EPISTEMIC NECESSITY AND OBLIGATION MODALS IN CHINESE EFL ACADEMIC WRITING: EXPLORING L1 INFLUENCE AND DISCIPLINARY VARIATION ON THE USE OF *MUST*, *HAVE TO*, AND *SHOULD*

by

QIUYI SUN

A thesis submitted to the University of Birmingham for the degree of

DOCTOR OF PHILOSOPHY

Department of English Language and Linguistics

School of English, Drama and Creative Studies

College of Arts and Law

University of Birmingham

May 2024

Abstract

The use of modals is a key feature in academic writing, yet Chinese EFL students often find them challenging. Substantial research has been undertaken on their use in short argumentative essays on general topics, but little attention has been given to discipline-specific academic writing. To address this, this study examines Chinese EFL undergraduate dissertations, focusing on three epistemic necessity and obligation modals: *must*, *have to*, and *should*, and explores how first languages and disciplines influence their use.

This study employs a corpus-based approach to examine the modals in terms of frequency, meaning distribution, and main verbs co-occurring with them. The semantics of the verb collocates are systematically examined through distributional semantic analysis, an approach not previously applied in the study of modality in EFL student writing. The quantitative analysis is complemented by a qualitative analysis exploring distinctive features of the modals in academic writing. A Chinese learner corpus (LC), consisting of dissertations in two disciplines, Business and Management (BM), and English Literature (EL), is compared to a reference corpus (RC) that includes essays written by British students in comparable disciplinary groups, Social Science (SS) and Arts and Humanities (AH), extracted from the British Academic Written English corpus.

Epistemic use of the modals does not show statistically significant differences between Chinese and British students, though there is a slight under-representation of epistemic *must* in the learner corpus, which contradicts previous literature. By contrast, the two student groups use root sense of the modals significantly differently, with root *must* being under-represented and root *should* being over-represented in the learner corpus. Chinese students in LC-BM use root *must* and *should* similarly with two clusters of verb

collocates to give practical suggestions for business, whereas British students in both disciplines tend to use them with verbs for assessing propositions. The variations between student groups may be indicative of influences from first languages, cultural values, and textbook presentations. In terms of disciplinary variations, epistemic *must* is under-represented in LC-BM and RC-SS compared to LC-EL and RC-AH, whereas root *must* and *should* show an opposite pattern. These variations could be explained by different analytical approaches and rhetorical conventions. The qualitative analysis provides a fine-grained view to explore the characteristic features of the modals in academic writing. Variations across sub-corpora are identified in terms of the textual voice expressed by the modals and their distribution across different parts of a text.

Following a mixed-methods approach, the study contributes to a more comprehensive view of *must*, *have to*, and *should* in Chinese EFL academic writing and underscores the influence of the first language and discipline.

Acknowledgements

I am deeply indebted to my supervisors, Dr. Paul Thompson and Dr. Florent Perek, whose extensive guidance and academic support have been paramount throughout my PhD journey. Dr. Thompson's expertise in academic writing and his knack for clarifying and formulating complex ideas have significantly shaped my research path. His meticulous feedback and ability to challenge and refine my thinking have been instrumental in honing my analytical skills and academic rigour. Dr. Perek's assistance with data analysis and technical insights have greatly contributed to the depth of my analysis.

My appreciation goes to Dr. Yanli Zou for providing access to the essential Chinese EFL learner corpus data, which has been a cornerstone of my thesis. I am also grateful to my fellow postgraduate students of the PG Tips group for their support and the enriching discussions that have helped sustain my motivation and intellectual curiosity throughout my research. Special thanks to Dr. David Roxburgh for his unwavering support throughout my PhD, through regular discussion and ongoing engagement.

Lastly, I am deeply thankful to my parents, Hong and Qinghua, for their love, patience, and belief in my abilities. Their constant support has been my anchor during this challenging and rewarding journey.

Note: The reference corpus data in this study come from the British Academic Written English (BAWE) corpus, which was developed at the Universities of Warwick, Reading and Oxford Brookes under the directorship of Hilary Nesi and Sheena Gardner (formerly of the Centre for Applied Linguistics, Warwick), Paul Thompson (formerly of the Department of Applied Linguistics, Reading) and Paul Wickens (School of Education, Oxford Brookes), with funding from the ESRC (RES-000-23-0800).

