future events. A closer examination of the TE editorials reveals that *will* is much more prevalent than *shall*, and it

is often used to convey a sense of determination (e.g., Excerpt 11) and confidence in prediction (e.g., Excerpt 12).

Excerpt (11):

The Chinese people cherish peace and desire stability, but they *will* never sit idly by while China's sovereignty, security and development interests are undermined, and if such a situation arises, they *will* certainly deal with them head-on. (TE-ed-010)

Excerpt (12):

India *will* find that China *will* pose no threat to it in the Indian Ocean, and there are more areas of cooperation than differences between the two countries on the international stage. (TE-ed-009)

The underrepresented features in TE editorials appear to be primarily associated with noun phrases, particularly concerning nouns with post-modification, represented by prepositional phrases as post-modifiers, appositive modifiers, and non-finite relative clauses. The following examples show a typical use of these structures in EFL and NE editorials. As exemplified in the excerpts, the most common function of appositive modifiers (e.g., "Sir Fred Goodwin, *the chief executive*") is to identify the individuals or entities. In other times, prepositional phrases (e.g., "wishes *of many*", "voters *in three countries*", and "engagement *with Europe*"), appositives and non-finite relative clauses are often used to provide additional information and clarification (e.g., Expert 13 and Excerpt 14). TE editorials are less reliant on these noun phrase structures.

Excerpt (13):

Energy efficiency, *the right to reduce carbon emissions*, has become an asset for countries to fight for during the last 10 years. (EFL-ed-003)

Excerpt (14):

And who exactly would talk, when this religious movement lacks a Sinn Fein*, a political arm doing the thinking*? (NE-ed-011)

TE editorials also display a reduced frequency of the modal verb *could*. In writing, *could* serves various functions, including expressing possibility, ability, permission, and occasionally referring to past time (Biber et al., 2002, p.176ff). However, it is sometimes perceived as less forceful or assertive compared to other modal verbs like *should* or *will*. In contexts where a stronger stance or a more direct message is desired, the use of *could* might be less frequent. Additionally, the occurrence of communication verbs is notably lower in TE editorials. Communication verbs belong to the broader category of activity verbs and specifically pertain to verbs that describe speech and writing actions, such as *describe*, *tell*, *ask*, and *claim* (Biber et al., 2002, p.107). The decreased usage of communication verbs in TE editorials might suggest a relative absence of direct speech or quoted statements compared to EFL and NE editorials. Besides, a noticeable decrease in the use of prepositions is observed in TE editorials. Prepositions play a crucial role in facilitating a wide range of sentence structures and serve diverse functions in language. Thus, a lower frequency of prepositions in TE editorials does not lend itself to a straightforward interpretation.

Table 5.6 Distinctive features of TE reports.

Over-represented features	Under-represented features
ACT	MDWS
PLACE	THATD
XX0	NNP
PP3	
ADVCL	
ADVMOD	

Table 5.6 presents the nine features that exhibit statistically significant differences in their distribution in TE reports compared to EFL and NE reports. TE reports are characterized by an overrepresentation of six features and an underrepresentation of three features.

The overrepresented features in TE reports include one verb semantics (activity verbs), negations, third person pronouns, adverbial clauses, and adverbial modifiers, especially place adverbials. The underrepresented features include modal verbs *will* and *shall*, proper nouns, and one syntactic feature (subordinator *that* deletion).

One notable observation in TE reports is the prevalence of adverbial structures. Compared to NE and EFL reports, TE reports incorporate a significantly higher number of adverbial modifiers and adverbial clauses, and they make more frequent references to places, potentially aiming to provide a richer contextual background for the reported events.

Another distinguishing characteristic of TE reports is their greater usage of third person pronouns combined with fewer occurrences of proper nouns. Pronouns are deictic expressions whose interpretation relies on the shared knowledge between writers and readers. It appears that TE reports may display a preference for substituting proper nouns with pronouns to create textual cohesion.

However, TE reports exhibit fewer cases of subordinator *that* deletion, which is often considered an indicator of structural explicitness in language (Kruger, 2019). The presence of subordinator *that* is associated with a more formal and explicit writing convention. Meanwhile, it helps establish clear syntactic relationships between clauses, which may ease the cognitive load in language production and comprehension. This choice to retain the subordinator *that* more often indicates an increased effort of translators to maintain clarity and precision in the target language, aligning with the conventions of standard English. However, it should also be noted that the presence of subordinator *that* is conditioned by various factors related to the matrix clauses, e.g., the subjects and the overall complexity level. For instance, it is reported that high-frequency verbs are less likely to be followed by subordinator *that* (Wulff et al., 2014). The fact that TE reports involve fewer communication verbs, including some high-frequency verbs like *know* and *say*, may contribute to the observed lower frequency of *that* deletion.

In addition, TE reports also differ in their use of modal verbs *will* and *shall*, which suggests a reduced focus on predictions or intentions pertaining to future events. Meanwhile, TE reports employ more activity verbs that refer to volitional activities, indicating a writing style that is more oriented towards actions and events. These preferences potentially reflect the intention of the translators to emphasize the immediacy and relevance of reported events rather than speculating about future developments.

5.3 Summary

This section investigates the variations at the feature level of the two constrained varieties of English, spanning across two specific sub-registers of news writing. The analysis unveils the distributions of the linguistic features loading on the identified textual dimensions across the language varieties. Furthermore, it identifies the linguistic features that are distinctively distributed in EFL and TE.

Overall, the examination does not reveal significant distributional disparities between the constrained and non-constrained varieties of English, evident by a small number of features exhibit distinctive patterns across sub-registers. What emerges from this observation is that the distributions of individual features are sensitive to register, as unique patterns manifest within the two sub-registers. The register-specific variations at the feature level of the two constrained varieties are quantitative, albeit nuanced and subtle.

However, this analysis is not without its limitations. Since individual features might correspond to multiple textual properties, they should be considered in conjunction with the dimensional variations identified by the multidimensional analysis. Additionally, the broad categorization of some selected features might also obscure specific differences of individual elements under a particular category. For instance, a preliminary qualitative probe identifies some contrasts in the types of adverbials used between constrained and non-constrained varieties. More reliable conclusions can only be drawn upon more in-depth qualitative investigation.

Despite these limitations, the feature-level analysis supplements the multidimensional perspective. By grouping over- and under-represented features in the constrained varieties, it uncovers new distinctive clusters of features that shed light on the characteristics of constrained varieties in contrast to the non-constrained one. This adds a novel, complementary layer to the understanding of textual dimensions previously delineated by the multidimensional analysis.

Chapter 6 Discussion

6.1 Implications for constrained language universals

The feature-level analysis uncovers distinctive distributional patterns of linguistic features in the three varieties, providing insights into the typicality of each variety under examination. Specifically, over- and under-represented linguistic characteristics in both TE and EFL are identified, revealing traits that set them apart from non-mediated, native English production. These findings carry significant implications, underscoring the similarities between TE and EFL. This section delves into the implications of these commonalities, especially in relation to two widely explored "universals" in translation and SLA studies: simplification and explicitation. The discussion will assess whether they could be considered as "universals" in constrained English varieties.

6.1.1 Simplification hypothesis

The simplification hypothesis appears to be partially contradicted by the feature-level variations identified, particularly at the lexical and phrasal aspects. Compared to NE, TE and EFL editorials exhibited a higher level of lexical complexity, indicated by higher average word length, greater lexical density, and increased use of nouns. These intriguing findings are in contrast to previous studies that have generally shown translations to have lower lexical complexity, often measured through lexical density and average word length (Laviosa, 1998, 2002; Xiao, 2010; Bernardini et al., 2016). Additionally, non-native production in SLA studies consistently demonstrates lower lexical diversity compared to native production (Jarvis, 2002; McWhorter, 2007, 2011). However, it is worth noting that contradictory findings have been reported when examining different language pairs and text types (Pastor et al., 2008; Grabowski, 2013; Ferraresi et al., 2019).

Conversely, the results concerning syntactic simplification in translation studies are less definitive (Liu et al., 2022), which is consistent with the findings of the current study. TE and EFL editorials are characterized by a higher level of syntactic complexity, evidenced

by a marked increase in noun phrases embedding adjectival modifiers, split structures, and coordination structures. However, the underrepresentation of relative clauses indicates a reduced level of syntactic complexity in terms of this particular clausal structure.

6.2.2 Explicitation hypothesis

The findings offer partial support for the explicitation hypothesis, which proposes that there is an "explicitation of information through elaboration and specification" (Kruger & Van Rooy, 2016b, p.26).

In terms of individual linguistic features, both TE and EFL editorials display lower frequencies of function words with a deictic function. These deictic expressions encompass personal, temporal, and spatial ones, represented by features such as determiners (quantifiers and demonstrative pronouns), pronouns (first person pronouns, third person pronouns, and quantifying pronouns), adverbial modifiers and particularly time adverbials. The reduced usage of these deictic expressions may imply that constrained language users assume little shared context with the addressee of the texts. Notably, the combination of underusing these deictic features and overusing nouns implies that constrained language users rely less on deictic expressions and instead employ nouns to explicitly identify the subjects under discussion.

Explicitation in constrained language use has been shown in previous studies, evidenced by more frequent occurrence of the subordinator *that* (De Sutter & Lefer, 2020; Kruger & De Sutter, 2018), heightened use of cohesive devices such as demonstrative pronouns (Kajzer-Wietrzny, 2021; Rabinovich et al., 2016), and a preference for analyticity over syntheticity (Szmrecsanyi & Kortmann, 2011). Moreover, an underrepresentation of adverbs has been identified in non-native English production, suggesting a reduced reliance on situation-dependent referencing (Van Rooy et al., 2010). Consistent with prior research, our findings also indicate limited use of context-dependent referencing and a preference for explicit referencing in both TE and EFL editorials.

However, the simultaneous underuse of relative clauses and overuse of noun phrases with adjectival modifiers deviates from the general trend of explicitation. Relative clauses, as clausal expansions, are considered to convey information in a more elaborate manner by explicitly stating the relationship between propositions (Biber et al., 2002). In this regard, the constrained language varieties appear to be less explicit regarding the use of noun-modifying structures. This finding suggests that explicitation in constrained language use might not be uniform across all linguistic features.

6.2 Textual and feature-level variations as consequences of constraints

The five-dimension constraint matrix introduced in Chapter 3 has illuminated the main difference between constrained and non-constrained language varieties under examination. The primary distinction resides in the first dimension "language activation": bilingual activation presents in both EFL and TE but does not in NE. Additionally, the status of mediation emerges as the key factor that differentiates TE from EFL. While this current study does not aim to be explanatory, the analysis does enable some assumptions. Specifically, it's possible to infer how these shared and unique constraints might interact in a complex manner with each other, leading to the textual and feature-level variations within the examined constrained varieties of English.

6.2.1 Bilingual activation: interference versus normalization

Language activation encompasses various aspects that can significantly impact language use. In a bilingual action mode, both linguistic and corresponding cultural systems are co-activated, which may either conflict or reinforce each other in shaping linguistic choices. Factors such as the specific language pair involved, the relative status of the two languages, and the typological and cultural differences between them play essential roles in this interaction. In translation studies, the relationship between different types of translation-related behavior towards source and target language norms is describe "as a continuum

ranging from shining through, i.e., orientation towards source language norms, to normalization, orientation towards target language norms" (Hansen-Schirra & Steiner, 2012, p.272). Various models have been proposed to elucidate the dual forces exerted by the two languages. For instance, in translation studies, Toury's dual law of translation (2012) posits the law of interference and the law of growing standardization, placing emphasis on the cultural dynamic between source and target languages. From a cognitive standpoint, Halverson's (2017) revised Gravitational Pull Hypothesis identifies the gravitational pull from the source language juxtaposed with the target language's magnetism. These forces are also explored in SLA studies under the terms of cross-linguistic influence or transfer (Jarvis & Pavlenko, 2008). In the subsequent discussion, the term "normalization" is employed to describe the pulling force of target language for TE and L2 for EFL, and "interference" is adopted to delineate the influence of the source language for TE and L1 for EFL.

Notably, normalization has been categorized as one of the translation universals "which typically occur in translated texts rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker, 1993, p.243). However, this study adopts an opposing view, which regards normalization as a language-dependent property similar to interference. Specifically, normalization is target-language dependent, while interference is source-language dependent, which aligns with the distinction between T-universals and S-universals proposed in Chesterman (2004). Different from simplification and explicitation which are both language-independent, interference and normalization are used in the current study to address both linguistic and cultural aspects associated with specific language pairs which collectively lead to the variations at both the feature and textual levels.

The feature-level variations seem to stem from a robust effect of interference, particularly covert transfer (Mougeon et al., 2005; Heine & Kuteva, 2005), referring to instances where the distribution of lexical or syntactic features in translation or L2 production differs from those in the original texts. A notable example is the prevalent adjectival modification in constrained varieties, evident across sub-registers. Within the current dataset, constrained

language users are influenced by Mandarin Chinese where noun modification is often placed before the noun head, regardless of the complexity of modifiers (Huang, 1998; Cheng & Sybesma, 2014). This preference for left-branching noun modification in Chinese may transfer into the English produced by EFL writers and translators, leading them to underutilize alternative forms of noun modification, such as phrasal and clausal post-modification, as evident in the underrepresentation of relative clauses in EFL and TE editorials. Similar preference for pre-modification for nouns has been reported in previous studies on Chinese English (Liu et al., 2017) and English translated from Chinese (B. Wang & Zou, 2018).

Covert transfer may also explain another two distinctive patterns in TE and EFL, i.e., more conjunctions and fewer existential *there* in constrained language use, though these distributional patterns do not hold across sub-registers. Coordinating structures are prevalent in Mandarin Chinese, with Chinese speakers frequently employing coordination to conjoin parallel items in meaning, function, and form (Xu, 2010). Xu (2010) identified the parallel use of multiple conjunctions within a sentence as one of the prominent features of China English. Meanwhile, the lower frequency of existential *there* may be the result of a lack of equivalent structures for expressing existence in Chinese. Similar patterns concerning this structure have been observed in previous studies on French-English translation, which found the difference in the frequencies of existential *there* in English and its equivalence *il y a* in French leads to different distributions of the latter structure in translated and original French (Cappelle & Loock, 2013).

At the textual level, interference prominently affects Dimension 1, 'Elaborated-involved versus Integrated-formal production', where both TE and EFL diverge from NE. Specifically, while NE editorials tend to be elaborate and involving, constrained ones are more formal. This corroborates previous comparative studies on English editorials in China and western countries (e.g., Huang & Ren, 2020; Wang, 2008a, 2008b), which identified a more formal and impersonal style in Chinese editorials. This contrast may be attributed to the difference in the writing conventions of argumentative writing between English and Chinese (Wang, 2008b). Editorials is argumentative writing in essence, and in Chinese

public writing such as commentary or argumentative writings, personal voice is limited "to maintain social harmony and to inform readers of established opinions rather than negotiate opinions with readers" (Wang, 2008b, p.184). Such conventions may influence English writing by native Chinese speakers and translations from Chinese into English.

On the other hand, the effect of normalization seems to be more pronounced at the textual level, evident in the results on Dimension 2, Dimension 4, and Dimension 6 where both TE and EFL display an inclination to exaggerate typical NE features. NE editorials display characteristics of an evaluative tone and the use of irrealis, while NE reports are characterized by depictive narration with spatial-temporal focus and information density. In comparison, both TE and EFL editorials accentuate these features, exhibiting a heightened level of evaluative tone and irrealis usage. Likewise, TE and EFL reports extend the characteristics of NE reports, with more pronounced depictive narration, spatial-temporal focus, and increased information density. In other words, there is a preference in the constrained varieties for the prototypical and salient textual features of non-constrained variety. As observed in Xiao & Cao (2013) on research abstract produced by native and non-native language users, "L2 writers normalize with respect to what they believe to be typical genre-specific norms and writing conventions holding in L2" (Lefer & Vogeleer, 2013, p.8). The finding of the current study shows that this statement also applies to translators, though the degree of normalization varies between the two groups.

Previous research has illuminated the dynamic nature of these opposing forces in language production. The extent to which these forces exert influence depends on various factors, such as the professional expertise of the translators (Dimitrova, 2005; Lapshinova-Koltunski et al., 2022) and the specific text register (Toury, 1995; Kunilovskaya & Corpas Pastor, 2021). The present study resonates with these findings, emphasizing that the impact of these two forces on EFL and translation is sensitive to register. Moreover, the current research suggests that these forces affect different linguistic features and textual dimensions in diverse ways.

6.2.2 Interwoven constraints: register and bilingual activation

Register refers to a collection of texts that exhibit specific linguistic features, reflecting a particular situational context and a unifying communicative purpose (Biber & Conrad, 2009, p.6). This concept is vital in understanding how language adapts to various communicative scenarios. In the current analysis, the nuanced interaction between register and the constraint of bilingual activation is demonstrated mainly in two aspects.

First, constrained language variations differ across sub-registers. While distinctive distributions of features can be observed in both sub-registers, only a single feature stands out as being statistically overrepresented in both TE and EFL across the two sub-registers. This observation underscores two crucial aspects: on the one hand, constrained and non-constrained language varieties under examination are largely homogeneous at the feature level. On the other hand, the characteristics of constrained language varieties is sensitive to specific sub-registers in question. This sensitivity extends beyond the feature-level and is also reflected on the textual dimensions, as the textual characteristics shared by TE and EFL across dimensions also vary in the two sub-registers.

Such findings align with earlier research that recognizes register as a significant factor in shaping constrained language varieties such as translation and L2 production (Kruger & Van Rooy, 2012, Neumann, 2014; Delaere, 2015). Notably, although current study focuses on news registers, it shows that newspaper discourse is not a single or homogeneous entity (Biber & Conrad, 2009). Different sub-registers, specifically editorials and news reports, reveal significant variations. Seen from a functionalist perspective, news reports or hard news aim to report on events and inform readers, often using quoted materials or attribution to convey multiple views to maintain a "factual and objective" tone (White, 2000, p. 379). In contrast, editorials, driven by the specific purpose of stating an opinion and argumentation, may adopt different linguistic features.

Moreover, the data reveal more variations in constrained editorials compared to constrained news reports at both feature and textual levels. For example, 14 features differentiate TE

and EFL from NE editorials, compared to only three for reports. Similarly, at the textual level, the two constrained varieties demonstrate more variations in their factor scores for editorials than reports, with sharper contrasts between the two and NE.

These results suggest that the more distinct a register's performances are across two languages or cultural systems, the more variation is exhibited in constrained languages, which corroborates previous studies (e.g., Nikolaev et al. 2020; Kunilovskaya & Pastor, 2021). Defined as "a culturally recognised artifact" (Lee, 2001, p.46), register is closely connected to social conventions, displaying recognizable features that may vary across linguistic and cultural systems (Santini, 2007). This leads to potential variations in writing conventions even within the same register. Particularly, as an editorial embodies the author's cultural perspectives, it manifests differences across cultures in terms of lexical choices, syntactic structures, and rhetorical organization (see Ansary & Babaii, 2009). The current analysis supports these observations, indicating a relationship between the disparity in register performance across linguistic or cultural systems and the degree of variations among language varieties.

6.2.3 Mitigating effect of mediation

Disparities can be discerned between the two constrained varieties of English, particularly at the textual level. These discrepancies seem intrinsically relate to the mediation status of translation, which is considered a prominent difference between the two constrained varieties and is thus posited as a factor leading to these disparities.

For instance, the mediation status of TE appears to lessen the impact of normalization. Though EFL and TE demonstrate parallel tendencies on D2, D4, and D6 compared to NE, EFL notably diverges more on D4 ('Descriptive narration with a spatial-temporal focus') and D6 ('Information density versus Irrealis') by exhibiting a larger score difference. In essence, while both TE and EFL tend to exaggerate typical NE features, it is EFL that takes a more extreme approach. Additionally, EFL displays a contrasting tendency on D5 ('Activity focus versus Referential precision') compared to NE and TE.

This pattern reveals that compared to EFL, TE manifest a more consistent tendency to exaggerate typical features of non-mediated, native writing, with a relatively narrower range of deviation. Conversely, EFL's deviations from NE are more pronounced, and occasionally present inconsistent patterns. This outcome may imply that translators, perhaps due to their professional status as mediators, align more closely with the writing conventions of the target language. The alignment may be a result of heightened awareness of subtle linguistic and cultural differences, coupled with increased adaptability. This finding corroborates Kotze (2020:117), which stated that "the high degree of linguistic proficiency, biliteracy, task expertise and professional training may lead to particularly aware of CLI (crosslinguistic influence) and develop conscious strategies to avoid such effects."

Although TE generally aligns more closely with NE writing conventions, the precise impact of mediation appears to be more intricate, as contrasting patterns are evident on D1 where TE diverges more significantly from NE. These findings underline that the influence of mediation is multifaceted, warranting further in-depth exploration to uncover the nuanced mechanisms driving the differences between TE and EFL.

Chapter 7 Conclusion

The study initiated an in-depth exploration into the textual variations of two constrained varieties of English, Translated English (TE), and English as a Foreign Language (EFL), in contrast to non-mediated, native English writing (NE). TE and EFL are hypothesized to display similar textual characteristics due to their shared constraints, most notably the bilingual activation involved in language production which distinguishes them from NE. Meanwhile, the mediation aspect of translation defines the difference between TE and EFL, presumed to cause divergence between the two constrained varieties.

To verify these assumptions, a corpus-based approach is adopted to contrast English translated from Chinese and EFL writings by native Chinese speakers, with the non-mediated native English production as a benchmark. Based on the constrained language framework, various constraints such as register, mode of production, language proficiency, and the task expertise level of the language users are acknowledged as intertwined constraints in conditioning language use. For this reason, the current investigation narrows its focus, concentrating on writings and translations produced by professionals with advanced level of language proficiency and professional expertise within the context of published news writing including editorials and news reports.

The textual variations of the two constrained English varieties are investigated through a multidimensional analysis, which illustrates their commonalities and divergences regarding the co-occurring patterns of a group of 69 features at various linguistic levels. This macro-level examination is supplemented by a micro-level analysis of the distributional patterns of linguistic features, aimed at gathering finer details on how textual variations are achieved through feature-level nuances and providing insights into the characteristics of constrained language use at the feature level. Following this dual approach analysis, the study moves on to interpret the implications of these findings. It examines whether the observed commonalities imply any “universal features” of constrained language use and explores how these attributes might be mapped onto the

shared and distinctive constraints of TE and EFL. The ensuing section outlines the major findings addressing the research questions posed in Chapter One, elucidates the broader impact of the investigation, acknowledges its limitations, and discusses potential avenues for future research.

7.1 Major findings

7.1.1 Textual-level variations of constrained language

To address the first research question, namely, the exploration of textual characteristics of two constrained varieties of English (represented by Translated English, or TE, and English as a Foreign Language, or EFL) as compared to NE, a multidimensional analysis was conducted. Specifically, the co-occurring patterns of 69 meticulously chosen linguistic features at the lexical, syntactic, and textual levels were identified through exploratory factor analysis. The results led to the extraction of six factors, accounting for approximately 40% of the total variance in the language varieties. Subsequently, six dimensions were interpreted based on the functions that were commonly shared by the co-occurring features associated with each factor.

The similarities and differences between the two constrained varieties of English in contrast to NE were understood through the relationships of the factor scores of the three varieties. The examination of these factor scores reveals a strong register effect, indicating that the textual characteristics of the three language varieties are significantly different in the two sub-registers under examination, namely, editorials and news reports. An ensuing regression analysis not only confirmed this register effect but also revealed that it is challenging to generalize any consistent textual characteristics of TE and EFL across the sub-registers.

Intriguingly, however, the analysis uncovers notable similarities between the two constrained varieties in several compelling ways. Firstly, both constrained varieties frequently exhibit a shared tendency to exaggerate the typical textual features of the non-

constrained variety. This pattern reveals that the conventional features of certain registers are not merely adopted but are also intensified in both TE and EFL. Moreover, this intensification appears to be more pronounced in EFL, suggesting subtle differences between these two constrained varieties. On certain textual dimensions, however, both constrained varieties share patterns opposite to that of the non-constrained one. Below is a summary of the textual variations of TE and EFL compared to NE along the six identified dimensions, followed by explanations on how these textual variations are embodied in the specific linguistic features that define each dimension.

Dimension 1, labeled as 'Elaborated-involved versus Integrated-formal production', notably differentiates constrained language varieties from non-constrained native production. This distinction is evident as the median scores of TE and EFL pattern closely together, contrasting with NE. However, these distinctive textual features of constrained language varieties are register-sensitive. While NE editorials demonstrate elaborated and involved production, NE reports are more aligned with integrated and formal production. TE and EFL both demonstrate a reverse trend, exhibiting integration and formality in editorials, and elaboration and involvement in reports.

In TE and EFL editorials, integration and formality materialize through the co-occurring reduced usage of positive features on D1 (e.g., adverbial modifiers, quantifiers, relative clauses, downtoners, first person pronouns, and quantifying pronouns, etc.) and an increased level of negative features (e.g., lexical density, average word length, total nouns, and occurrence verbs, etc.). Conversely, the tendency towards elaboration and involvement in TE and EFL reports are embodied by an increased presence of positive features (e.g., downtoners, possessive modifiers, and mental verbs) and reduced use of negative features (e.g., progressive aspect, proper nouns, and passive structures).

Dimension 2, titled 'Evaluative discourse versus Reporting/retelling discourse', reveals a clear distinction between the two sub-registers, reflecting the nature of this textual dimension. All varieties show a tendency towards evaluative discourse in editorials, but this inclination is more pronounced in the constrained varieties, particularly in EFL. For

reports, both constrained and non-constrained varieties demonstrate reporting and retelling characteristics, but EFL reports are distinguished by an increased focus on evaluative tone.

The more pronounced evaluative tone in TE and EFL editorials is discernible at the feature level, demonstrated by the concurrent presence of more positive features on D2 (e.g., necessary modal verbs, pronoun *it*, present aspect, nominalization, split structures, and adjective modifiers) and fewer negative features (e.g., complement clauses, third person pronouns and past tense). For EFL reports, the more evaluative tone is achieved with the grouping of more positive features (such as modal verbs *can*, *will* and *shall*, and superlatives) and fewer negative features such as third person pronouns and past tense.

Dimension 3, which captures 'Depictive and detailed narration', reveals differences among the three language varieties. In the context of NE, editorials exhibit a subtle positive trend in being more depictive and detailed, whereas reports follow a marginal negative trend. In EFL, both editorials and reports reveal a more pronounced trend in the same direction as NE. Conversely, TE distinguishes itself uniquely by manifesting inverse trends. Specifically, TE editorials are characterized by a reduction in depictive and detailed narration, while TE reports display an increased tendency for detailed narration.

EFL editorials' tendency towards detailed narration is signified by the co-occurrence of increased usage of positive features on D3 such as adverbial clauses, non-finite *-ed* verb forms, and non-finite *-ing* verb forms, while TE editorials show the opposite trend, demonstrated by a lack of these positive features and a significantly higher frequency of the negative feature of proper nouns. In contrast, EFL reports align with NE, exhibiting a diminished use of a group of positive features (e.g., non-finite relative clauses, particles, causal conjunctions, concessive conjunctions, and existential verbs). While the enhanced depictive narration in TE reports is achieved by more frequent use of the above positive features as well as predictive and attributive ajectives, and prepositional modifiers.

Dimension 4, labeled as 'Descriptive narration with a spatial-temporal focus', highlights a distinction between constrained and non-constrained language use, and meanwhile

showcases a clear register effect. All three varieties exhibit descriptive narration with a spatial-temporal focus in reports, and this feature is more pronounced in the constrained varieties. Conversely, all three display a lack of this focus in editorials, with this tendency being especially strong in EFL.

The emphasis on spatial-temporal information in both TE and EFL reports is evident by the grouping of fewer nominalizations and more positive features on D4 such as numeric modifiers and hedges. Additionally, TE reports show higher scores contributed by more place adverbials, verbal contractions, and concessive conjunctions, while EFL reports rely more on appositional modifiers. Conversely, both TE and EFL editorials display lower frequencies of positive features (e.g., time adverbials, second person pronouns, and hedges) and higher frequency of nominalizations. The more pronounced tendency in EFL reports is discernable through reduced frequencies of positive features such as verbal contractions and place adverbials.

Dimension 5, tentatively labeled as 'Activity focus versus Referential precision', reveals contrasts between the two constrained language varieties. TE aligns with NE showing similar patterns: reports are characterized by a heightened emphasis on activity focus, while editorials prioritize referential precision, though the tendencies are more prominent in TE. Conversely, EFL display different patterns, with editorials showing more activity focus and reports marked by greater referential precision.

The divergence between TE and EFL on D5 could be observed at the feature level. The activity focus in EFL editorials is embodied by the co-occurring higher frequencies of positive features (e.g., noun compounds, activity verbs and modal verb *could*) and lower frequencies of negative features (e.g., proper nouns and copula verb *be*). Conversely, TE editorials is marked by referential precision, demonstrated by significantly more negative features such as proper nouns and fewer positive features (e.g., modal verb *could* and activity verbs). TE reports are marked by an increased activity focus and less referential precision, shown by more positive feature such as activity verbs and fewer negative feature of proper nouns. Contrarily, EFL reports demonstrate a reverse tendency with fewer

positive features such as the modal verb *could* and more negative features such as propositional phrase modifiers.

Dimension 6, tentatively named 'Information density versus Irrealis', identifies differences between constrained and non-constrained language varieties with a register effect. In the context of editorials, the utilization of irrealis is a shared characteristic among all three language varieties, but this feature is particularly pronounced in the constrained varieties, most notably in EFL. In reports, a shared characteristic among the three is information density, with this trend being noticeably stronger in the constrained varieties.

The parallelism between TE and EFL editorials in displaying lower scores is the result of a reduced level of positive features such as mental verbs and type-token ratio. Moreover, EFL editorials employ more negative features such as determiners. Both TE and EFL reports, however, appear to be more informational, marked by an increased level of lexical density and a greater usage of mental verbs, as well as a reduced presence of existential *there* and fewer determiners.

7.1.2 Feature-level variations of constrained language

Overall, there is an observable feature-level distinction between the constrained and non-constrained varieties of English. Approximately one-fifth of all the linguistic features examined are distributed differently across these two categories in editorials, while only three demonstrate distinction in reports. This reveals that the differences between the constrained and non-constrained varieties are register-sensitive. In editorials, the overused features in constrained varieties predominantly relate to linguistic complexity. Specifically, this complexity manifests in three ways: noun phrase complexity as indicated by higher frequencies of nouns and adjectival noun modifiers; syntactic complexity in terms of coordination and split structures; and overall lexical complexity represented by an increased average word length and lexical density. Meanwhile, the underrepresented features are primarily deictic words, such as third-person pronouns, quantifying pronouns, and demonstrative pronouns. This underrepresentation indicates that constrained language

users may rely less on features that require the shared knowledge between readers and writers/translators. In contrast, the distinction between reports produced by constrained and non-constrained language users is less prominent. Specifically, TE and EFL reports demonstrate more use of adjectival modifiers for nouns, with a corresponding reduction in the use of existential *there* and progressive aspect. Only one linguistic feature, adjectival modifiers for nouns, is consistently overused in the two constrained English varieties for both editorials and reports.

In summary, the distributional patterns of individual linguistic features in TE and EFL do not deviate significantly from the non-constrained variety. As found in the multidimensional analysis, the distribution of linguistic features are also register-sensitive, and the constrained and non-constrained varieties exhibit more differences in editorials than in reports. The single consistently overused feature in the constrained varieties reflects a preference for adjectival modification for nouns, which is "the most common building block of complex noun phrase", corresponds to the tendency of information condensation in contemporary English news writing (Leech et al., 2009, p.216), and could also be linked to the preference for premodification in mandarin Chinese.

Despite the general alignment in terms of linguistic feature distribution between constrained and non-constrained varieties, EFL and TE also demonstrate unique patterns respectively. EFL presents distinctive patterns in 12 features for editorials and eight for reports. In editorials, these distinctions appear to be verb-related, represented by overused features such as existential verbs, causative verbs, non-finite *-ing* verb forms, and non-finite *-ed* verb forms, and underused features such as the copula verb *be*, auxiliary verb *do*, and verbal contractions. In reports, the overused features, including lexical density, total nouns, appositive modifiers, and numeric modifiers, point to higher lexical richness and a greater complexity in noun phrases. Only one feature is consistently underrepresented across both editorials and reports, specifically the verbal contractions. This feature has been operationalized as an indicator of explicitness in translation studies, and its underrepresentation may suggest an increased level of explicitness in EFL production.

TE exhibits distinctive patterns in eight features for editorials and nine for reports. In editorials, the two overused features are modal verbs *will* and *shall* and proper nouns, while the six underused features include communication verbs, prepositions, modal verb *could*, prepositional phrase noun modifiers, appositive noun modifiers, and non-finite relative clauses. In reports, the six overused features are activity verbs, negations, third person pronouns, adverbial clauses, place adverbials, and adverbial modifiers; and three underused features are modal verbs *will* and *shall*, subordinator *that* deletion, and proper nouns. Interestingly, no feature is consistently overused or underused across sub-registers. Instead, the two overused linguistic features in TE editorials, namely, modal verbs *will* and *shall* and proper nouns, are underrepresented in TE reports. This distinctive distribution of features may suggest that translators are more attuned to register differences, resulting in translations that exhibit varying linguistic patterns across different registers.

7.1.3 Implications for constrained language universals

The feature-level analysis elucidates distinct distributional patterns of linguistic features in TE and EFL compared to NE, highlighting similarities between TE and EFL in relation to two widely discussed "universals" in translation and SLA studies, namely simplification and explicitation.

With regard to the simplification hypothesis, the analysis partially contradicts prior findings at the lexical and phrasal level. Unlike previous studies (e.g., Laviosa, 1998, 2002; Xiao, 2010) which generally showed less lexical complexity in translations, the current study found that TE and EFL editorials demonstrate a higher level of complexity, marked by increased word length and lexical density. Mixed results are found in syntactic complexity, as TE and EFL are characterized by increased complexity in structures such as noun phrases with adjectival modifiers but a decrease in relative clauses, a finding that aligns with recent research (e.g., Liu et al., 2022).

The findings offer partial support to the explicitation hypothesis. Both TE and EFL editorials show reduced frequencies of deictic expressions such as demonstratives and

pronouns, aligning with previous observations in Kruger and Van Rooy (2016b) and Van Rooy et al. (2010). This reduction, coupled with the overuse of nouns, implies that constrained language users opt for more explicit identification. However, certain deviations from the general trend suggest that explicitation might not be uniform across all features. For instance, in the case of noun modification, relative clauses are underrepresented, and adjectival modifiers are overrepresented. Relative clauses function as clausal extensions and are thought to be more explicit by overtly delineating the connection between propositions (Leech et al., 2009, p.226). In this sense, the constrained language varieties appear to be less explicit concerning the use of noun modification structures.

7.1.4 Variations as consequences of shared and distinctive constraints

The textual and feature level variations of constrained varieties of English reveal the intricate interplay of various constraint dimensions. Particularly noteworthy is the shared constraint of bilingual activation present in both TE and EFL, along with the constraint of mediated production unique to translation. The interaction between these constraints, coupled with the influence of register, underscores the complexity of the linguistic phenomena observed in constrained varieties of English.

The findings illustrate the complexities of bilingual activation which gives rise to two competing tendencies towards the source/first language and target/second language in translation and EFL. At the feature level, interference from Chinese into English may account for the noticeable patterns such as the prevalent use of adjectival modification, more conjunctions but fewer existential *there* in constrained varieties. At the textual level, the more formal writing style in constrained editorials captured by Dimension 1 may be attributed to the different writing conventions in Chinese and English cultures. The effect of normalization is more pronounced at the textual level, with both TE and EFL tend to exaggerate typical features of the non-constrained native production for each sub-register. The findings highlight the dynamic nature of the two opposing forces and their sensitivity to constraint dimensions such as register. Overall, the parallel patterns at both feature-level

and textual-level as the result of multifaceted influence of bilingual activation can be observed in both translated English and EFL.

The findings also demonstrate the nuanced interaction between register and bilingual activation. One key observation is that the variations at both feature and textual levels of the two constrained varieties are sensitive to specific sub-registers under investigation. Such findings align with earlier research (e.g., Ivaska & Bernardini, 2020) recognizing register as a significant constraint dimension in shaping constrained language varieties. Another notable finding is that editorials demonstrate more inter-group variations than news reports, which may be attributed to the diverse functions and writing conventions of argumentative writing in culturally distant languages such as English and Chinese. As a culturally recognized artifact, register connects closely to social conventions, and this connection may lead to variations within the same register due to different cultural influences. Thus, the current analysis also provides support to the claim that the more distinct a register's performances are across two languages or cultural systems, the more variation of constrained languages is exhibited (Nikolaev et al. 2020; Kunilovskaya & Pastor, 2021).

Disparities between the two constrained varieties of English can be observed, particularly at the textual level, intrinsically linked to the mediation status of translation. For instance, the mediation status of TE seems to reduce the impact of normalization and is found to be more consistent in exaggerating native writing features with a narrower deviation range. Conversely, EFL's deviations from NE are more pronounced. This pattern may imply that translators align more closely to the writing conventions of the target language. The analysis also notes that TE's alignment to NE writing conventions is complex and shows contrasting patterns in certain textual dimensions. These findings underline that the influence of mediation is multifaceted, warranting further in-depth exploration to uncover the nuanced mechanisms driving the differences between TE and EFL.

7.2 Significance of the study

A corpus-based methodology has traditionally been employed to examine distinctive characteristics of translation and non-native language production. However, the present research diverges from conventional analyses by incorporating a comprehensive theoretical paradigm known as the constrained communication framework. This framework enables an integrative investigation into two specific varieties of English shaped by bilingualism: translated English and English as a Foreign Language (EFL). By highlighting their commonalities, the research serves to bridge these two areas of study, thereby offering a theoretical vantage point for both areas. This approach also addresses existing gaps in translation studies, which have largely focused on elucidating linguistic differences rather than similarities, and have often overlooked the importance of rigorous statistical methodologies and interdisciplinary collaboration (De Sutter & Lefer, 2020).