Table of Contents

List of Figures

List of Tables

List of Key Abbreviations

AF	Absolute frequency
AH	Arts and Humanities
BAWE	British Academic Written English (corpus)
BM	Business and Management
BNC	British National Corpus
CA	Contrastive Analysis
CAEL-CAWE	Chinese Advanced English Learner Corpus of Academic Written English
CEFR	Common European Framework of Reference for Languages
CIA	Contrastive Interlanguage Analysis
CLEC	Chinese Learner English Corpus
COCA	Corpus of Contemporary American English
COHA	Corpus of Historical American English
CSE	Chinese Standards of English Language Ability
EAP	English for Academic Purposes
EFL	English as a Foreign Language
EL	English Literature
ESL	English as a Second Language
ICE	International Corpus of English
ICLE	International Corpus of Learner English
ICNALE	International Corpus Network of Asian Learners of English
L1	First language
L2	Second language
LC	Learner corpus
LOCNESS	Louvain Corpus of Native English Essays
LS	Learner sample
LSWE	Longman Spoken and Written English Corpus
MICUSP	Michigan Corpus of Upper-level Student Papers
NF	Normalised frequency
RC	Reference corpus
RS	Reference sample
SS	Social Science
TLE	Treebank of Learner English

1 INTRODUCTION

1.1 Introduction

1-1 So why not try to find a part time work? And then our college life *must* be more colorful. (example provided in Hu & Li, 2015, p. 26)

1-2 If your friend is sick, you *must* visit him and cook for him and take care of him. You *have to* talk to him about gossip to give him amusement. If you don't do these things, he will think you are not a friend for him. (example provided in Hinkel, 1995, p. 331)

The two examples are taken from argumentative texts written by Chinese EFL (English as a Foreign Language) university students. These examples are grammatically correct, but pragmatically, there seems to be some inappropriateness. *Must* in Example 1-1 expresses a strong certainty about the impact of finding part-time work on the colourfulness of college life. Nonetheless, this level of certainty might be excessively definitive, failing to account for individual variations in experiencing college life. Substituting *must* with *should* modifies the statement to indicate an expected outcome rather than a guaranteed result, thereby allowing for greater individual differences.

Must and *have to* in Example 1-2 both convey a strong sense of obligation. In the context of advising on how to care for a sick friend, these modals impose a high level

of duty on the addressee, suggesting these actions are not just recommended but required for the maintenance of the friendship. However, the addressee may think that this perspective overlooks the variety of ways in which friends can express their care and support, implying a limited view of friendship that does not accommodate different cultural or personal preferences.

These examples show that the nuanced use of modals is not merely a grammatical concern but also a pragmatic one, requiring an understanding of subtle differences in meaning and context. For Chinese EFL students, this aspect of English language learning is particularly challenging in academic writing (Yang, 2018). Previous literature, including the studies from which the examples are drawn, has acknowledged that Chinese EFL students use modals in a manner distinct from native English speakers. However, most of the investigations focus on short argumentative essays (under 400 words) covering diverse topics rather than being discipline-specific, largely due to the accessibility of Chinese learner corpora. This raises two questions. Firstly, are the findings of these argumentative essays applicable to discipline-specific academic writings, such as dissertations? Secondly, does modal use in Chinese EFL academic writing vary by discipline, as observed in research articles (e.g., Peacock, 2014; Vičič & Petek, 2016; Yang et al., 2015)?

This study aims to answer these two questions, providing a more comprehensive profile of modals in Chinese EFL academic writing with a focus on three epistemic necessity and obligation modals, *must*, *have to*, and *should*. Specifically, it explores how first language and discipline may impact modal use in dissertations using a corpus-based approach. The analysis extends beyond mere frequency and meaning distribution, delving into the semantics of main verbs co-occurring with the modals and the distinctive features of modals in academic writing by mixing the quantitative and qualitative approaches. It is hoped that these insights will contribute to a deeper understanding of modality in academic writing.

The remainder of the chapter is structured as follows: Section 1.2 provides a brief review of previous literature and introduces the background of the study. It is followed by a statement of the aims and research questions to be addressed. Section 1.4 presents the significance of the study and the last section offers an outline of the whole thesis.

1.2 Background of the study

A detailed literature review will be found in Chapter 2. This section discusses key issues identified by previous studies and introduces the background of the study.

Modality has been a major area of interest in linguistics in the past few decades. While there is a lack of consensus about how the different types of modality should be classified, most researchers agree that modality is a semantic category expressing a speaker's 'opinion or attitude' (Lyon, 1977, p. 452). The range of modal expressions includes modal verbs, modal adverbs, modal lexical verbs, among others. Specifically, modal verbs receive considerable attention in linguistics and are the primary focus of this study. Regarding terminology, *modal* is applied broadly, referring to both central and semi-modal verbs. Thus, all three verbs under discussion are referred to as *modals*.