Furthermore, the study adopts a multidimensional perspective on language use by conducting a novel Multidimensional Analysis (MDA). This systematic method examines the textual-level characteristics of constrained language, and the study complements it with feature-level variation analysis. The current study utilizes an emerging tagging tool (MFTE) and embraces an unconventional normalisation method for the frequencies of linguistic features (different normalisation baselines for different linguistic features). This innovative approach not only facilitates the identification of the textual variations of constrained English varieties but also seeks to discover nuanced understanding on language dynamics. Methodologically, the study advances a descriptive approach, employing statistical tools to examine the quantitative distributions of linguistic features. It also incorporates text-based exploration to uncover qualitative findings, adding depth to the quantitative analysis. The use of multidimensional analysis and innovative normalisation techniques further exemplifies the potential for new tools and approaches to enrich and expand existing research paradigms.

In addition to extending the theoretical framework and refining the methodology, this study has broader implications for the fields of translation and linguistic studies. By revealing

the underlying commonalities between translated English and EFL, it contributes to a more integrated understanding of bilingualism-influenced communication phenomena. This nuanced view challenges traditional notions that have compartmentalized these language varieties as distinctly separate entities. By elucidating their shared features, the study serves as part of a pioneering initiative that redirects scholarly attention towards the convergence of languages which have been previously conceptualized as divergent. This shift in perspective is not merely an academic exercise but represents a profound reconfiguration of how we comprehend linguistic phenomena.

This exploration into constrained varieties of English also offers practical advantages. By delineating the similarities and differences across English varieties, the research equips news writers and translators with the knowledge to navigate the complexities of writing and translating with greater proficiency. Besides, in an era where writing and translation tasks are increasingly supported by machine and AI assistance, the findings of this study are potentially beneficial for the development of algorithms, so that the production could not only be more sophisticated but also specifically tailored to meet the nuanced demands of diverse linguistic tasks.

From a pedagogical perspective, this study casts light on the intricacies of news writing and translation, serving as a vital resource for students engaged in these disciplines. It challenges the notion that adherence to “native” English standards is the sole marker of legitimacy in writing English as a second or foreign language or translating into English. Instead, the findings advocate for a broadened linguistic awareness, encouraging an appreciation for the unique attributes of various English varieties. This shift in perspective not only enriches students’ linguistic competencies but also prepares them for a globalized communication landscape where multiple English varieties coexist and complement each other.

7.3 Limitations and future directions

7.3.1 Limitations

The current study explores the relations between translation and EFL, with an emphasis on their shared constraint of bilingual activation. Given that the complexity of intertwining constraint dimensions necessitates highly comparable data, the study focuses on a specific language pair, register and professional level of the language users covered by a relatively small corpus. Consequently, it is only possible to make tentative statements rather than more confident assertions.

As further shown by the findings, constrained language use is substantially influenced by register at both textual and feature levels. The linguistic makeup of the constrained language discovered in the study may also be language pair specific. This suggests that research involving different language pairs and registers could potentially yield conflicting results. Thus, for a comprehensive understanding of a phenomenon as complex as constrained communication, it is essential to gather empirical evidence across a broader range of language pairs and registers to reach more robust conclusions.

There is an inherent trade-off between maintaining data comparability and constructing a large-scale corpus encompassing diverse registers and languages. The real-life functioning of language also needs consideration, especially in the context of Chinese - English language pair, limiting the data further. The current investigation does not aim to provide a holistic picture that could be generalized to constrained communication as a whole. Instead, it serves as an initial attempt to shed light on a specific register and language pair which is less explored in this research line, anticipating validation and comparison with future endeavors.

Another limitation related to the corpus involves the inadequacy of metadata. In terms of the sub-corpus of EFL writing, while the writers are assumed to be native Chinese speakers who write in English as a foreign language, more precise linguistic profiles are not

available. Factors such as individual variability in the environment where they acquire English and their levels of English proficiency add further complexity to the categorization of EFL. These unaccounted variables could potentially introduce bias into the study's findings. Similar challenges appeared for the construction of the translation sub-corpus. For some editorials incorporated, the translators are often collectively referred to as the board of the news agency, leaving it unclear who exactly the translators are. This obscurity extends to the directionality of the translation (i.e., whether translators are translating from their native language to their second language, or the reverse direction), a factor which may significantly influence the linguistic features of the translated text.

In terms of methodology, the limitations are rooted primarily in the inherent challenges associated with multivariate analysis, where the choice of features can heavily impact the results. Though the current study conscientiously draws on previous literature to select 69 linguistic features, the nature of this selection process renders it prone to potential biases. Furthermore, while the selected features are comprehensive, some are categorical in nature, grouping together an array of individual characteristics (e.g., complement clauses to verbs/nouns/adjectives are grouped together). This categorization may obscure the variations in specific feature distribution within a given category. Furthermore, relying on an automatic tool to tag and extract linguistic features is efficient yet introduces another layer of complexity. Inaccuracies in annotation tools can also lead to errors in feature extraction, which may in turn affect the overall conclusions drawn from the data.

The awareness of these limitations not only tempers the conclusions but also paves the way for further research that can build upon the current study and address these challenges and continue to enrich our understanding of the constrained language.

7.3.2 Future directions

The present study serves as an exploratory and descriptive examination of two constrained language varieties, placing emphasis on the implications of the shared and distinctive constraints involved in TE and EFL. These constraints encompass complex multifaceted

elements including the dynamic interplay of bilingual activation, mediated text production and register. However, this initial investigation offers more of a speculative glimpse into these effects, thereby signaling the necessity for more focused and exhaustive explanatory research on the mechanisms underlying such language phenomenon.

For example, to thoroughly investigate the influence of interference from source/first language—a potentially critical aspect of bilingual activation—a parallel approach would be required. As noted by Evert and Neumann (2017: 3), "it's methodologically impossible to determine differences between translated and non-translated texts without comparing the realization of a feature in the matching source text". As such, future research could strategically employ a parallel methodology, acquiring data such as the source texts of translations or original comparable productions in the source/first language. This would facilitate a deeper exploration of the typical features within specific registers of the two language varieties under scrutiny, and thereby ascertain whether the characteristics identified in the output are indeed transferred from the L1 or source language.

Additionally, the complex intertwining of multiple constraints that shape constrained language use calls for an intricate, multifactorial, and multimethodological approach. By leveraging advanced models and statistical tools that accommodate numerous factors simultaneously, it is possible to unravel the interrelated influences that underpin specific phenomena. Here, studies focused on constrained language use may find inspiration in research related to contact language and variationist linguistics. These fields traditionally explore linguistic alternations across diverse language varieties, utilizing sophisticated methodological tools such as clustering techniques and logistic regression analysis (Edwards & Laporte, 2015; Szmrecsanyi & Kortmann, 2011; Gries & Deshors, 2015). In pursuing this course, it becomes feasible to precisely identify the contextual factors, or "constraints", that condition the selection of a particular linguistic variant. The proposed future directions not only build upon the foundational work of this study but also point towards a robust, multifaceted exploration towards a more enriched and comprehensive insight into constrained language use.

Appendices

Appendix 1 Selected linguistic features and their normalisation baselines

No.	Code	Normalisation unit	No.	Code	Normalisation unit
(A)	**General Text Properties**		(H)	**Prepositions**	
1	AWL	Words	37	IN	Words
2	TTR	Words (by default first 400)	(I)	**Adjectives**	
3	LDE	Words	38	JJPR	Finite verbs
(B)	**Verb Semantics**		(J)	**Pronouns**	
4	ACT	Finite verbs	39	PIT	Finite verbs
5	ASPECT	Finite verbs	40	QUPR	Finite verbs
6	CAUSE	Finite verbs	41	PP1	Finite verbs
7	COMM	Finite verbs	42	PP2	Finite verbs
8	EXIST	Finite verbs	43	PP3	Finite verbs
9	MENTAL	Finite verbs	(K)	**Adverbials**	
10	OCCUR	Finite verbs	44	PLACE	Finite verbs
11	DOAUX	Finite verbs	45	TIME	Finite verbs
(C)	**Verb Features**		46	ADVMOD	Words
12	CONT	Finite verbs	(L)	**Negation**	
13	PEAS	Finite verbs	47	XX0	Finite verbs
14	PROG	Finite verbs	(M)	**Noun Semantics**	
15	RP	Finite verbs	48	NNP	Nouns
16	VBD	Finite verbs	49	NOMZ	Nouns
17	VBG	Finite verbs	50	NCOMP	Nouns
18	VBN	Finite verbs	51	NN	Words
19	VPRT	Finite verbs	(N)	**Noun Modifications**	
20	PASS	Finite verbs	52	AMOD	Nouns
(D)	**Modal Verbs**		53	POSS	Nouns
21	MDCA	Finite verbs	54	PPMOD	Nouns
22	MDCO	Finite verbs	55	APPOS	Nouns
23	MDMM	Finite verbs	56	NUM	Nouns
24	MDNE	Finite verbs	57	RC	Nouns
25	MDWO	Finite verbs	58	NFRC	Nouns
26	MDWS	Finite verbs	(O)	**Lexis**	
(E)	**Stative Forms**		59	COMPAR	Words
27	BEMA	Finite verbs	60	SUPER	Words

28	EX	Finite verbs	61	AMP	Words
(F)	**Coordinators and Conjuncts**		62	DWNT	Words
29	CONC	Finite verbs	63	EMPH	Words
30	COND	Finite verbs	64	HDG	Words
31	CUZ	Finite verbs	(P)	**Syntax**	
32	ELAB	Finite verbs	65	SPLIT	Finite verbs
33	CC	Finite verbs	66	THATD	Finite verbs
(G)	**Determinatives**		67	CSUBJ	Finite verbs
34	DEMO	Words	68	ADVCL	Finite verbs
35	DT	Nouns	69	CCOMP	Finite verbs
36	QUAN	Nouns			

Appendix 2 Tests of normality for EFL, NE, and TE

Tests of Normality (EFL)						
	Kolmogorov-Smirnov[b]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AWL	0.074	52	.200*	0.975	52	0.340
TTR	0.059	52	.200*	0.989	52	0.908
LDE	0.090	52	.200*	0.966	52	0.138
ACT	0.096	52	.200*	0.987	52	0.842
AMP	0.130	52	0.028	0.929	52	0.004
ASPECT	0.132	52	0.025	0.917	52	0.001
BEMA	0.071	52	.200*	0.984	52	0.715
CAUSE	0.102	52	.200*	0.953	52	0.040
COMM	0.108	52	0.192	0.978	52	0.459
CONC	0.108	52	0.189	0.948	52	0.023
COND	0.141	52	0.012	0.935	52	0.007
CONT	0.150	52	0.005	0.915	52	0.001
CUZ	0.129	52	0.032	0.930	52	0.004
DEMO	0.097	52	.200*	0.968	52	0.171
DOAUX	0.142	52	0.010	0.913	52	0.001
DT	0.075	52	.200*	0.977	52	0.424
DWNT	0.114	52	0.091	0.962	52	0.100
ELAB	0.145	52	0.008	0.936	52	0.008
EMPH	0.088	52	.200*	0.986	52	0.794
EX	0.084	52	.200*	0.944	52	0.016
EXIST	0.095	52	.200*	0.955	52	0.046
HDG	0.138	52	0.015	0.936	52	0.008
IN	0.078	52	.200*	0.985	52	0.743
JJPR	0.069	52	.200*	0.978	52	0.457
MDCA	0.083	52	.200*	0.960	52	0.082
MDCO	0.128	52	0.034	0.943	52	0.014

MDMM	0.106	52	.200*	0.961	52	0.083
MDNE	0.213	52	0.000	0.690	52	0.000
MDWO	0.146	52	0.008	0.924	52	0.003
MDWS	0.172	52	0.001	0.813	52	0.000
MENTAL	0.104	52	.200*	0.973	52	0.292
NCOMP	0.097	52	.200*	0.984	52	0.706
NN	0.080	52	.200*	0.976	52	0.360
OCCUR	0.197	52	0.000	0.779	52	0.000
PEAS	0.176	52	0.000	0.912	52	0.001
PIT	0.186	52	0.000	0.722	52	0.000
PLACE	0.101	52	.200*	0.951	52	0.032
PP2	0.193	52	0.000	0.813	52	0.000
PROG	0.129	52	0.031	0.963	52	0.110
QUAN	0.076	52	.200*	0.967	52	0.151
QUPR	0.108	52	0.191	0.933	52	0.006
RP	0.116	52	0.080	0.954	52	0.044
SPLIT	0.132	52	0.024	0.960	52	0.076
THATD	0.110	52	0.162	0.971	52	0.227
TIME	0.061	52	.200*	0.981	52	0.559
VBD	0.072	52	.200*	0.973	52	0.276
VBG	0.097	52	.200*	0.937	52	0.009
VBN	0.068	52	.200*	0.990	52	0.930
VPRT	0.077	52	.200*	0.979	52	0.489
XX0	0.102	52	.200*	0.970	52	0.222
COMPAR	0.068	52	.200*	0.980	52	0.530
NNP	0.066	52	.200*	0.990	52	0.936
NOMZ	0.189	52	0.000	0.850	52	0.000
PASS	0.120	52	0.058	0.915	52	0.001
PP1	0.179	52	0.000	0.916	52	0.001
PP3	0.114	52	0.087	0.969	52	0.183
SUPER	0.118	52	0.068	0.885	52	0.000
AMOD	0.159	52	0.002	0.904	52	0.000

POSS	0.108	52	0.182	0.981	52	0.587
PPMOD	0.072	52	.200*	0.976	52	0.390
APPOS	0.075	52	.200*	0.975	52	0.336
NUM	0.076	52	.200*	0.985	52	0.744
NFRC	0.105	52	.200*	0.965	52	0.130
RC	0.077	52	.200*	0.974	52	0.312
ADVCL	0.091	52	.200*	0.966	52	0.144
CCOMP	0.070	52	.200*	0.978	52	0.456
CSUBJ	0.163	52	0.002	0.863	52	0.000
ADVCL	0.080	52	.200*	0.976	52	0.359
CC	0.103	52	.200*	0.955	52	0.046
*. This is a lower bound of the true significance.						
b. Lilliefors Significance Correction						

Tests of Normality (NE)						
	Kolmogorov-Smirnov[b]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AWL	0.074	65	.200*	0.984	65	0.566
TTR	0.103	65	0.085	0.971	65	0.130
LDE	0.061	65	.200*	0.966	65	0.073
ACT	0.129	65	0.009	0.984	65	0.572
AMP	0.091	65	.200*	0.953	65	0.015
ASPECT	0.132	65	0.006	0.879	65	0.000
BEMA	0.090	65	.200*	0.980	65	0.389
CAUSE	0.128	65	0.010	0.949	65	0.010
COMM	0.131	65	0.007	0.963	65	0.048
CONC	0.099	65	0.190	0.949	65	0.009
COND	0.139	65	0.003	0.885	65	0.000
CONT	0.131	65	0.007	0.832	65	0.000
CUZ	0.090	65	.200*	0.971	65	0.130
DEMO	0.097	65	.200*	0.968	65	0.089

DOAUX	0.127	65	0.011	0.904	65	0.000
DT	0.099	65	0.191	0.961	65	0.039
DWNT	0.089	65	.200*	0.974	65	0.180
ELAB	0.158	65	0.000	0.835	65	0.000
EMPH	0.065	65	.200*	0.991	65	0.922
EX	0.150	65	0.001	0.942	65	0.004
EXIST	0.099	65	0.190	0.971	65	0.131
HDG	0.175	65	0.000	0.815	65	0.000
IN	0.049	65	.200*	0.991	65	0.916
JJPR	0.091	65	.200*	0.987	65	0.757
MDCA	0.107	65	0.061	0.901	65	0.000
MDCO	0.064	65	.200*	0.975	65	0.204
MDMM	0.115	65	0.033	0.914	65	0.000
MDNE	0.138	65	0.004	0.919	65	0.000
MDWO	0.167	65	0.000	0.854	65	0.000
MDWS	0.140	65	0.003	0.800	65	0.000
MENTAL	0.119	65	0.024	0.951	65	0.012
NCOMP	0.069	65	.200*	0.981	65	0.411
NN	0.104	65	0.078	0.969	65	0.105
OCCUR	0.105	65	0.071	0.926	65	0.001
PEAS	0.062	65	.200*	0.958	65	0.028
PIT	0.083	65	.200*	0.972	65	0.139
PLACE	0.101	65	0.094	0.944	65	0.005
PP2	0.292	65	0.000	0.503	65	0.000
PROG	0.113	65	0.037	0.964	65	0.054
QUAN	0.119	65	0.022	0.902	65	0.000
QUPR	0.094	65	.200*	0.969	65	0.101
RP	0.099	65	0.184	0.953	65	0.014
SPLIT	0.089	65	.200*	0.972	65	0.144
THATD	0.073	65	.200*	0.959	65	0.032
TIME	0.185	65	0.000	0.760	65	0.000
VBD	0.117	65	0.027	0.964	65	0.055

VBG	0.138	65	0.004	0.936	65	0.002
VBN	0.108	65	0.058	0.935	65	0.002
VPRT	0.097	65	.200*	0.969	65	0.107
XX0	0.094	65	.200*	0.963	65	0.049
COMPAR	0.078	65	.200*	0.963	65	0.047
NNP	0.078	65	.200*	0.986	65	0.702
NOMZ	0.058	65	.200*	0.973	65	0.170
PASS	0.076	65	.200*	0.976	65	0.232
PP1	0.125	65	0.014	0.882	65	0.000
PP3	0.075	65	.200*	0.968	65	0.091
SUPER	0.127	65	0.011	0.940	65	0.003
AMOD	0.115	65	0.032	0.953	65	0.015
POSS	0.085	65	.200*	0.945	65	0.006
PPMOD	0.151	65	0.001	0.929	65	0.001
APPOS	0.133	65	0.006	0.810	65	0.000
NUM	0.163	65	0.000	0.787	65	0.000
NFRC	0.086	65	.200*	0.975	65	0.214
RC	0.067	65	.200*	0.988	65	0.803
ADVCL	0.084	65	.200*	0.978	65	0.294
CCOMP	0.077	65	.200*	0.988	65	0.778
CSUBJ	0.147	65	0.001	0.911	65	0.000
ADVCL	0.094	65	.200*	0.966	65	0.067
CC	0.098	65	.200*	0.972	65	0.153
*. This is a lower bound of the true significance.						
b. Lilliefors Significance Correction						

Tests of Normality (TE)						
	Kolmogorov-Smirnov[b]			Shapiro-Wilk		
	Statistic	df	Sig.	Statistic	df	Sig.
AWL	0.161	51	0.002	0.881	51	0.000
TTR	0.105	51	.200*	0.976	51	0.402