Initially, research into modals relied on a small number of texts (e.g., Joos, 1964) or invented examples (Halliday, 1970). As corpus studies have advanced, analysis now involves a broader dataset with authentic examples. This progress allows for a more detailed and representative investigation of modal use. In addition, the investigation shifts from a general discussion to a finer-grained assessment of how modality is used in various registers, genres, and disciplines.

Among the studies, modality in academic writing has gained much attention. Academic writing is initially viewed as impersonal and neutral prose, reflecting the prioritisation of objectivity in scholarly work. However, through the years, the perception has shifted towards recognising academic writing as both interactive and personal. This view

highlights the importance of conveying the writer's identity, along with their confidence in their assessments and dedication to their propositions (Hyland, 2002a), with modality being the key in achieving these aims.

Generally, modality can be classified into two categories: epistemic and non-epistemic (see Section 2.2.2 for a full discussion). Epistemic modality shows the writer's judgement of the truth of a proposition (Coates, 1983). In academic writing, it enables writers to situate their arguments on a spectrum of certainty, reflecting the rigour of their research, the openness to further investigation and the willingness to engage with readers. As for non-epistemic modality, they express a range of meanings, including permission, obligation, ability, and volition. In particular, permission and obligation modals play a key role in the interpersonal aspect of academic writing, balancing the writer's authority and the reader's expectations. For instance, by modulating arguments with obligation modals, the writers can position themselves as a member of the disciplinary community and lay obligations on knowledge construction and practical implications.

Previous research has established that the profile of modals in academic writing differs from that in other contexts such as general written English (e.g., Parkinson, 2020) or conversations (Biber et al., 1999) in terms of frequency and meaning distribution.

Another distinctive feature is related to textual voice, which involves the identification of the voice expressed by modals. Unlike conversations in which the voice is almost exclusively from the speaker, the case in academic writing is more complicated, which may include the voice of the writer, the antecedent authors, and other third parties such as organisations or the government.

Although studies such as Huddleston (1971) and Thompson (2001) mention this aspect in passing, little research gives a satisfactory definition of modality in academic writing. Most of them simply apply the definition discussed in previous literature, which prioritises spoken language data and omits the inclusion of textual voice (e.g., Coates, 1983). The present study argues that the extent to which these descriptions can be applied to the discussion of modality in academic writing requires further consideration and gives a working definition in Section 2.3.1. Given the difficulty of quantitatively analysing textual voice due to the extensive time required for manual annotation, this study also employs a qualitative analysis to explore this aspect and the distribution of modals across different parts of a text. This mixed-methods approach allows for a nuanced examination beyond numerical data.

It should also be noted that most studies on modality in academic writing have examined research articles, possibly due to the greater accessibility of these texts

compared to other academic genres. Additionally, there is a discernibly uneven coverage of the different meanings of modality being explored. Epistemic modality, particularly that expressing possibility, has gained extensive focus since it deals with the degree of certainty, which is central to constructing arguments and positioning the findings in an ongoing academic discussion. In contrast, non-epistemic modality such as that expresses obligation and suggestion is relatively underexamined, with only a handful of studies addressing these areas such as Vincent (2020) and Parkinson (2020). This research gap can also be observed in studies of modality in EFL student writing, and in particular, Chinese EFL student writing.

Thus, this study focuses on epistemic necessity and obligation modals since they are largely neglected by previous studies. There are a handful of such modals. To delimit the population of the study and operationalise the investigation, this study targets three: *must*, *have to*, and *should*. This selection is guided by a form-based approach, rather than a function-based one, to facilitate straightforward extraction from large corpora. If taking a function-based approach, there is a danger that the annotation may differ across raters and will therefore be hard to generalise and compare the findings. The relatively small number of specific modals under consideration allows for a fine-grained assessment of each modal and permits a distinction between their uses, which is vital given their similarities in meanings and certain features that pose challenges for

language learners, as illustrated in Examples 1-1 and 1-2. Another reason is that during the pilot study (see Section 3.3), *must*, *have to*, and *should* are found to be used markedly more frequently compared to other modals that convey similar meanings. For instance, *have got to* appears only four times in the learner corpus, whereas *need* is used 19 times. Focusing on *must*, *have to*, and *should* could provide a robust dataset for examining the nuances of modal use in academic writing.