LDE	0.175	51	0.000	0.854	51	0.000
ACT	0.082	51	.200*	0.968	51	0.186
AMP	0.136	51	0.019	0.940	51	0.013
ASPECT	0.107	51	.200*	0.955	51	0.052
BEMA	0.073	51	.200*	0.981	51	0.592
CAUSE	0.150	51	0.006	0.949	51	0.028
COMM	0.137	51	0.018	0.944	51	0.018
CONC	0.088	51	.200*	0.961	51	0.089
COND	0.077	51	.200*	0.952	51	0.038
CONT	0.127	51	0.039	0.912	51	0.001
CUZ	0.121	51	0.061	0.947	51	0.025
DEMO	0.088	51	.200*	0.980	51	0.538
DOAUX	0.139	51	0.016	0.955	51	0.053
DT	0.081	51	.200*	0.987	51	0.840
DWNT	0.108	51	0.196	0.979	51	0.491
ELAB	0.169	51	0.001	0.859	51	0.000
EMPH	0.102	51	.200*	0.972	51	0.270
EX	0.076	51	.200*	0.980	51	0.546
EXIST	0.162	51	0.002	0.904	51	0.001
HDG	0.160	51	0.002	0.945	51	0.020
IN	0.078	51	.200*	0.974	51	0.311
JJPR	0.098	51	.200*	0.963	51	0.115
MDCA	0.104	51	.200*	0.942	51	0.015
MDCO	0.129	51	0.034	0.919	51	0.002
MDMM	0.199	51	0.000	0.845	51	0.000
MDNE	0.161	51	0.002	0.915	51	0.001
MDWO	0.191	51	0.000	0.868	51	0.000
MDWS	0.162	51	0.002	0.895	51	0.000
MENTAL	0.097	51	.200*	0.974	51	0.323
NCOMP	0.121	51	0.060	0.963	51	0.112
NN	0.148	51	0.007	0.836	51	0.000
OCCUR	0.073	51	.200*	0.989	51	0.914

PEAS	0.193	51	0.000	0.897	51	0.000
PIT	0.086	51	.200*	0.973	51	0.293
PLACE	0.106	51	.200*	0.947	51	0.025
PP2	0.292	51	0.000	0.657	51	0.000
PROG	0.111	51	0.164	0.956	51	0.058
QUAN	0.177	51	0.000	0.767	51	0.000
QUPR	0.128	51	0.035	0.919	51	0.002
RP	0.113	51	0.100	0.964	51	0.122
SPLIT	0.096	51	.200*	0.976	51	0.396
THATD	0.150	51	0.006	0.933	51	0.006
TIME	0.080	51	.200*	0.962	51	0.104
VBD	0.146	51	0.008	0.904	51	0.001
VBG	0.095	51	.200*	0.970	51	0.230
VBN	0.133	51	0.024	0.956	51	0.057
VPRT	0.151	51	0.005	0.896	51	0.000
XX0	0.069	51	.200*	0.980	51	0.530
COMPAR	0.130	51	0.031	0.831	51	0.000
NNP	0.105	51	.200*	0.962	51	0.103
NOMZ	0.089	51	.200*	0.940	51	0.013
PASS	0.088	51	.200*	0.978	51	0.471
PP1	0.266	51	0.000	0.719	51	0.000
PP3	0.170	51	0.001	0.833	51	0.000
SUPER	0.066	51	.200*	0.979	51	0.497
AMOD	0.117	51	0.076	0.975	51	0.352
POSS	0.159	51	0.003	0.779	51	0.000
PPMOD	0.063	51	.200*	0.978	51	0.453
APPOS	0.112	51	0.153	0.907	51	0.001
NUM	0.105	51	.200*	0.963	51	0.114
NFRC	0.066	51	.200*	0.986	51	0.786
RC	0.131	51	0.029	0.948	51	0.026
ADVCL	0.091	51	.200*	0.971	51	0.246
CCOMP	0.077	51	.200*	0.976	51	0.402

CSUBJ	0.123	51	0.052	0.928	51	0.004
ADVCL	0.118	51	0.073	0.954	51	0.046
CC	0.090	51	.200*	0.982	51	0.634
*. This is a lower bound of the true significance.						
b. Lilliefors Significance Correction						

Appendix 3 Total variance explained based on Principal Axis Factoring

Factor	Initial Eigenvalues			Extraction Sums of Squared Loadings		
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %
1	9.342	13.539	13.539	9.022	13.075	13.075
2	8.019	11.621	25.161	7.718	11.185	24.260
3	4.226	6.125	31.285	3.850	5.579	29.839
4	2.579	3.738	35.023	2.226	3.227	33.066
5	2.504	3.630	38.653	2.073	3.004	36.070
6	2.103	3.048	41.701	1.662	2.408	38.478
7	2.022	2.931	44.632	1.554	2.252	40.731
8	1.874	2.716	47.347	1.473	2.135	42.866
9	1.734	2.514	49.861	1.313	1.903	44.769
10	1.606	2.327	52.188	1.204	1.745	46.514
11	1.552	2.249	54.438	1.112	1.612	48.126
12	1.501	2.175	56.613	1.079	1.563	49.689
13	1.448	2.099	58.712	1.009	1.463	51.152
14	1.343	1.947	60.659	0.883	1.280	52.431
15	1.289	1.867	62.526	0.811	1.176	53.607
16	1.267	1.837	64.363	0.792	1.148	54.755
17	1.203	1.744	66.107	0.749	1.085	55.841
18	1.130	1.638	67.745	0.700	1.014	56.855
19	1.074	1.556	69.301	0.670	0.970	57.825
20	1.022	1.481	70.782	0.591	0.856	58.681
21	0.981	1.422	72.204			
22	0.950	1.377	73.581			
23	0.935	1.354	74.936			
24	0.904	1.310	76.246			
25	0.856	1.241	77.487			
26	0.817	1.184	78.671			
27	0.794	1.150	79.822			
28	0.774	1.121	80.943			

29	0.735	1.065	82.008			
30	0.714	1.035	83.043			
31	0.701	1.016	84.059			
32	0.643	0.932	84.991			
33	0.593	0.860	85.851			
34	0.569	0.825	86.677			
35	0.536	0.777	87.453			
36	0.521	0.756	88.209			
37	0.508	0.737	88.946			
38	0.503	0.729	89.675			
39	0.479	0.695	90.370			
40	0.452	0.655	91.024			
41	0.443	0.642	91.666			
42	0.410	0.595	92.261			
43	0.397	0.575	92.836			
44	0.383	0.555	93.391			
45	0.348	0.505	93.896			
46	0.338	0.489	94.385			
47	0.330	0.478	94.863			
48	0.302	0.437	95.300			
49	0.296	0.429	95.729			
50	0.286	0.414	96.143			
51	0.271	0.393	96.535			
52	0.263	0.381	96.916			
53	0.238	0.345	97.261			
54	0.220	0.319	97.580			
55	0.201	0.292	97.871			
56	0.188	0.273	98.144			
57	0.181	0.262	98.406			
58	0.179	0.259	98.665			
59	0.161	0.233	98.899			
60	0.148	0.214	99.112			
61	0.128	0.186	99.299			
62	0.117	0.170	99.468			

63	0.102	0.147	99.616			
64	0.080	0.116	99.732			
65	0.060	0.087	99.819			
66	0.058	0.084	99.903			
67	0.038	0.055	99.958			
68	0.027	0.039	99.996			
69	0.003	0.004	100.000			

Appendix 4 Descriptive dimension statistics for EFL, NE, and TE for two sub-registers

Editorials									
	Dimension 1			Dimension 2			Dimension 3		
	EFL	NE	TE	EFL	NE	TE	EFL	NE	TE
Median	-0.56	0.39	-0.69	1.27	0.90	1.10	1.88	0.28	-0.45
MIN	-2.44	-1.32	-1.26	-1.11	-0.29	0.13	-0.27	-1.32	-0.95
MAX	1.31	3.60	-0.21	2.81	2.20	1.69	3.02	2.02	0.38
Q3	0.10	1.13	-0.54	1.89	1.24	1.39	2.22	0.79	-0.17
Q1	-0.92	-0.03	-0.93	0.62	0.44	0.86	1.37	-0.24	-0.55
IQR	1.02	1.16	0.39	1.27	0.80	0.53	0.84	1.03	0.37
	Dimension 4			Dimension 5			Dimension 6		
	EFL	NE	TE	EFL	NE	TE	EFL	NE	TE
Median	-1.85	-0.41	-0.67	0.41	-0.66	-0.54	-1.03	-0.21	-0.49
MIN	-2.50	-1.61	-1.56	-1.96	-1.50	-1.92	-2.25	-2.68	-1.50
MAX	-0.11	3.43	0.07	2.37	4.95	0.75	1.78	1.62	0.55
Q3	-1.18	0.25	-0.31	0.90	-0.02	-0.11	0.28	0.48	-0.25
Q1	-2.25	-0.71	-0.92	-0.38	-0.98	-0.89	-1.61	-1.07	-0.83
IQR	1.07	0.97	0.61	1.28	0.96	0.78	1.89	1.56	0.58
Reports									
	Dimension 1			Dimension 2			Dimension 3		
	EFL	NE	TE	EFL	NE	TE	EFL	NE	TE
Median	0.11	-0.31	0.33	-0.31	-0.75	-0.75	-0.59	-0.17	0.42
MIN	2.67	2.01	3.49	1.21	1.43	0.58	0.83	2.52	2.50
MAX	4.21	4.69	4.72	2.68	3.42	2.89	3.00	4.91	4.59
Q3	0.62	0.30	0.88	0.01	-0.14	-0.21	-0.27	0.37	1.09
Q1	-0.54	-0.84	-0.23	-0.80	-1.17	-1.34	-0.79	-0.69	-0.49
IQR	1.16	1.15	1.10	0.81	1.03	1.13	0.52	1.06	1.58
	Dimension 4			Dimension 5			Dimension 6		
	EFL	NE	TE	EFL	NE	TE	EFL	NE	TE
Median	0.39	-0.15	0.48	-0.04	-0.07	0.47	0.25	0.22	0.46
MIN	2.69	5.81	2.09	1.69	5.44	2.21	3.15	2.31	2.64
MAX	4.83	6.83	3.11	3.40	7.55	3.48	5.48	3.99	4.53
Q3	1.20	0.19	1.06	0.52	0.46	0.79	0.92	0.77	1.28
Q1	0.09	-0.39	0.08	-0.56	-0.35	-0.22	-0.54	-0.56	-0.60
IQR	1.11	0.58	0.98	1.08	0.80	1.01	1.46	1.33	1.89

Appendix 5 Results of regression model comparison

Dimension 1

Analysis of Variance Table

Model 1: D1 ~ Variety * Register

Model 2: D1 ~ Variety + Register

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	162	141.04				
2	164	171.55	-2	-30.504	17.518	1.296e-07***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dimension 2

Analysis of Variance Table

Model 1: D2 ~ Variety * Register

Model 2: D2 ~ Variety + Register

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	162	82.708				
2	164	84.239	-2	-1.5304	1.4987	0.2265

Dimension 3

Analysis of Variance Table

Model 1: D3 ~ Variety * Register

Model 2: D3 ~ Variety + Register

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	162	126.89				
2	164	173.29	-2	-46.397	29.617	1.091e-11***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dimension 4

Analysis of Variance Table

Model 1: D4 ~ Variety * Register

Model 2: D4 ~ Variety + Register

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	162	133.04				
2	164	153.91	-2	-20.873	12.709	7.467e-06***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dimension 5

Analysis of Variance Table

Model 1: D5 ~ Variety * Register

Model 2: D5 ~ Variety + Register

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	162	195.14				
2	164	202.50	-2	-7.3532	3.0522	0.04998*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Dimension 6

Analysis of Variance Table

Model 1: D6 ~ Variety * Register

Model 2: D6 ~ Variety + Register

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	162	188.74				
2	164	190.97	-2	-2.2308	0.9574	0.3861

Appendix 6 Results of regression models

Dimension 1

Call:

lm(formula = D1 ~ Variety * Register, data = data)

Residuals:

Min	1Q	Median	3Q	Max
-2.37487	-0.52243	-0.06843	0.47818	3.11659

Coefficients:

	Estimate	Std.Error	T value	Pr(>|t|)
(Intercept)	-0.4016	0.2588	-1.552	0.122678
VarietyNE	0.8827	0.3191	2.766	0.006325**
VarietyTE	-0.3276	0.3264	-1.004	0.317036
Registerreport	0.5354	0.2988	1.792	0.075034
VarietyNE:Registerreport	-1.3172	0.3820	-3.449	0.000718***
VarietyTE:Registerreport	0.7469	0.3986	1.874	0.062757

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.9331 on 162 degrees of freedom

Multiple R-squared: 0.1885, Adjusted R-squared: 0.1635

F-statistic: 7.528 on 5 and 162 DF, p-value: 2.222e-06

Dimension 2

Call:

lm(formula = D2 ~ Variety * Register, data = data)

Residuals:

Min	1Q	Median	3Q	Max
-2.17961	-0.48665	0.03684	0.40308	2.00672

Coefficients:

	Estimate	Std.Error	T value	Pr(>|t|)
(Intercept)	1.067524	0.198174	5.387	2.49e-07***
VarietyNE	-0.150671	0.244325	-0.617	0.538
VarietyTE	0.006279	0.249959	0.025	0.980
Registerreport	-1.423364	0.228831	-6.220	4.07e-09***
VarietyNE:Registerreport	-0.066521	0.292488	-0.227	0.820
VarietyTE:Registerreport	-0.465047	0.305246	-1.524	0.130

Signif. codes: 0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

Residual standard error: 0.7145 on 162 degrees of freedom

Multiple R-squared: 0.5436, Adjusted R-squared: 0.5295
F-statistic: 38.59 on 5 and 162 DF, p-value: < 2.2e-16

Dimension 3

Call:
lm(formula = D3 ~ Variety * Register, data = data)
Residuals:

Min	1Q	Median	3Q	Max
-2.37727	-0.48534	-0.00292	0.52959	2.72882

Coefficients:

	Estimate	Std.Error	T value	Pr(>|t|)
(Intercept)	1.6840	0.2455	6.861	1.38e-10***
VarietyNE	-1.3544	0.3026	-4.475	1.43e-05***
VarietyTE	-2.0610	0.3096	-6.657	4.12e-10***
Registerreport	-2.2454	0.2834	-7.922	3.54e-13***
VarietyNE:Registerreport	1.7096	0.3623	4.719	5.09e-06***
VarietyTE:Registerreport	2.9083	0.3781	7.692	1.33e-12***

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 0.885 on 162 degrees of freedom
Multiple R-squared: 0.3176, Adjusted R-squared: 0.2966
F-statistic: 15.08 on 5 and 162 DF, p-value: 3.747e-12

Dimension 4

Call:
lm(formula = D4 ~ Variety * Register, data = data)
Residuals:

Min	1Q	Median	3Q	Max
-2.6842	-0.4919	-0.1327	0.4511	5.7000

Coefficients:

	Estimate	Std.Error	T value	Pr(>|t|)
(Intercept)	-1.6206	0.2513	-6.448	1.25e-09***
VarietyNE	1.4372	0.3099	4.638	7.21e-06***
VarietyTE	0.9602	0.3170	3.029	0.00286**
Registerreport	2.1608	0.2902	7.446	5.40e-12***
VarietyNE:Registerreport	-1.8627	0.3710	-5.022	1.34e-06***
VarietyTE:Registerreport	-0.9994	0.3871	-2.582	0.01072*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.9062 on 162 degrees of freedom
Multiple R-squared: 0.324, Adjusted R-squared: 0.3031
F-statistic: 15.53 on 5 and 162 DF, p-value: 1.806e-12

Dimension 5

Call:
lm(formula = D5 ~ Variety * Register, data = data)
Residuals:

Min	1Q	Median	3Q	Max
-2.2764	-0.5315	-0.1177	0.4297	5.2786

Coefficients:

	Estimate	Std.Error	T value	Pr(>|t|)
(Intercept)	0.2313	0.3044	0.760	0.4484
VarietyNE	-0.4839	0.3753	-1.289	0.1991
VarietyTE	-0.7127	0.3839	-1.856	0.0652
Registerreport	-0.3084	0.3515	-0.877	0.3815
VarietyNE:Registerreport	0.7189	0.4493	1.600	0.1115
VarietyTE:Registerreport	1.1549	0.4689	2.463	0.0148*

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.098 on 162 degrees of freedom
Multiple R-squared: 0.06012, Adjusted R-squared: 0.03112
F-statistic: 2.073 on 5 and 162 DF, p-value: 0.07143

Dimension 6

Call:
lm(formula = D6 ~ Variety * Register, data = data)
Residuals:

Min	1Q	Median	3Q	Max
-2.56124	-0.70224	0.02567	0.71425	2.91837

Coefficients:

	Estimate	Std.Error	T value	Pr(>|t|)
(Intercept)	-0.68142	0.29937	-2.276	0.02414*
VarietyNE	0.45148	0.36909	1.223	0.22302
VarietyTE	0.20864	0.37760	0.553	0.58133
Registerreport	0.90856	0.34568	2.628	0.00941**
VarietyNE:Registerreport	-0.53491	0.44184	-1.211	0.22780
VarietyTE:Registerreport	-0.07712	0.46112	-0.167	0.86738

Signif. codes: 0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
Residual standard error: 1.079 on 162 degrees of freedom
Multiple R-squared: 0.09078, Adjusted R-squared: 0.06271
F-statistic: 3.235 on 5 and 162 DF, p-value: 0.008227

Appendix 7 Kruskal-Wallis tests and post-hoc tests among EFL, NE, and TE

Kruskal-Wallis tests among EFL, NE, and TE (editorials)

Feature	Test Statistic	Degree Of Freedom	Asymptotic Sig. (2-sided test)
AWL	37.494	2	0.000
TTR	12.519	2	0.002
LDE	30.082	2	0.000
ACT	3.288	2	0.193
AMP	1.103	2	0.576
ASPECT	1.104	2	0.576
BEMA	10.192	2	0.006
CAUSE	14.978	2	0.001
COMM	21.446	2	0.000
CONC	0.381	2	0.827
COND	5.687	2	0.058
CONT	10.636	2	0.005
CUZ	22.377	2	0.000
DEMO	31.734	2	0.000
DOAUX	12.337	2	0.002
DT	5.291	2	0.071
DWNT	6.238	2	0.044
ELAB	11.049	2	0.004
EMPH	18.770	2	0.000
EX	5.818	2	0.055
EXIST	15.514	2	0.000
HDG	8.665	2	0.013
IN	18.068	2	0.000
JJPR	1.352	2	0.509
MDCA	7.028	2	0.030
MDCO	15.712	2	0.000
MDMM	7.534	2	0.023
MDNE	4.046	2	0.132
MDWO	13.068	2	0.001
MDWS	12.255	2	0.002
MENTAL	1.900	2	0.387
NCOMP	7.077	2	0.029
NN	25.762	2	0.000
OCCUR	13.674	2	0.001

PEAS	20.613	2	0.000
PIT	1.388	2	0.500
PLACE	3.949	2	0.139
PP2	13.170	2	0.001
PROG	10.050	2	0.007
QUAN	18.816	2	0.000
QUPR	12.079	2	0.002
RP	5.096	2	0.078
SPLIT	21.887	2	0.000
THATD	1.208	2	0.547
TIME	16.926	2	0.000
VBD	5.319	2	0.070
VBG	19.347	2	0.000
VBN	30.634	2	0.000
VPRT	10.353	2	0.006
XX0	6.855	2	0.032
COMPAR	3.520	2	0.172
NNP	20.318	2	0.000
NOMZ	10.551	2	0.005
PASS	11.233	2	0.004
PP1	22.137	2	0.000
PP3	11.509	2	0.003
SUPER	2.411	2	0.300
AMOD	22.933	2	0.000
POSS	4.238	2	0.120
PPMOD	20.053	2	0.000
APPOS	23.672	2	0.000
NUM	3.314	2	0.191
NFRC	12.726	2	0.002
RC	34.190	2	0.000
ADVCL	17.993	2	0.000
CCOMP	10.569	2	0.005
CSUBJ	6.407	2	0.041
ADVCL	26.334	2	0.000
CC	17.710	2	0.000

Pairwise comparisons among EFL, NE, and TE (editorials)

Feature		Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.
AWL	NE-ed-TE-ed	-22.711	5.105	-4.449	0.000	0.000
	NE-ed-EFL-ed	33.658	5.972	5.636	0.000	0.000
	TE-ed-EFL-ed	10.948	6.109	1.792	0.073	0.219
TTR	TE-ed-EFL-ed	5.369	6.107	0.879	0.379	1.000
	TE-ed-NE-ed	17.663	5.104	3.461	0.001	0.002
	EFL-ed-NE-ed	-12.294	5.970	-2.059	0.039	0.118
LDE	NE-ed-TE-ed	-18.495	5.105	-3.623	0.000	0.001
	NE-ed-EFL-ed	31.194	5.972	5.224	0.000	0.000
	TE-ed-EFL-ed	12.699	6.109	2.079	0.038	0.113
ACT	TE-ed-NE-ed	7.202	5.105	1.411	0.158	0.475
	TE-ed-EFL-ed	10.066	6.109	1.648	0.099	0.298
	NE-ed-EFL-ed	2.865	5.972	0.480	0.631	1.000
AMP	EFL-ed-NE-ed	-1.823	5.971	-0.305	0.760	1.000
	EFL-ed-TE-ed	-5.900	6.109	-0.966	0.334	1.000
	NE-ed-TE-ed	-4.077	5.105	-0.799	0.424	1.000
ASPECT	NE-ed-EFL-ed	2.462	5.972	0.412	0.680	1.000
	NE-ed-TE-ed	-5.364	5.105	-1.051	0.293	0.880
	EFL-ed-TE-ed	-2.902	6.109	-0.475	0.635	1.000
BEMA	EFL-ed-TE-ed	-16.502	6.109	-2.701	0.007	0.021
	EFL-ed-NE-ed	-18.155	5.972	-3.040	0.002	0.007
	TE-ed-NE-ed	1.654	5.105	0.324	0.746	1.000
CAUSE	NE-ed-TE-ed	-5.998	5.105	-1.175	0.240	0.720
	NE-ed-EFL-ed	22.988	5.972	3.850	0.000	0.000
	TE-ed-EFL-ed	16.990	6.109	2.781	0.005	0.016
COMM	TE-ed-NE-ed	20.882	5.105	4.090	0.000	0.000
	TE-ed-EFL-ed	22.990	6.109	3.763	0.000	0.001
	NE-ed-EFL-ed	2.108	5.971	0.353	0.724	1.000
CONC	TE-ed-NE-ed	1.261	5.104	0.247	0.805	1.000
	TE-ed-EFL-ed	3.764	6.108	0.616	0.538	1.000
	NE-ed-EFL-ed	2.503	5.971	0.419	0.675	1.000
COND	EFL-ed-NE-ed	-11.165	5.971	-1.870	0.062	0.185
	EFL-ed-TE-ed	-14.271	6.109	-2.336	0.019	0.058
	NE-ed-TE-ed	-3.106	5.105	-0.609	0.543	1.000
CONT	EFL-ed-NE-ed	-17.502	5.968	-2.932	0.003	0.010
	EFL-ed-TE-ed	-18.189	6.106	-2.979	0.003	0.009
	NE-ed-TE-ed	-0.687	5.102	-0.135	0.893	1.000
CUZ	TE-ed-NE-ed	0.697	5.105	0.137	0.891	1.000

	TE-ed-EFL-ed	26.247	6.109	4.296	0.000	0.000
	NE-ed-EFL-ed	25.549	5.971	4.279	0.000	0.000
DEMO	EFL-ed-TE-ed	-13.941	6.109	-2.282	0.023	0.068
	EFL-ed-NE-ed	-32.317	5.972	-5.412	0.000	0.000
	TE-ed-NE-ed	18.376	5.105	3.599	0.000	0.001
DOAUX	EFL-ed-TE-ed	-15.673	6.109	-2.565	0.010	0.031
	EFL-ed-NE-ed	-20.823	5.972	-3.487	0.000	0.001
	TE-ed-NE-ed	5.150	5.105	1.009	0.313	0.939
DT	TE-ed-NE-ed	9.622	5.105	1.885	0.059	0.178
	TE-ed-EFL-ed	12.336	6.109	2.019	0.043	0.130
	NE-ed-EFL-ed	2.714	5.972	0.454	0.650	1.000
DWNT	TE-ed-EFL-ed	2.399	6.109	0.393	0.695	1.000
	TE-ed-NE-ed	12.171	5.105	2.384	0.017	0.051
	EFL-ed-NE-ed	-9.772	5.972	-1.636	0.102	0.305
ELAB	TE-ed-NE-ed	7.018	5.101	1.376	0.169	0.507
	TE-ed-EFL-ed	20.280	6.104	3.322	0.001	0.003
	NE-ed-EFL-ed	13.262	5.967	2.223	0.026	0.079
EMPH	EFL-ed-TE-ed	-18.881	6.109	-3.090	0.002	0.006
	EFL-ed-NE-ed	-25.754	5.972	-4.313	0.000	0.000
	TE-ed-NE-ed	6.873	5.105	1.346	0.178	0.535
EX	EFL-ed-TE-ed	-9.439	6.109	-1.545	0.122	0.367
	EFL-ed-NE-ed	-14.402	5.971	-2.412	0.016	0.048
	TE-ed-NE-ed	4.963	5.105	0.972	0.331	0.993
EXIST	TE-ed-NE-ed	0.065	5.105	0.013	0.990	1.000
	TE-ed-EFL-ed	21.589	6.109	3.534	0.000	0.001
	NE-ed-EFL-ed	21.525	5.971	3.605	0.000	0.001
HDG	EFL-ed-TE-ed	-6.512	6.095	-1.068	0.285	0.856
	EFL-ed-NE-ed	-16.608	5.958	-2.788	0.005	0.016
	TE-ed-NE-ed	10.095	5.093	1.982	0.047	0.142
IN	TE-ed-NE-ed	14.266	5.105	2.794	0.005	0.016
	TE-ed-EFL-ed	25.117	6.109	4.111	0.000	0.000
	NE-ed-EFL-ed	10.851	5.972	1.817	0.069	0.208
JJPR	NE-ed-TE-ed	-4.036	5.105	-0.791	0.429	1.000
	NE-ed-EFL-ed	6.554	5.972	1.097	0.272	0.817
	TE-ed-EFL-ed	2.517	6.109	0.412	0.680	1.000
MDCA	EFL-ed-NE-ed	-1.538	5.971	-0.258	0.797	1.000
	EFL-ed-TE-ed	-13.357	6.109	-2.186	0.029	0.086
	NE-ed-TE-ed	-11.818	5.105	-2.315	0.021	0.062
MDCO	TE-ed-NE-ed	18.265	5.089	3.589	0.000	0.001
	TE-ed-EFL-ed	18.897	6.090	3.103	0.002	0.006
	NE-ed-EFL-ed	0.632	5.953	0.106	0.915	1.000
MDMM	TE-ed-EFL-ed	1.014	6.102	0.166	0.868	1.000

	TE-ed-NE-ed	12.891	5.099	2.528	0.011	0.034
	EFL-ed-NE-ed	-11.877	5.964	-1.991	0.046	0.139
MDNE	NE-ed-TE-ed	-5.663	5.105	-1.109	0.267	0.802
	NE-ed-EFL-ed	11.832	5.971	1.982	0.048	0.143
	TE-ed-EFL-ed	6.170	6.109	1.010	0.313	0.938
MDWO	TE-ed-EFL-ed	8.163	6.109	1.336	0.181	0.544
	TE-ed-NE-ed	18.392	5.105	3.603	0.000	0.001
	EFL-ed-NE-ed	-10.229	5.971	-1.713	0.087	0.260
MDWS	EFL-ed-NE-ed	-5.386	5.972	-0.902	0.367	1.000
	EFL-ed-TE-ed	-19.369	6.109	-3.170	0.002	0.005
	NE-ed-TE-ed	-13.983	5.105	-2.739	0.006	0.018
MENTAL	EFL-ed-TE-ed	-1.187	6.109	-0.194	0.846	1.000
	EFL-ed-NE-ed	-6.986	5.972	-1.170	0.242	0.726
	TE-ed-NE-ed	5.799	5.105	1.136	0.256	0.768
NCOMP	TE-ed-NE-ed	1.364	5.105	0.267	0.789	1.000
	TE-ed-EFL-ed	15.210	6.109	2.490	0.013	0.038
	NE-ed-EFL-ed	13.846	5.972	2.319	0.020	0.061
NN	NE-ed-TE-ed	-23.131	5.105	-4.531	0.000	0.000
	NE-ed-EFL-ed	23.348	5.972	3.910	0.000	0.000
	TE-ed-EFL-ed	0.217	6.109	0.035	0.972	1.000
OCCUR	NE-ed-TE-ed	-9.274	5.105	-1.817	0.069	0.208
	NE-ed-EFL-ed	21.968	5.972	3.679	0.000	0.001
	TE-ed-EFL-ed	12.694	6.109	2.078	0.038	0.113
PEAS	NE-ed-TE-ed	-11.036	5.105	-2.162	0.031	0.092
	NE-ed-EFL-ed	27.015	5.972	4.524	0.000	0.000
	TE-ed-EFL-ed	15.979	6.109	2.615	0.009	0.027
PIT	NE-ed-TE-ed	-5.276	5.105	-1.034	0.301	0.904
	NE-ed-EFL-ed	5.563	5.972	0.932	0.352	1.000
	TE-ed-EFL-ed	0.287	6.109	0.047	0.963	1.000
PLACE	EFL-ed-TE-ed	-9.573	6.109	-1.567	0.117	0.351
	EFL-ed-NE-ed	-11.606	5.972	-1.944	0.052	0.156
	TE-ed-NE-ed	2.033	5.105	0.398	0.691	1.000
PP2	TE-ed-EFL-ed	3.776	5.526	0.683	0.494	1.000
	TE-ed-NE-ed	16.145	4.618	3.496	0.000	0.001
	EFL-ed-NE-ed	-12.369	5.401	-2.290	0.022	0.066
PROG	EFL-ed-NE-ed	-10.562	5.971	-1.769	0.077	0.231
	EFL-ed-TE-ed	-19.257	6.109	-3.152	0.002	0.005
	NE-ed-TE-ed	-8.695	5.105	-1.703	0.089	0.266
QUAN	EFL-ed-TE-ed	-4.278	6.109	-0.700	0.484	1.000
	EFL-ed-NE-ed	-22.266	5.972	-3.729	0.000	0.001
	TE-ed-NE-ed	17.988	5.105	3.524	0.000	0.001
QUPR	TE-ed-EFL-ed	1.038	6.109	0.170	0.865	1.000

	TE-ed-NE-ed	16.260	5.105	3.185	0.001	0.004
	EFL-ed-NE-ed	-15.222	5.971	-2.549	0.011	0.032
RP	TE-ed-NE-ed	10.342	5.105	2.026	0.043	0.128
	TE-ed-EFL-ed	10.951	6.109	1.793	0.073	0.219
	NE-ed-EFL-ed	0.609	5.972	0.102	0.919	1.000
SPLIT	NE-ed-TE-ed	-21.366	5.105	-4.185	0.000	0.000
	NE-ed-EFL-ed	21.442	5.972	3.591	0.000	0.001
	TE-ed-EFL-ed	0.075	6.109	0.012	0.990	1.000
THATD	TE-ed-NE-ed	1.930	5.104	0.378	0.705	1.000
	TE-ed-EFL-ed	6.673	6.108	1.092	0.275	0.824
	NE-ed-EFL-ed	4.743	5.970	0.794	0.427	1.000
TIME	EFL-ed-TE-ed	-9.503	6.109	-1.556	0.120	0.359
	EFL-ed-NE-ed	-23.391	5.971	-3.917	0.000	0.000
	TE-ed-NE-ed	13.887	5.105	2.720	0.007	0.020
VBD	TE-ed-EFL-ed	5.414	6.109	0.886	0.375	1.000
	TE-ed-NE-ed	11.748	5.105	2.301	0.021	0.064
	EFL-ed-NE-ed	-6.334	5.972	-1.061	0.289	0.867
VBG	NE-ed-TE-ed	-9.521	5.105	-1.865	0.062	0.187
	NE-ed-EFL-ed	26.257	5.971	4.397	0.000	0.000
	TE-ed-EFL-ed	16.736	6.109	2.739	0.006	0.018
VBN	TE-ed-NE-ed	12.146	5.105	2.379	0.017	0.052
	TE-ed-EFL-ed	33.809	6.109	5.534	0.000	0.000
	NE-ed-EFL-ed	21.663	5.972	3.628	0.000	0.001
VPRT	NE-ed-EFL-ed	6.342	5.972	1.062	0.288	0.865
	NE-ed-TE-ed	-16.380	5.105	-3.208	0.001	0.004
	EFL-ed-TE-ed	-10.038	6.109	-1.643	0.100	0.301
XX0	EFL-ed-NE-ed	-12.800	5.972	-2.143	0.032	0.096
	EFL-ed-TE-ed	-15.455	6.109	-2.530	0.011	0.034
	NE-ed-TE-ed	-2.655	5.105	-0.520	0.603	1.000
COMPAR	EFL-ed-TE-ed	-5.463	6.109	-0.894	0.371	1.000
	EFL-ed-NE-ed	-10.977	5.972	-1.838	0.066	0.198
	TE-ed-NE-ed	5.514	5.105	1.080	0.280	0.840
NNP	EFL-ed-NE-ed	-4.868	5.972	-0.815	0.415	1.000
	EFL-ed-TE-ed	-23.944	6.109	-3.919	0.000	0.000
	NE-ed-TE-ed	-19.076	5.105	-3.737	0.000	0.001
NOMZ	NE-ed-TE-ed	-6.922	5.105	-1.356	0.175	0.525
	NE-ed-EFL-ed	19.394	5.972	3.248	0.001	0.003
	TE-ed-EFL-ed	12.472	6.109	2.041	0.041	0.124
PASS	TE-ed-NE-ed	4.402	5.105	0.862	0.389	1.000
	TE-ed-EFL-ed	20.066	6.109	3.285	0.001	0.003
	NE-ed-EFL-ed	15.665	5.972	2.623	0.009	0.026
PP1	TE-ed-EFL-ed	1.217	6.109	0.199	0.842	1.000

	TE-ed-NE-ed	21.949	5.105	4.299	0.000	0.000
	EFL-ed-NE-ed	-20.732	5.972	-3.472	0.001	0.002
PP3	TE-ed-EFL-ed	0.972	6.109	0.159	0.874	1.000
	TE-ed-NE-ed	15.858	5.105	3.106	0.002	0.006
	EFL-ed-NE-ed	-14.886	5.972	-2.493	0.013	0.038
SUPER	EFL-ed-NE-ed	-8.394	5.972	-1.406	0.160	0.480
	EFL-ed-TE-ed	-8.608	6.109	-1.409	0.159	0.476
	NE-ed-TE-ed	-0.215	5.105	-0.042	0.966	1.000
AMOD	NE-ed-TE-ed	-14.038	5.105	-2.750	0.006	0.018
	NE-ed-EFL-ed	28.028	5.972	4.693	0.000	0.000
	TE-ed-EFL-ed	13.990	6.109	2.290	0.022	0.066
POSS	EFL-ed-NE-ed	-0.662	5.972	-0.111	0.912	1.000
	EFL-ed-TE-ed	-10.052	6.109	-1.645	0.100	0.300
	NE-ed-TE-ed	-9.391	5.105	-1.839	0.066	0.198
PPMOD	TE-ed-NE-ed	13.976	5.105	2.738	0.006	0.019
	TE-ed-EFL-ed	26.829	6.109	4.391	0.000	0.000
	NE-ed-EFL-ed	12.852	5.972	2.152	0.031	0.094
APPOS	TE-ed-NE-ed	22.429	5.105	4.393	0.000	0.000
	TE-ed-EFL-ed	23.371	6.109	3.825	0.000	0.000
	NE-ed-EFL-ed	0.942	5.972	0.158	0.875	1.000
NUM	TE-ed-NE-ed	6.209	5.105	1.216	0.224	0.672
	TE-ed-EFL-ed	10.717	6.109	1.754	0.079	0.238
	NE-ed-EFL-ed	4.508	5.972	0.755	0.450	1.000
NFRC	TE-ed-NE-ed	16.528	5.105	3.238	0.001	0.004
	TE-ed-EFL-ed	16.991	6.109	2.781	0.005	0.016
	NE-ed-EFL-ed	0.463	5.972	0.078	0.938	1.000
RC	TE-ed-EFL-ed	1.734	6.109	0.284	0.777	1.000
	TE-ed-NE-ed	27.353	5.105	5.358	0.000	0.000
	EFL-ed-NE-ed	-25.618	5.972	-4.290	0.000	0.000
ADVCL	NE-ed-TE-ed	-3.916	5.105	-0.767	0.443	1.000
	NE-ed-EFL-ed	24.665	5.972	4.130	0.000	0.000
	TE-ed-EFL-ed	20.748	6.109	3.396	0.001	0.002
CCOMP	TE-ed-EFL-ed	13.353	6.109	2.186	0.029	0.087
	TE-ed-NE-ed	15.965	5.105	3.127	0.002	0.005
	EFL-ed-NE-ed	-2.612	5.972	-0.437	0.662	1.000
CSUBJ	NE-ed-EFL-ed	10.777	5.941	1.814	0.070	0.209
	NE-ed-TE-ed	-11.905	5.079	-2.344	0.019	0.057
	EFL-ed-TE-ed	-1.128	6.078	-0.186	0.853	1.000
ADVCL	EFL-ed-TE-ed	-14.222	6.109	-2.328	0.020	0.060
	EFL-ed-NE-ed	-29.854	5.972	-4.999	0.000	0.000
	TE-ed-NE-ed	15.632	5.105	3.062	0.002	0.007
CC	NE-ed-TE-ed	-16.002	5.105	-3.134	0.002	0.005

	NE-ed-EFL-ed	22.858	5.972	3.828	0.000	0.000
	TE-ed-EFL-ed	6.857	6.109	1.122	0.262	0.785