While examining modals in research articles may offer insights applicable to student academic writing, such an approach could be misleading due to the substantial differences between these two genres. Therefore, focusing on student writing, as initially suggested by Dudley-Evans (1999), seems to be a more suitable approach for informing teaching practices. In the case of Chinese EFL writing, there has been substantial research undertaken on the frequency distribution of modals in short argumentative essays on various topics such as family and friendship (e.g., Liang, 2008; Tang, 2013), which may result from the scarcity of other accessible Chinese learner corpora. It has been revealed that factors potentially causing difficulty in using modals include the differences in cultural value (Li, 2016), textbook presentation (Li, 2020), and L1 (first language) influence (Yang, 2018). However, to date, discipline-specific academic writing by Chinese EFL students has not been closely studied. One exception is Yang (2018), who compiled a corpus of Chinese undergraduates' research

reports for the course of International Business and Trade to examine the use of central modals. The current study seeks to address this gap by examining dissertations of Chinese EFL students in two disciplines and explores whether Chinese students use modals differently compared to native English speakers.

A further question is, do the profiles of modals differ between disciplines? It has been found that modality shows disciplinary variations in research articles (e.g., Peacock, 2014; Yang et al., 2015). Could we find similar patterns in Chinese student academic writing? Disciplines have been categorised by Becher and Trowler (2001) into four basic dimensions along two axes: soft-hard, and pure-soft. Research on modality in research articles has a preference to compare disciplines between soft and hard disciplines, potentially due to their distinct features in research approaches, interpretations of data and rhetorical conventions. Disciplines that are both categorised as, for example, soft disciplines are less examined, with a few exceptions such as Takimoto (2015), who compares writings of Humanities and Social Science.

In the context of Chinese EFL academic writing, disciplinary variations in modals are rarely examined since previous research has predominantly focused on argumentative essays that are not specific to any discipline, as mentioned before. There is a need to conduct a more systematic and well-rounded analysis of modals used in Chinese EFL

academic writing and explore their disciplinary variations. As pointed out by Zou (2018), students in most disciplines in China are not required to submit dissertations in English, and this leads to a limited number of texts that can be collected. Two disciplines, Business and Management, and English Literature, are the exceptions. The present study compares the modals between these two disciplines in a learner corpus compiled by Zou (2018), aiming to offer insights into disciplinary variations in modal use in Chinese EFL academic writing. Upon detailed examination, texts from these disciplines in the learner corpus could be categorised under Social Sciences and Arts and Humanities (see Section 3.2.2 for discussion), respectively, which are generally considered as soft disciplines.

At a more specific level, profiles of modals can be examined in various aspects. Research on modals initially focused on frequency and meaning distribution. Over time, studies have expanded to explore their co-textual features. For instance, Coates (1983) identifies the association between meanings of modal verbs and their co-occurring syntactic features such as subjects, voice, and aspect. However, studies on Chinese EFL writing still seem to primarily focus on frequency and meaning distribution, while overlooking the co-textual features of modals. Consequently, our understanding of the profiles of modals in Chinese EFL academic writing remains superficial and incomplete. The present study aims to provide a more comprehensive profile of modals in Chinese

EFL academic writing, including a discussion on syntactic features (e.g., negation, voice, aspect and tense) and semantics of their verb collocates. Among these features, the semantics of the main verbs used with the modals are the primary focus. These verbs convey information about what action is modulated by the modals, which can, on the one hand, offer new perspectives into the profiles of the modals, and on the other hand, potentially explain variations in modal use between student groups and between disciplines. In the case of epistemic use, these verbs could show what the writers' certainty is about. As for the root use, the semantics of verbs can illustrate what obligation or suggestion is presented by the writer. The verb collocates of modals are briefly discussed in previous literature, with few semantic patterns noted such as the association between the stative verbs and the epistemic use of the modals (Biber et al., 1999), but they have not been systematically analysed. The present study applies distributional semantic analysis to explore them on a broader scale, offering a new perspective and a more detailed profile of the similarities and differences in the modal use between student groups and between disciplines.

1.3 Aims and research questions

The study uses a corpus-based approach along with a qualitative analysis of selected texts to examine the profiles of *must*, *have to*, and *should* in Chinese EFL undergraduate academic writing. The research compares two corpora, a learner

corpus and a reference corpus in comparable disciplines to explore the similarities and differences of the modal use between student groups and between disciplines. Texts in the learner corpus (LC) are extracted from the Chinese Advanced English Learner Corpus of Academic Written English (CAEL-CAWE; Zou, 2018), which consists of dissertations written by Chinese EFL students in mainland China. Two of the sub-corpora are used in the present study, the academic texts written by the undergraduates majoring in Business and Management (BM) and those in English Literature (EL). Comparable texts in the British Academic Written English (BAWE) corpus (Nesi et al., 2008) are compiled as the reference corpus (RC), consisting of essays written by L1 English undergraduates in Social Science (SS) and Arts and Humanities (AH).