Kruskal-Wallis tests among EFL, NE and TE (reports)

Feature	Test Statistic	Degree Of Freedom	Asymptotic Sig. (2-sided test)
AWL	7.196	2	0.027
TTR	2.125	2	0.346
LDE	27.759	2	0.000
ACT	11.474	2	0.003
AMP	7.066	2	0.029
ASPECT	4.096	2	0.129
BEMA	1.334	2	0.513
CAUSE	1.172	2	0.556
COMM	28.649	2	0.000
CONC	6.509	2	0.039
COND	5.205	2	0.074
CONT	15.515	2	0.000
CUZ	0.696	2	0.706
DEMO	9.448	2	0.009
DOAUX	4.861	2	0.088
DT	13.129	2	0.001
DWNT	13.433	2	0.001
ELAB	10.986	2	0.004
EMPH	8.388	2	0.015
EX	17.040	2	0.000
EXIST	2.706	2	0.258
HDG	0.149	2	0.928
IN	0.457	2	0.796
JJPR	6.213	2	0.045
MDCA	6.624	2	0.036
MDCO	2.778	2	0.249
MDMM	0.467	2	0.792
MDNE	0.650	2	0.722
MDWO	9.500	2	0.009
MDWS	29.656	2	0.000
MENTAL	1.856	2	0.395
NCOMP	6.347	2	0.042
NN	25.470	2	0.000
OCCUR	1.230	2	0.541
PEAS	5.007	2	0.082
PIT	7.140	2	0.028
PLACE	14.527	2	0.001
PP2	0.368	2	0.832
PROG	20.113	2	0.000

QUAN	6.406	2	0.041
QUPR	14.342	2	0.001
RP	23.543	2	0.000
SPLIT	4.683	2	0.096
THATD	24.961	2	0.000
TIME	3.184	2	0.204
VBD	2.744	2	0.254
VBG	8.895	2	0.012
VBN	3.129	2	0.209
VPRT	1.472	2	0.479
XX0	16.114	2	0.000
COMPAR	4.083	2	0.130
NNP	32.064	2	0.000
NOMZ	2.844	2	0.241
PASS	13.726	2	0.001
PP1	0.557	2	0.757
PP3	11.353	2	0.003
SUPER	6.733	2	0.035
AMOD	16.936	2	0.000
POSS	2.245	2	0.325
PPMOD	2.099	2	0.350
APPOS	32.874	2	0.000
NUM	36.092	2	0.000
NFRC	1.075	2	0.584
RC	14.368	2	0.001
ADVCL	20.309	2	0.000
CCOMP	3.048	2	0.218
CSUBJ	6.363	2	0.042
ADVCL	43.822	2	0.000
CC	11.970	2	0.003

Pairwise comparisons among EFL, NE, and TE (reports)

Feature		Test Statistic	Std. Error	Std. Test Statistic	Sig.	Adj. Sig.
AWL	NE-R-EFL-R	14.773	7.048	2.096	0.036	0.108
	NE-R-TE-R	-18.678	7.639	-2.445	0.014	0.043
	EFL-R-TE-R	-3.905	7.680	-0.508	0.611	1.000
	TE-R-NE-R	9.449	7.638	1.237	0.216	0.648
TTR	TE-R-EFL-R	10.314	7.679	1.343	0.179	0.538
	NE-R-EFL-R	0.865	7.047	0.123	0.902	1.000
	NE-R-TE-R	-5.997	7.639	-0.785	0.432	1.000
LDE	NE-R-EFL-R	35.210	7.048	4.996	0.000	0.000
	TE-R-EFL-R	29.212	7.680	3.804	0.000	0.000
	NE-R-EFL-R	0.590	7.048	0.084	0.933	1.000
ACT	NE-R-TE-R	-23.320	7.639	-3.053	0.002	0.007
	EFL-R-TE-R	-22.729	7.680	-2.960	0.003	0.009
	TE-R-EFL-R	15.182	7.676	1.978	0.048	0.144
AMP	TE-R-NE-R	19.785	7.635	2.591	0.010	0.029
	EFL-R-NE-R	-4.603	7.045	-0.653	0.513	1.000
	NE-R-EFL-R	1.844	7.048	0.262	0.794	1.000
ASPECT	NE-R-TE-R	-14.558	7.639	-1.906	0.057	0.170
	EFL-R-TE-R	-12.714	7.680	-1.655	0.098	0.293
	TE-R-EFL-R	2.264	7.680	0.295	0.768	1.000
BEMA	TE-R-NE-R	8.267	7.639	1.082	0.279	0.837
	EFL-R-NE-R	-6.003	7.048	-0.852	0.394	1.000
	NE-R-TE-R	-6.526	7.639	-0.854	0.393	1.000
CAUSE	NE-R-EFL-R	6.917	7.048	0.981	0.326	0.979
	TE-R-EFL-R	0.391	7.680	0.051	0.959	1.000
	TE-R-NE-R	25.690	7.639	3.363	0.001	0.002
COMM	TE-R-EFL-R	41.020	7.680	5.341	0.000	0.000
	NE-R-EFL-R	15.330	7.048	2.175	0.030	0.089
	NE-R-EFL-R	0.009	7.048	0.001	0.999	1.000
CONC	NE-R-TE-R	-17.352	7.638	-2.272	0.023	0.069
	EFL-R-TE-R	-17.344	7.679	-2.259	0.024	0.072
	TE-R-EFL-R	2.248	7.677	0.293	0.770	1.000
COND	TE-R-NE-R	15.405	7.636	2.017	0.044	0.131
	EFL-R-NE-R	-13.157	7.045	-1.867	0.062	0.186
	EFL-R-NE-R	-20.858	7.048	-2.959	0.003	0.009
CONT	EFL-R-TE-R	-28.239	7.680	-3.677	0.000	0.001
	NE-R-TE-R	-7.381	7.639	-0.966	0.334	1.000
	EFL-R-NE-R	-0.659	7.045	-0.093	0.926	1.000

CUZ	EFL-R-TE-R	-5.967	7.676	-0.777	0.437	1.000
	NE-R-TE-R	-5.308	7.635	-0.695	0.487	1.000
	EFL-R-TE-R	-17.942	7.680	-2.336	0.019	0.058
DEMO	EFL-R-NE-R	-20.119	7.048	-2.854	0.004	0.013
	TE-R-NE-R	2.178	7.639	0.285	0.776	1.000
	EFL-R-NE-R	-0.279	7.048	-0.040	0.968	1.000
DOAUX	EFL-R-TE-R	-15.132	7.680	-1.970	0.049	0.146
	NE-R-TE-R	-14.853	7.638	-1.944	0.052	0.156
	EFL-R-TE-R	-18.796	7.680	-2.447	0.014	0.043
DT	EFL-R-NE-R	-24.692	7.048	-3.503	0.000	0.001
	TE-R-NE-R	5.897	7.639	0.772	0.440	1.000
	NE-R-EFL-R	16.340	7.048	2.318	0.020	0.061
DWNT	NE-R-TE-R	-27.371	7.639	-3.583	0.000	0.001
	EFL-R-TE-R	-11.031	7.680	-1.436	0.151	0.453
	NE-R-TE-R	-16.440	7.632	-2.154	0.031	0.094
ELAB	NE-R-EFL-R	22.702	7.042	3.224	0.001	0.004
	TE-R-EFL-R	6.262	7.673	0.816	0.414	1.000
	EFL-R-NE-R	-8.377	7.048	-1.188	0.235	0.704
EMPH	EFL-R-TE-R	-22.202	7.680	-2.891	0.004	0.012
	NE-R-TE-R	-13.825	7.639	-1.810	0.070	0.211
	EFL-R-TE-R	-7.848	7.678	-1.022	0.307	0.920
EX	EFL-R-NE-R	-28.302	7.047	-4.016	0.000	0.000
	TE-R-NE-R	20.454	7.637	2.678	0.007	0.022
	NE-R-EFL-R	0.950	7.048	0.135	0.893	1.000
EXIST	NE-R-TE-R	-11.618	7.639	-1.521	0.128	0.385
	EFL-R-TE-R	-10.668	7.680	-1.389	0.165	0.494
	NE-R-EFL-R	1.492	7.046	0.212	0.832	1.000
HDG	NE-R-TE-R	-2.928	7.637	-0.383	0.701	1.000
	EFL-R-TE-R	-1.437	7.678	-0.187	0.852	1.000
	TE-R-EFL-R	4.567	7.680	0.595	0.552	1.000
IN	TE-R-NE-R	4.625	7.639	0.605	0.545	1.000
	EFL-R-NE-R	-0.057	7.048	-0.008	0.994	1.000
	NE-R-EFL-R	5.347	7.048	0.759	0.448	1.000
JJPR	NE-R-TE-R	-18.785	7.639	-2.459	0.014	0.042
	EFL-R-TE-R	-13.439	7.680	-1.750	0.080	0.240
	NE-R-TE-R	-11.143	7.639	-1.459	0.145	0.434
MDCA	NE-R-EFL-R	17.988	7.048	2.552	0.011	0.032
	TE-R-EFL-R	6.845	7.680	0.891	0.373	1.000
	EFL-R-TE-R	-7.704	7.675	-1.004	0.315	0.946
MDCO	EFL-R-NE-R	-11.584	7.043	-1.645	0.100	0.300
	TE-R-NE-R	3.880	7.633	0.508	0.611	1.000
	NE-R-TE-R	-3.119	7.626	-0.409	0.683	1.000

MDMM	NE-R-EFL-R	4.742	7.037	0.674	0.500	1.000
	TE-R-EFL-R	1.623	7.667	0.212	0.832	1.000
	TE-R-EFL-R	3.243	7.680	0.422	0.673	1.000
MDNE	TE-R-NE-R	6.148	7.638	0.805	0.421	1.000
	EFL-R-NE-R	-2.905	7.048	-0.412	0.680	1.000
	EFL-R-NE-R	-17.931	7.048	-2.544	0.011	0.033
MDWO	EFL-R-TE-R	-20.912	7.680	-2.723	0.006	0.019
	NE-R-TE-R	-2.980	7.639	-0.390	0.696	1.000
	TE-R-NE-R	34.923	7.638	4.572	0.000	0.000
MDWS	TE-R-EFL-R	38.807	7.679	5.053	0.000	0.000
	NE-R-EFL-R	3.884	7.048	0.551	0.582	1.000
	NE-R-EFL-R	0.000	7.048	0.000	1.000	1.000
MENTAL	NE-R-TE-R	-9.263	7.639	-1.213	0.225	0.676
	EFL-R-TE-R	-9.263	7.680	-1.206	0.228	0.683
	TE-R-NE-R	10.082	7.639	1.320	0.187	0.561
NCOMP	TE-R-EFL-R	19.309	7.680	2.514	0.012	0.036
	NE-R-EFL-R	9.228	7.048	1.309	0.190	0.571
	TE-R-NE-R	11.901	7.639	1.558	0.119	0.358
NN	TE-R-EFL-R	37.019	7.680	4.820	0.000	0.000
	NE-R-EFL-R	25.119	7.048	3.564	0.000	0.001
	TE-R-NE-R	0.582	7.639	0.076	0.939	1.000
OCCUR	TE-R-EFL-R	7.281	7.680	0.948	0.343	1.000
	NE-R-EFL-R	6.698	7.048	0.950	0.342	1.000
	TE-R-EFL-R	1.698	7.680	0.221	0.825	1.000
PEAS	TE-R-NE-R	14.871	7.639	1.947	0.052	0.155
	EFL-R-NE-R	-13.173	7.048	-1.869	0.062	0.185
	TE-R-EFL-R	11.489	7.680	1.496	0.135	0.404
PIT	TE-R-NE-R	20.407	7.639	2.671	0.008	0.023
	EFL-R-NE-R	-8.918	7.048	-1.265	0.206	0.617
	EFL-R-NE-R	-4.061	7.048	-0.576	0.565	1.000
PLACE	EFL-R-TE-R	-27.677	7.680	-3.604	0.000	0.001
	NE-R-TE-R	-23.616	7.639	-3.092	0.002	0.006
	EFL-R-TE-R	-1.419	7.656	-0.185	0.853	1.000
PP2	EFL-R-NE-R	-4.198	7.026	-0.598	0.550	1.000
	TE-R-NE-R	2.780	7.615	0.365	0.715	1.000
	TE-R-EFL-R	7.388	7.680	0.962	0.336	1.000
PROG	TE-R-NE-R	31.575	7.639	4.133	0.000	0.000
	EFL-R-NE-R	-24.187	7.048	-3.432	0.001	0.002
	EFL-R-NE-R	-0.421	7.048	-0.060	0.952	1.000
QUAN	EFL-R-TE-R	-17.420	7.680	-2.268	0.023	0.070
	NE-R-TE-R	-16.999	7.639	-2.225	0.026	0.078
	EFL-R-NE-R	-3.339	7.048	-0.474	0.636	1.000

QUPR	EFL-R-TE-R	-27.241	7.680	-3.547	0.000	0.001
	NE-R-TE-R	-23.902	7.639	-3.129	0.002	0.005
	EFL-R-NE-R	-25.231	7.048	-3.580	0.000	0.001
RP	EFL-R-TE-R	-35.050	7.680	-4.564	0.000	0.000
	NE-R-TE-R	-9.818	7.639	-1.285	0.199	0.596
	NE-R-EFL-R	1.990	7.048	0.282	0.778	1.000
SPLIT	NE-R-TE-R	-15.572	7.639	-2.039	0.041	0.124
	EFL-R-TE-R	-13.583	7.680	-1.769	0.077	0.231
	TE-R-NE-R	33.178	7.639	4.343	0.000	0.000
THATD	TE-R-EFL-R	34.725	7.680	4.522	0.000	0.000
	NE-R-EFL-R	1.547	7.048	0.220	0.826	1.000
	TE-R-NE-R	6.147	7.639	0.805	0.421	1.000
TIME	TE-R-EFL-R	13.557	7.680	1.765	0.078	0.233
	NE-R-EFL-R	7.410	7.048	1.051	0.293	0.879
	EFL-R-NE-R	-3.251	7.048	-0.461	0.645	1.000
VBD	EFL-R-TE-R	-12.465	7.680	-1.623	0.105	0.314
	NE-R-TE-R	-9.214	7.639	-1.206	0.228	0.683
	NE-R-EFL-R	9.994	7.048	1.418	0.156	0.469
VBG	NE-R-TE-R	-22.777	7.639	-2.982	0.003	0.009
	EFL-R-TE-R	-12.782	7.680	-1.664	0.096	0.288
	NE-R-TE-R	-10.322	7.639	-1.351	0.177	0.530
VBN	NE-R-EFL-R	11.502	7.048	1.632	0.103	0.308
	TE-R-EFL-R	1.180	7.680	0.154	0.878	1.000
	TE-R-NE-R	2.342	7.639	0.307	0.759	1.000
VPRT	TE-R-EFL-R	8.722	7.680	1.136	0.256	0.768
	NE-R-EFL-R	6.380	7.048	0.905	0.365	1.000
	EFL-R-NE-R	-11.286	7.048	-1.601	0.109	0.328
XX0	EFL-R-TE-R	-30.746	7.680	-4.003	0.000	0.000
	NE-R-TE-R	-19.460	7.639	-2.548	0.011	0.033
	EFL-R-NE-R	-7.032	7.048	-0.998	0.318	0.955
COMPAR	EFL-R-TE-R	-15.511	7.680	-2.020	0.043	0.130
	NE-R-TE-R	-8.479	7.639	-1.110	0.267	0.801
	TE-R-EFL-R	32.032	7.680	4.171	0.000	0.000
NNP	TE-R-NE-R	42.274	7.639	5.534	0.000	0.000
	EFL-R-NE-R	-10.242	7.048	-1.453	0.146	0.439
	TE-R-EFL-R	0.270	7.680	0.035	0.972	1.000
NOMZ	TE-R-NE-R	10.677	7.639	1.398	0.162	0.487
	EFL-R-NE-R	-10.407	7.048	-1.477	0.140	0.419
	TE-R-EFL-R	10.770	7.680	1.402	0.161	0.482
PASS	TE-R-NE-R	27.578	7.639	3.610	0.000	0.001
	EFL-R-NE-R	-16.808	7.048	-2.385	0.017	0.051
	TE-R-EFL-R	3.225	7.680	0.420	0.675	1.000

PP1	TE-R-NE-R	5.701	7.639	0.746	0.455	1.000
	EFL-R-NE-R	-2.476	7.048	-0.351	0.725	1.000
	EFL-R-NE-R	-4.925	7.048	-0.699	0.485	1.000
PP3	EFL-R-TE-R	-24.909	7.680	-3.243	0.001	0.004
	NE-R-TE-R	-19.984	7.639	-2.616	0.009	0.027
	TE-R-NE-R	16.410	7.639	2.148	0.032	0.095
SUPER	TE-R-EFL-R	18.644	7.680	2.428	0.015	0.046
	NE-R-EFL-R	2.233	7.048	0.317	0.751	1.000
	NE-R-EFL-R	18.337	7.048	2.602	0.009	0.028
AMOD	NE-R-TE-R	-30.737	7.639	-4.024	0.000	0.000
	EFL-R-TE-R	-12.401	7.680	-1.615	0.106	0.319
	NE-R-TE-R	-2.045	7.639	-0.268	0.789	1.000
POSS	NE-R-EFL-R	10.110	7.048	1.434	0.151	0.454
	TE-R-EFL-R	8.065	7.680	1.050	0.294	0.881
	NE-R-TE-R	-0.160	7.639	-0.021	0.983	1.000
PPMOD	NE-R-EFL-R	9.158	7.048	1.299	0.194	0.582
	TE-R-EFL-R	8.997	7.680	1.172	0.241	0.724
	TE-R-NE-R	21.235	7.639	2.780	0.005	0.016
APPOS	TE-R-EFL-R	43.775	7.680	5.700	0.000	0.000
	NE-R-EFL-R	22.540	7.048	3.198	0.001	0.004
	NE-R-TE-R	-8.881	7.639	-1.163	0.245	0.735
NUM	NE-R-EFL-R	40.716	7.048	5.777	0.000	0.000
	TE-R-EFL-R	31.835	7.680	4.145	0.000	0.000
	EFL-R-NE-R	-5.413	7.048	-0.768	0.443	1.000
NFRC	EFL-R-TE-R	-7.478	7.680	-0.974	0.330	0.991
	NE-R-TE-R	-2.066	7.639	-0.270	0.787	1.000
	EFL-R-TE-R	-15.603	7.680	-2.032	0.042	0.127
RC	EFL-R-NE-R	-26.626	7.048	-3.778	0.000	0.000
	TE-R-NE-R	11.023	7.639	1.443	0.149	0.447
	NE-R-EFL-R	7.127	7.048	1.011	0.312	0.936
ADVCL	NE-R-TE-R	-33.382	7.639	-4.370	0.000	0.000
	EFL-R-TE-R	-26.255	7.680	-3.419	0.001	0.002
	TE-R-NE-R	3.291	7.639	0.431	0.667	1.000
CCOMP	TE-R-EFL-R	12.523	7.680	1.631	0.103	0.309
	NE-R-EFL-R	9.232	7.048	1.310	0.190	0.571
	EFL-R-NE-R	-1.040	7.014	-0.148	0.882	1.000
CSUBJ	EFL-R-TE-R	-17.568	7.643	-2.299	0.022	0.065
	NE-R-TE-R	-16.528	7.602	-2.174	0.030	0.089
	EFL-R-NE-R	-10.490	7.048	-1.488	0.137	0.410
ADVCL	EFL-R-TE-R	-49.176	7.680	-6.403	0.000	0.000
	NE-R-TE-R	-38.686	7.639	-5.064	0.000	0.000
	NE-R-EFL-R	15.308	7.048	2.172	0.030	0.090

CC	NE-R-TE-R	-25.872	7.639	-3.387	0.001	0.002
	EFL-R-TE-R	-10.563	7.680	-1.375	0.169	0.507
	NE-R-EFL-R	14.773	7.048	2.096	0.036	0.108

References

Alamri, B. (2023). A Multidimensional Comparative Analysis of MENA and International English Research Article Abstracts in Applied Linguistics. *SAGE Open*, *13*(1), 215824402211456. https://doi.org/10.1177/21582440221145669

Albrecht, S. (2021). Current research on the linguistic features of Chinese English. *World Englishes*, *42*(3), 487–506. https://doi.org/10.1111/weng.12572

Ansary, H., & Babaii, E. (2009). A Cross-cultural Analysis of English Newspaper Editorials: A Systemic-Functional View of Text for Contrastive Rhetoric Research. *RELC Journal*, *40*(2), 211–249. https://doi.org/10.1177/0033688209105867

Baker, M. (1993). Corpus Linguistics and Translation Studies—Implications and Applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology* (p. 233). John Benjamins Publishing Company. https://doi.org/10.1075/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.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Benjamins Translation Library* (Vol. 18, p. 175). John Benjamins Publishing Company. https://doi.org/10.1075/btl.18.17bak

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

Becher, V. (2011). *Explicitation and implicitation in translation. A corpus-based study of English-German and German-English translations of business texts*. (Doctoral dissertation). Staats-und Universitätsbibliothek Hamburg Carl von Ossietzky.

Berber Sardinha, T. (2018). Dimensions of variation across Internet registers. *International Journal of Corpus Linguistics*, *23*(2), 125-157.

Bernardini, S., & Ferraresi, A. (2011). Practice, Description and Theory Come Together – Normalization or Interference in Italian Technical Translation? *Meta*, *56*(2), 226–246. https://doi.org/10.7202/1006174ar

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.), *Benjamins Translation Library* (Vol. 48, pp. 51–62). John Benjamins Publishing Company. https://doi.org/10.1075/btl.48.05ber

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

Biber, D. (1988). *Variation across Speech and Writing* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511621024

Biber, D. (2006). *University Language: A corpus-based study of spoken and written registers* (Vol. 23). John Benjamins Publishing Company. https://doi.org/10.1075/scl.23

Biber, D., & Conrad, S. (2009). *Register, genre, and style*. Cambridge University Press.

Biber, D., & Egbert, J. (2016). Register Variation on the Searchable Web: A Multi-Dimensional Analysis. *Journal of English Linguistics*, *44*(2), 95–137. https://doi.org/10.1177/0075424216628955

Biber, D., Conrad, S., & Leech, G. (2002). *Longman student grammar of spoken and written English.* (1st edition). Longman.

Biewer, C. (2011). Modal auxiliaries in second language varieties of English: A learner's perspective. In J. Mukherjee & M. Hundt (Eds.), *Studies in Corpus Linguistics* (Vol. 44, pp. 7–34). John Benjamins Publishing Company. https://doi.org/10.1075/scl.44.02bie

Bisiada, M. (2017). *Universals Of Editing and Translation*. Zenodo. https://doi.org/10.5281/ZENODO.1090972

Blum, S., & Levenston, E. A. (1978). UNIVERSALS OF LEXICAL SIMPLIFICATION. *Language Learning*, *28*(2), 399–415. https://doi.org/10.1111/j.1467-1770.1978.tb00143.x

Blum-Kulka, S. (1986). Shifts of Cohesion and Coherence in Translation Shoshana Blum-Kulka, Hebrew University of Jerusalem. *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies*, *272*, 17.

Bolton, K., & Botha, W. (2015). Researching English in contemporary China: Researching English in contemporary China. *World Englishes*, *34*(2), 169–174. https://doi.org/10.1111/weng.12131

Bolton, K., Botha, W., & Zhang, W. (2020). English in China. In K. Bolton, W. Botha, & A. Kirkpatrick (Eds.), *The Handbook of Asian Englishes* (1st ed., pp. 501–528). Wiley. https://doi.org/10.1002/9781118791882.ch21

Cappelle, B., & Loock, R. (2013). Is there interference of usage constraints?: A frequency study of existential *there is* and its French equivalent *il y a* in translated vs. Non-translated texts. *Target. International Journal of Translation Studies*, *25*(2), 252–275. https://doi.org/10.1075/target.25.2.05cap

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

Chan, A. Y. W. (2004). Syntactic Transfer: Evidence from the Interlanguage of Hong Kong Chinese ESL Learners. *The Modern Language Journal*, *88*(1), 56–74. https://doi.org/10.1111/j.0026-7902.2004.00218.x

Cheng, L. L. -S., & Sybesma, R. (2014). The Syntactic Structure of Noun Phrases. In C. -T. J. Huang, Y. -H. A. Li, & A. Simpson (Eds.), *The Handbook of Chinese Linguistics* (1st ed., pp. 248–274). Wiley. https://doi.org/10.1002/9781118584552.ch10

Chesterman, A. (2004). Hypotheses about translation universals. In G. Hansen, K. Malmkjaer, & D. Gile (Eds.), *Benjamins Translation Library* (Vol. 50, pp. 1–13). John Benjamins Publishing Company. https://doi.org/10.1075/btl.50.02che

Conrad, S., & Biber, D. (Eds.). (2001). *Variation in English: Multi-dimensional studies*. Longman.

Conway, J. M., & Huffcutt, A. I. (2003). A Review and Evaluation of Exploratory Factor Analysis Practices in Organizational Research. *Organizational Research Methods*, *6*(2), 147–168. https://doi.org/10.1177/1094428103251541

Costa, A., & Sebastián-Gallés, N. (2014). How does the bilingual experience sculpt the brain? *Nature Reviews Neuroscience*, *15*(5), 336–345. https://doi.org/10.1038/nrn3709

Croft, W. (2000). *Explaining language change: An evolutionary approach*. Longman.

Dayter, D. (2018). Describing lexical patterns in simultaneously interpreted discourse in a parallel aligned corpus of Russian-English interpreting (SIREN). In *Forum* (Vol. 16, No. 2, pp. 241-264). Amsterdam/Philadelphia: John Benjamins Publishing Company.

De Groot, A. M. B., & Christoffels, I. K. (2006). Language control in bilinguals: Monolingual tasks and simultaneous interpreting. *Bilingualism: Language and Cognition*, *9*(2), 189–201. https://doi.org/10.1017/S1366728906002537

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

Delaere, I. (2015). *Do translations walk the line?: visually exploring translated and non-translated texts in search of norm conformity* (Doctoral dissertation), Ghent University.

Delaere, I., Sutter, G. D., & 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

Dimitrova, B. E. (2005). *Expertise and explicitation in the translation process* (Vol. 64). John Benjamins Publishing.

DiStefano, C., Zhu, M., & Mîndrilã, D. (2019) Understanding and Using Factor Scores: Considerations for the Applied Researcher, *Practical Assessment, Research, and Evaluation*: Vol. 14, Article 20. https://doi.org/10.7275/da8t-4g52

Duff, A. (1981). *The Third Language: recurrent problems of translation into English*. Oxford: Pergamon Press.

Edwards, A., & Laporte, S. (2015). Outer and expanding circle Englishes: The competing roles of norm orientation and proficiency levels. *English World-Wide. A Journal of Varieties of English*, *36*(2), 135–169. https://doi.org/10.1075/eww.36.2.01edw

Egbert, J., & Staples, S. (2019). Doing Multi-Dimensional Analysis in SPSS, SAS, and R. In T.B. Sardinha & M.V. Pinto, *Multi-Dimensional Analysis: Research Methods and Current*

Issues (pp. 125–144). London: Bloomsbury Academic. Retrieved August 7, 2023, from http://dx.doi.org/10.5040/9781350023857.0015

Evert, S., & Neumann, S. (2017). 2 The impact of translation direction on characteristics of translated texts. A multivariate analysis for English and German. In G. D. Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical Translation Studies* (pp. 47–80). De Gruyter. https://doi.org/10.1515/9783110459586-003

Ferraresi, A., Bernardini, S., Petrović, M. M., & Lefer, M.-A. (2019). Simplified or not Simplified? The Different Guises of Mediated English at the European Parliament. *Meta*, *63*(3), 717–738. https://doi.org/10.7202/1060170ar

Field, A. (2009). *Discovering statistics using SPSS.* London:Sage.

Filipović, L., & Hawkins, J. A. (2013). Multiple factors in second language acquisition: The CASP model. *Linguistics*, *51*(1). https://doi.org/10.1515/ling-2013-0005

Filppula, M., Klemola, J., & Paulasto, H. (Eds.). (2009). *Vernacular universals and language contacts: Evidence from varieties of English and beyond*. Routledge.

François, T., & Lefer, M.-A. (2022). Revisiting simplification in corpus-based translation studies: Insights from readability research. *Meta*, *67*(1), 50–70. https://doi.org/10.7202/1092190ar

Frawley, W. (1984/2000). Prolegomenon to a theory of translation. In L. Venuti (ed.) *The Translation Studies Reader*. London & New York: Routledge, 250-263.

Gaspari, F., & Bernardini, S. (2010). Comparing Non-native and Translated Language. In R. Xiao (ed.), *Using Corpora in Contrastive and Translation Studies*, 215–34, Newcastle: Cambridge Scholars Publishing.

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course (3rd ed.)*. New York, NY: Routledge.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In Wollin, L. and Lindquist, H. (Eds.), *Translation studies in Scandinavia: Poceedings from the Scandinavian Symposium on Translation Theory (SSOTT) II*, 88–95. Lund: CWK Gleerup.

Gellerstam, M. (2005). Chapter 13. Fingerprints in Translation. In G. Anderman & M. Rogers (Eds.), *In and Out of English* (pp. 201–213). Multilingual Matters. https://doi.org/10.21832/9781853597893-016

Gilquin, G. (2011). *Corpus linguistics to bridge the gap between World Englishes and Learner Englishes.*12th International Symposium on Social Communication (Santiago de Cuba, du 17/01/2011 au 21/01/2011). In: L. Ruiz Miyares & M.R. Álvarez Silva, *Comunicación Social en el siglo XXI, Vol. II*, Centro de Lingüística Aplicada: Santiago de Cuba, p.638-642. http://hdl.handle.net/2078.1/112509

Gilquin, G. (forthcoming). Lexical use in spoken New Englishes and Learner Englishes: The effects of shared and distinct communicative constraints. In Van Rooy. B, & Kruger. H. (Eds). *Constraints on Language Variation and Change in Complex Multilingual Contact Settings*. (Contact Language Library). John Benjamins.

Gilquin, G., & Granger, S. (2011). From EFL to ESL: Evidence from the *International Corpus of Learner English*. In J. Mukherjee & M. Hundt (Eds.), *Studies in Corpus Linguistics* (Vol. 44, pp. 55–78). John Benjamins Publishing Company. https://doi.org/10.1075/scl.44.04gra

Götz, S., & Schilk, M. (2011). Formulaic sequences in spoken ENL, ESL and EFL: Focus on British English, Indian English and learner English of advanced German learners. In J. Mukherjee & M. Hundt (Eds.), *Studies in Corpus Linguistics* (Vol. 44, pp. 79–100). John Benjamins Publishing Company. https://doi.org/10.1075/scl.44.05sch

Goulart, L., & Wood, M. (2021). Methodological synthesis of research using multi-dimensional analysis. *Journal of Research Design and Statistics in Linguistics and Communication Science*, *6*(2). https://doi.org/10.1558/jrds.18454

Grabowski, L. (2013). Interfacing corpus linguistics and computational stylistics: Translation universals in translational literary Polish. *International Journal of Corpus Linguistics*, *18*(2), 254–280. https://doi.org/10.1075/ijcl.18.2.04gra

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra

Granger, S. (2018). Tracking the third code: a crosslinguistic corpus-driven approach to metadiscursive markers. In A. Čermáková and M. Mahlberg (Eds.), *The Corpus Linguistics Discourse: In Honour of Wolfgang Teubert*, 185–204. Amsterdam: John Benjamins.

Granger, S., & Lefer, M.-A. (2022). *Extending the Scope of Corpus-Based Translation Studies*. Bloomsbury Academic. https://doi.org/10.5040/9781350143289

Gries, S. Th., & Deshors, S. C. (2015). EFL and/vs. ESL?: A multi-level regression modeling perspective on bridging the paradigm gap. *International Journal of Learner Corpus Research*, *1*(1), 130–159. https://doi.org/10.1075/ijlcr.1.1.05gri

Grieve, J. (2007). Quantitative Authorship Attribution: An Evaluation of Techniques. *Literary and Linguistic Computing*, *22*(3), 251–270. https://doi.org/10.1093/llc/fqm020

Halverson, S. L. (2017). 1 Gravitational pull in translation. Testing a revised model. In G. D. Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical Translation Studies* (pp. 9–46). De Gruyter. https://doi.org/10.1515/9783110459586-002

Halverson, S. L., & Kotze, H. (2021). Sociocognitive Constructs in Translation and Interpreting Studies (TIS). In S. L. Halverson & Á. M. García, *Contesting Epistemologies in Cognitive Translation and Interpreting Studies* (1st ed., pp. 51–79). Routledge. https://doi.org/10.4324/9781003125792-5

Hansen-Schirra, S., & Steiner, E. (2012). 14 Towards a typology of translation properties. In S. Hansen-Schirra, S. Neumann, & E. Steiner, *Cross-Linguistic Corpora for the Study of Translations* (pp. 255–280). DE GRUYTER. https://doi.org/10.1515/9783110260328.255

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.

Heine, B., & Kuteva, T. (2005). *Language contact and grammatical change*. New York: Cambridge University Press.

Hershberger, S. L. (2005). Factor Score Estimation. In B. S. Everitt & D. C. Howell (Eds.), *Encyclopedia of Statistics in Behavioral Science* (p. bsa726). John Wiley & Sons, Ltd. https://doi.org/10.1002/0470013192.bsa726

Hinkel, E. (1992). L2 Tense and Time Reference. *TESOL Quarterly*, *26*(3), 557. https://doi.org/10.2307/3587178

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

Hu, H., & Kübler, S. (2021). Investigating translated Chinese and its variants using machine learning. *Natural Language Engineering*, *27*(3), 339–372. https://doi.org/10.1017/S1351324920000182

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.

Huang, C. T. J. (1998). *Logical relations in Chinese and the theory of grammar*. New York: Garland Publishing.

Huang, J. (1988). Yingdang kending 'xiyi hanhua'xianxiang de jijimian (The positive role of Sinicism in the English translated version). *Zhongguo fanyi [Chinese Translators Journal]*, 1, 39–47.

Huang, Y., & Ren, W. (2020). A novel multidimensional analysis of writing styles of editorials from China Daily and The New York Times. *Lingua*, *235*, 102781. https://doi.org/10.1016/j.lingua.2019.102781

Ilisei, I., & Inkpen, D. (2011). Translationese Traits in Romanian Newspapers: A Machine Learning Approach. *Int. J. Comput. Linguistics Appl.*, *2*(1-2), 319-332.

Ilisei, I., Inkpen, D., Corpas Pastor, G., & Mitkov, R. (2010). Identification of Translationese: A Machine Learning Approach. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (Vol. 6008, pp. 503–511). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-12116-6_43

Ivaska, I., & Bernardini, S. (2020). Constrained language use in Finnish: A corpus-driven approach. *Nordic Journal of Linguistics*, *43*(1), 33–57. https://doi.org/10.1017/S0332586520000013

Jantunen, J. H. (2001). Synonymity and Lexical Simplification in Translations: A Corpus-Based Approach. *Across Languages and Cultures*, *2*(1), 97–112. https://doi.org/10.1556/Acr.2.2001.1.7

Jarvis, S. (2002). Short Texts, Best-Fitting Curves and New Measures of Lexical Diversity. *Language Testing*. 19(1):57-84.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic Influence in Language and Cognition* (0 ed.). Routledge. https://doi.org/10.4324/9780203935927

Jiang, W. (2017). A Study on Modified-Modifying Sequence in the Compositions by Chinese Advanced Users of English. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching Chinese*

English: The State of the Art (Vol. 22, pp. 93–107). Springer International Publishing. https://doi.org/10.1007/978-3-319-53110-6_7

Kachru, B. B. (Ed.). (1982). *The Other Tongue: English across Cultures*. Urbana: University of Illinois Press.

Kaiser, H. F. (1960). The Application of Electronic Computers to Factor Analysis. *Educational and Psychological Measurement*, *20*(1), 141–151. https://doi.org/10.1177/001316446002000116

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika*, *39*(1), 31–36. https://doi.org/10.1007/BF02291575

Kajzer-Wietrzny, M. (2022). An intermodal approach to cohesion in constrained and unconstrained language. *Target. International Journal of Translation Studies*. *Target*, *34*(1), 130-162.https://doi.org/10.1075/target.19186.kaj

Kajzer-Wietrzny, M., & Ivaska, I. (2020). A Multivariate Approach to Lexical Diversity in Constrained Language. *Across Languages and Cultures*, *21*(2), 169–194. https://doi.org/10.1556/084.2020.00011

Kajzer-Wietrzny, M., Whyatt, B., & Stachowiak, K. (2016). Simplification in inter- and intralingual translation – combining corpus linguistics, key logging and eye-tracking. *Poznan Studies in Contemporary Linguistics*, *52*(2). https://doi.org/10.1515/psicl-2016-0009

Kirkpatrick, A., & Xu, Z. (2002). Chinese pragmatic norms and "China English." *World Englishes*, *21*(2), 269–279. https://doi.org/10.1111/1467-971X.00247

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

Kolehmainen, L., Meriläinen, L. & H. Riionheimo. (2014). Interlingual Reduction: Evidence from Language Contacts, Translation and Second Language Acquisition. In H. Paulasto, L. Meriläinen, H. Riionheimo & M. Kok (Eds.). *Language Contacts at the Crossroads of Disciplines*, 3–32, Cambridge: Cambridge Scholars

Koppel, M., & Ordan, N. (2011). Translationese and Its Dialects. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1318–1326, Portland, Oregon, USA. Association for Computational Linguistics.

Kortmann, B., & Szmrecsanyi, B. (2009). World Englishes between simplification and complexification. In T. Hoffmann & L. Siebers (Eds.), *Varieties of English Around the World* (Vol. G40, pp. 263–286). John Benjamins Publishing Company. https://doi.org/10.1075/veaw.g40.17kor

Kotze, H. (2020). Translation, contact linguistics and cognition (pp. 113-132). In Alves, F., & Jakobsen, A. L. (Eds.). (2020). *The Routledge handbook of translation and cognition*. Abingdon: Routledge.

Kotze, H. (2022). Translation as constrained communication: Principles, concepts and methods. In S. Granger & M.-A. Lefer (Eds.), *Extending the Scope of Corpus-Based Translation Studies* (1st ed., pp. 67–97). Bloomsbury Academic; Bloomsbury Collections.

Kotze, H., & Van Rooy, B. (Eds.). (forthcoming). *Constraints on Language Variation and Change in Complex Multilingual Contact Settings*. (Contact Language Library). John Benjamins.

Kranich, S. (2014). Translations as a Locus of Language Contact. In J. House (Ed.), *Translation: A Multidisciplinary Approach* (pp. 96–115). Palgrave Macmillan UK. https://doi.org/10.1057/9781137025487_6

Kroll, J. F., & Tokowicz, N. (2005). Models of Bilingual Representation and Processing: Looking Back and to the Future. In J. F. Kroll & A. M. B. de Groot (Eds.), *Handbook of bilingualism: Psycholinguistic approaches* (pp. 531–553). Oxford University Press.

Kroll, J. F., Bobb, S. C., & Hoshino, N. (2014). Two Languages in Mind: Bilingualism as a Tool to Investigate Language, Cognition, and the Brain. *Current Directions in Psychological Science*, *23*(3), 159–163. https://doi.org/10.1177/0963721414528511

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. (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. (2018). Register variation in written contact varieties of English: A multidimensional analysis. *English World-Wide. A Journal of Varieties of English*, *39*(2), 214–242. https://doi.org/10.1075/eww.00011.kru

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. A Journal of Varieties of English*, *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

Kunilovskaya, M., & Corpas Pastor, G. (2021). Translationese and Register Variation in English-To-Russian Professional Translation. In V. X. Wang, L. Lim, & D. Li (Eds.), *New Perspectives on Corpus Translation Studies* (pp. 133–180). Springer Singapore. https://doi.org/10.1007/978-981-16-4918-9_6

Kunilovskaya, M., & Lapshinova-Koltunski, E. (2019). Translationese Features as Indicators of Quality in English-Russian Human Translation. *Proceedings of the Second Workshop Human-Informed Translation and Interpreting Technology Associated with RANLP 2019*, 47–56. https://doi.org/10.26615/issn.2683-0078.2019_006

Kurokawa, D., Cyril, G., & Pierre, I. (2009). Automatic detection of translated text and its impact on machine translation. *Machine Translation Summit XII*. http://www.mt-archive.info/MTS-2009-K

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

Laporte, S. (2012). Mind the gap!: Bridge between World Englishes and Learner Englishes in the making. *English Text Construction*, *5*(2), 264–291. https://doi.org/10.1075/etc.5.2.05lap

Lapshinova-Koltunski, E. (2022). Detecting normalization and shining-through in novice and professional translations. In S. Granger & M. -A. Lefer (Ed.). *Extending the Scope of Corpus-Based Translation Studies* (pp. 182–206). London: Bloomsbury Academic. Retrieved September 12, 2023, http://dx.doi.org/10.5040/9781350143289.0015

Lapshinova-Koltunski, E., Pollkläsener, C., & Przybyl, H. (2022). Exploring Explicitation and Implicitation in Parallel Interpreting and Translation Corpora. *Prague Bulletin of Mathematical Linguistics*, *119*(1), 5–22. https://doi.org/10.14712/00326585.020

Laviosa, S. (1998). The Corpus-based Approach: A New Paradigm in Translation Studies. *Meta: Journal des traducteurs*, *43*(4), 474. https://doi.org/10.7202/003424ar

Laviosa, S. (2002). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta*, *43*(4), 557–570. https://doi.org/10.7202/003425ar

Le Foll, E. (2021). *Introducing the Multi-Feature Tagger of English (MFTE) version 3.0*. https://github.com/elenlefoll/MultiFeatureTaggerEnglish.

Lee, D. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning and Technology* 5(3): 37–72.

Leech, G., Hundt, M., Mair, C., & Smith, N. (2009). *Change in Contemporary English: A Grammatical Study* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9780511642210

Lefer, M.-A., & Vogeleer, S. (2013). Interference and normalization in genre-controlled multilingual corpora: Introduction. *Belgian Journal of Linguistics*, *27*, 1–21. https://doi.org/10.1075/bjl.27.01lef

Lembersky, G., Ordan, N., & Wintner, S. (2011). Language Models for Machine Translation: Original vs. Translated Texts. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 363–374, Edinburgh, Scotland, UK. Association for Computational Linguistics.

Liang, J., & Li, D. C. (2017). Researching collocational features: Towards China English as a distinctive new variety. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching Chinese English: The State of the Art* (pp. 61–75). Springer International Publishing. https://doi.org/10.1007/978-3-319-53110-6_6

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

Liu, K., Liu, Z., & Lei, L. (2022). Simplification in translated Chinese: An entropy-based approach. *Lingua*, *275*, 103364. https://doi.org/10.1016/j.lingua.2022.103364

Liu, Y., & Li, D. (2022). The US-China battle over Coronavirus in the news media: Metaphor transfer as a representation of stance mediation. *Discourse & Society*, *33*(4), 456–477. https://doi.org/10.1177/09579265221088122

Liu, Y., Cheung, A. K. F., & Liu, K. (2023). Syntactic complexity of interpreted, L2 and L1 speech: A constrained language perspective. *Lingua*, *286*, 103509. https://doi.org/10.1016/j.lingua.2023.103509

Liu, Y., Fang, A. C., & Wei, N. (2017). A Corpus-Based Study of Syntactic Patterns of Nominalizations Across Chinese and British Media English. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching Chinese English: The State of the Art* (pp. 77–92). Springer International Publishing. https://doi.org/10.1007/978-3-319-53110-6_6

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics*, *15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

MacCallum, R. C., Widaman, K. F., Zhang, S., & Hong, S. (1999). Sample size in factor analysis. *Psychological Methods*, *4*(1), 84–99. https://doi.org/10.1037/1082-989X.4.1.84

Matras, Y., & Sakel, J. (2007). Investigating the mechanisms of pattern replication in language convergence. *Studies in Language*, *31*(4), 829–865. https://doi.org/10.1075/sl.31.4.05mat

Mauranen, A. (2007). Chapter 3. Universal Tendencies in Translation. In G. Anderman & M. Rogers (Eds.), *Incorporating Corpora* (pp. 32–48). Multilingual Matters. https://doi.org/10.21832/9781853599873-006

McWhorter, J. (2007). *Language Interrupted: Signs of Non-Native Acquisition in Standard Language Grammars* (1st ed.). Oxford University Press New York. https://doi.org/10.1093/acprof:oso/9780195309805.001.0001

McWhorter, J. H. (2011). *Linguistic Simplicity and Complexity: Why Do Languages Undress?* DE GRUYTER. https://doi.org/10.1515/9781934078402

Mesthrie, R. (2006). Anti-deletions in an L2 grammar: A study of Black South African English mesolect. *English world-wide*, *27*(2), 111-145.

Mougeon, R., & Beniak, E. (1991). *Linguistic consequences of language contact and restriction: The case of French in Ontario, Canada*. Clarendon Press. Oxford University Press.

Mougeon, R., Nadasdi, T., & Rehner, K. (2005). Contact- induced linguistic innovations on the continuum of language use: The case of French in Ontario. *Bilingualism: Language and Cognition*, 8, 99– 115.

Mukherjee, J., & Hundt, M. (Eds.). (2011). *Exploring second-language varieties of English and learner Englishes: Bridging a paradigm gap*. John Benjamins Pub. Company.

Nesselhauf, N. (2009). Co-selection phenomena across New Englishes: Parallels (and differences) to foreign learner varieties. *English World-Wide. A Journal of Varieties of English*, *30*(1), 1–25. https://doi.org/10.1075/eww.30.1.02nes

Neumann, S. (2014). *Contrastive register variation: A quantitative approach to the comparison of English and German*. De Gruyter Mouton.

Neumann, S., Freiwald, J., & Heilmann, A. (2022). On the use of multiple methods in empirical translation studies: A combined corpus and experimental analysis of subject identifiability in English and German. In S. Granger & M. -A. Lefer (ed.), *Extending the Scope of Corpus-Based Translation Studies* (pp. 98–129). London: Bloomsbury Academic. http://dx.doi.org/10.5040/9781350143289.0011

Newmark, P. (1988). Pragmatic translation and literalism. *TTR: traduction, terminologie, rédaction*, *1*(2), 133-145.

Ngai, J. (2022). *Evaluation Across Newspaper Genres: Hard News Stories, Editorials and Feature Articles* (1st ed.). Routledge. https://doi.org/10.4324/9781003150640

Ni, V. (2018). Is Shanghai's Sixth Tone a New Model for China's Overseas Propaganda? *Westminster Papers in Communication and Culture*, *13*(1), 37–40. https://doi.org/10.16997/wpcc.282

Nikolaev, D., Karidi, T., Kenneth, N., Mitnik, V., Saeboe, L., & Abend, O. (2020). Morphosyntactic predictability of translationese. *Linguistics Vanguard*, *6*(1), 20190077. https://doi.org/10.1515/lingvan-2019-0077

Nini, A. (2019). The Multidimensional Analysis Tagger. In Berber Sardinha, T. & Veirano Pinto M. (Eds.), *Multidimensional Analysis: Research Methods and Current Issues,* 67-94, London; New York: Bloomsbury Academic.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 4034–4043. https://aclanthology.org/2020.lrec-1.497

Odlin, T. (1989). *Language Transfer: Cross-Linguistic Influence in Language Learning* (1st ed.). Cambridge University Press. https://doi.org/10.1017/CBO9781139524537

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* (0 ed.). Routledge. https://doi.org/10.4324/9780203640005

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

Pápai, V. (2004). Explicitation: A universal of translated text? In A. Mauranen & P. Kujamäki (Eds.), *Benjamins Translation Library* (Vol. 48, pp. 143–164). John Benjamins Publishing Company. https://doi.org/10.1075/btl.48.12pap

Pastor, G. C., Mitkov, R., Afzal, N., & Pekar, V. (2008). Translation universals: do they exist? A corpus-based NLP study of convergence and simplification In *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas:*

Plonsky, L., & Gonulal, T. (2015). Methodological Synthesis in Quantitative L2 Research: A Review of Reviews and a Case Study of Exploratory Factor Analysis: Methodological Reviews and a Case Study of EFA. *Language Learning*, *65*(S1), 9–36. https://doi.org/10.1111/lang.12111

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. (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.), *Benjamins Translation Library* (Vol. 48, pp. 165–176). John Benjamins Publishing Company. https://doi.org/10.1075/btl.48.13puu

Pym, A. (2008). On Toury's laws of how translators translate. *Benjamins Translation Library*, *75*, 311.

Quirk, R., Greenbaum, S., Leech, G., & Svartoik, J. (Ed.). (1985). *A Comprehensive grammar of the English languag*e. Longman.

Rabinovich, E., Nisioi, S., Ordan, N., & Wintner, S. (2016). *On the Similarities Between Native, Non-native and Translated Texts* (arXiv:1609.03204). arXiv. http://arxiv.org/abs/1609.03204

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*(0). https://doi.org/10.5774/45-0-198

Ren, W. (2017). Pragmatics in Chinese graduate students' English gratitude emails. In Z. Xu, D. He, & D. Deterding (Eds.), *Researching Chinese English: The State of the Art* (pp. 109–124). Springer International Publishing.

Rubino, R., Lapshinova-Koltunski, E., & Van Genabith, J. (2016). Information Density and Quality Estimation Features as Translationese Indicators for Human Translation Classification. *Proceedings of the 2016 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, 960–970. https://doi.org/10.18653/v1/N16-1110

Santini, M. (2007). Automatic genre identification: Towards a flexible classification scheme. In *Proceedings of the First BCS IRSG Conference on Future Directions in Information Access*, Glasgow, 28-29 August 2007.

Sardinha, T. B., & Pinto, M. V. (2019). *Multi-Dimensional Analysis: Research Methods and Current Issues* (1st ed.). Bloomsbury Academic; Bloomsbury Collections. https://doi.org/10.5040/9781350023857

Schmid, H.J. (2015). A blueprint of the Entrenchment-and-Conventionalization Model. *Yearbook of the German Cognitive Linguistics Association*, 3(1), 3–25. https://doi.org/10.1515/gcla-2015-0002

Scott, M. (1998). *Normalisation and readers' expectations: A study of literary translation with reference to Lispector's A Hora Da Estrela*. Liverpool: University of Liverpool. (Doctoral dissertation).

Semino, E. (2009). The language of newspapers. In J. Culpeper, F. Katamba, P. Kerswill, R. Wodak, & T. McEnery (Eds.), *English Language: Description, Variation and Context* (pp. 439-453). Palgrave Macmillan.

Shlesinger, M., & Ordan, N. (2012). More *spoken* or more *translated*?: Exploring a known unknown of simultaneous interpreting. *Target. International Journal of Translation Studies*, *24*(1), 43–60. https://doi.org/10.1075/target.24.1.04shl

Szmrecsanyi, B., & Kortmann, B. (2011). Typological profiling: Learner Englishes versus indigenized L2 varieties of English. In J. Mukherjee & M. Hundt (Eds.), *Studies in Corpus Linguistics* (Vol. 44, pp. 167–188). John Benjamins Publishing Company. https://doi.org/10.1075/scl.44.09kor

Tabachnick, B. G., & Fidell, L. S. (2007). *Experimental designs using ANOVA*. Thomson/Brooks/Cole.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. (2013). *Using multivariate statistics* (Vol. 6, pp. 497-516). Boston, MA: pearson.

Teich, E. (2003). *Cross-Linguistic Variation in System and Text: A Methodology for the Investigation of Translations and Comparable Texts*. DE GRUYTER. https://doi.org/10.1515/9783110896541

Thompson, B. (2004). Exploratory factor analysis decision sequence. In B. Thompson, *Exploratory and confirmatory factor analysis: Understanding concepts and applications.* (pp. 27–48). American Psychological Association. https://doi.org/10.1037/10694-003

Tirkkonen-Condit, S. (2004). Unique items—Over- or under-represented in translated language? In A. Mauranen & P. Kujamäki (Eds.), *Benjamins Translation Library* (Vol. 48, pp. 177–184). John Benjamins Publishing Company. https://doi.org/10.1075/btl.48.14tir

Toury, G. (1995). *Descriptive Translation Studies and Beyond*. Amsterdam: John Benjamins.

Ulrych, M., & Murphy, A. C. (2008). Descriptive translation studies and the use of corpora: Investigating mediation universals. In *Corpora for university language teachers* (pp. 141-166). Peter Lang.

Van Doorslaer, L., & Gambier, Y. (2015). Measuring relationships in Translation Studies. On affiliations and keyword frequencies in the Translation Studies Bibliography. *Perspectives*, *23*(2), 305–319. https://doi.org/10.1080/0907676X.2015.1026360

Van Rooy, B., Terblanche, L., Haase, C., & Schmied, J. J. (2010). Register differentiation in East African English: A multidimensional study. *English World-Wide. A Journal of Varieties of English*, *31*(3), 311–349. https://doi.org/10.1075/eww.31.3.04van

Virtanen,T. (2005).“Polls and surveys show”: Public opinion as a persuasive device in editorial discourse. In H. Halmari & T. Virtanen (Eds.), *Persuasion across genres: A linguistic approach* (pp. 153–180). Amsterdam/Philadelphia: John Benjamins.

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

Wallis, S. (2020). *Statistics in Corpus Linguistics Research: A New Approach* (1st ed.). Routledge. https://doi.org/10.4324/9780429491696

Wang, B., & Zou, B. (2018). Exploring Language Specificity as a Variable in Chinese-English Interpreting. A Corpus-Based Investigation. In M. Russo, C. Bendazzoli, & B. Defrancq (Eds.), *Making Way in Corpus-based Interpreting Studies* (pp. 65–82). Springer Singapore. https://doi.org/10.1007/978-981-10-6199-8_4

Wang, W. (2008a). Intertextual aspects of Chinese newspaper commentaries on the events of 9/11. *Discourse Studies*, *10*(3), 361–381. https://doi.org/10.1177/1461445608089916

Wang, W. (2008b). Newspaper commentaries on terrorism in China and Australia: A contrastive genre study. In U. Connor, E. Nagelhout, & W. Rozycki (Eds.), *Pragmatics & Beyond New Series* (Vol. 169, pp. 169–191). John Benjamins Publishing Company. https://doi.org/10.1075/pbns.169.11wan

White, P. R. R. (2000). Dialog and Inter-subjectivity: Reinterpreting the semantics of modality and hedging. In M. Coulthard, J. Cotterill & F. Rock (Eds.), *Working with dialog* (pp. 67–80).Tübingen: Max Niemeyer Verlag.

Williams, D. A. (2005). *Recurrent features of translation in Canada: A corpus-based study*. (Doctoral dissertation), University of Ottawa. https://doi.org/10.20381/RUOR-12864

Williams, J. (1987). Non-native varieties of English: A special case of language acquisition. *English World-Wide*, *8*(2), 161-199.

Wulff, S., Lester, N., & Martinez-Garcia, M. T. (2014). *That* -variation in German and Spanish L2 English. *Language and Cognition*, *6*(2), 271–299. https://doi.org/10.1017/langcog.2014.5

Xiao, R. (2010). How different is translated Chinese from native Chinese?: A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, *15*(1), 5–35. https://doi.org/10.1075/ijcl.15.1.01xia

Xiao, R., & Cao, Y. (2013). Native and non-native English abstracts in contrast: A multidimensional move analysis. *Belgian Journal of Linguistics*, *27*(1), 111-134. https://doi-org.ezproxy.lb.polyu.edu.hk/10.1075/bjl.27.06xia

Xu, C. (2021). Identification of L2 interpretese: a corpus-based, intermodal, and multidimensional analysis. (Doctoral dissertation), the Hong Kong Polytechnic University. https://theses.lib.polyu.edu.hk/handle/200/11010

Xu, J., & Liang, M. (2013). A tale of two C's: Comparing English varieties with Crown and CLOB (the 2009 Brown family corpora). *ICAME Journal*, *37*(1), 175-183.

Xu, Z. (2008). Analysis of Syntactic Features of Chinese English. *Asian Englishes*, *11*(2), 4–31. https://doi.org/10.1080/13488678.2008.10801233

Xu, Z. (2010). *Chinese English: Features and implications*. Open University of Hong Kong Press.

Xu, Z. (2020). Chinese English——A future power? In *The Routledge Handbook of World Englishes* (2nd Edition, pp. 265–280). Routledge.

Zanettin, F., Saldanha, G., & Harding, S.-A. (2015). Sketching landscapes in translation studies: A bibliographic study. *Perspectives*, *23*(2), 161–182. https://doi.org/10.1080/0907676X.2015.1010551