For the quantitative analysis, the study compares the learner corpus with the reference corpus to investigate the profiles of *must*, *have to*, and *should*, identifying patterns between the two groups of students. Additionally, disciplinary variations are explored by contrasting the comparable sub-corpora in the two corpora. The corpus-based analysis provides an overview of how the target modals are used in terms of frequency, meaning distribution, dispersion and co-textual features with a focus on verb collocates. A fine-grained view is taken by conducting qualitative analysis on 16 sample texts. This examination not only revisits the aspects highlighted in the quantitative analysis but

also explores whose voice is expressed by these modals and their distribution across different parts of a text, thereby providing a deeper understanding of the distinctive features of modals in academic writing compared to other contexts.

The overall aim of the study is to examine how *must*, *have to*, and *should* are used in Chinese EFL student academic writing and in different disciplines. Specifically, the following research questions will be addressed:

RQ 1: How frequently are *must*, *have to*, and *should* used in the Chinese EFL learner corpus and the reference corpus?

RQ 2: How are the meanings of the three modals distributed?

RQ 3: What semantic patterns can be identified regarding the main verbs that collocate with the three modals?

RQ 4: Do the profiles of the three modals differ between the two student groups, Chinese and British students?

RQ 5: Do the profiles of the three modals differ between the disciplines?

The first three research questions each focus on one aspect of the profiles of *must*, *have to* and *should* for the analysis, and the last two questions serve as overarching themes for the study, exploring the variations in modal use between Chinese and

British students on the one hand and between students in different disciplines on the other.

1.4 Significance of the study

As mentioned in Section 1.2, the studies on modality in academic writing have tended to base their discussion on previous literature on modality used in spoken material, potentially omitting specific modal features in academic writing. The present study adds to our current understanding of how modality underpins the presentation of a writer's viewpoints and the construction of knowledge in academic writing. The specific focus on *must*, *have to*, and *should* fills a gap in the research on epistemic necessity and obligation modals and helps to provide a fine-grained assessment of these modals. In addition, previous studies in disciplinary variation mostly compare modality in disciplines with discrete features such as soft and hard disciplines, whereas this study examines disciplines that are not assumed to have salient differences. It is hoped that this research will contribute to a deeper understanding of modals used in various disciplines.

At a more specific level, this thesis also contributes to a comprehensive description of the three modals used by Chinese EFL students in discipline-specific academic writing. Previous studies mostly examine short argumentative writing on various topics, but

modals in discipline-specific dissertations have not yet been closely investigated. The present findings reveal differences in modal use in these dissertations compared to what previous studies have suggested. For example, *must* is observed to be over-represented in Chinese EFL argumentative writings (Cheng & Qiu, 2007; Liang, 2008; Long, 2013) in contrast to native English speakers, but it is under-represented in the present study.

By mixing quantitative analysis with qualitative interpretation, this research not only maps out the general profiles of the modals across sub-corpora, but also reveals the nuanced communicative functions these modalities serve in academic writing, such as the textual voice expressed by the modals and their distribution across different parts of a text. The investigation goes beyond frequency and meaning distribution, which is predominantly emphasised by previous studies, looking at the main verbs collocating with the modals. Semantic patterns of the verb collocates are identified to differ between student groups and between disciplines. For example, Chinese students in Business and Management are found to use root *must* and *should* with two clusters of verb collocates to suggest practical actions, whereas British students in both disciplines tend to use two other verb clusters to assess propositions. It is also identified that some semantically similar verbs are used more frequently with one of the target modals, such as the association between *have to* and verbs related to

unpleasant experiences (e.g., *suffer* and *tolerate*).

Methodologically, this thesis is the first investigation to use distributional semantic analysis in examining modality in EFL student academic writing and its disciplinary variations. Verbs co-occurring with the three modals in the sub-corpora are placed in the semantic plots based on their semantic similarity and comparisons are made between two student groups and between disciplines. Prior to the present study, this approach has been mostly used to explore the diachronic development of linguistic devices (e.g., Hilpert & Perek, 2015) or to distinguish between synonyms (Hilpert & Flach, 2021; Levshina & Heylen, 2014).

Finally, this study sheds new light on teaching material and English for Academic Purposes (EAP) practices, such as the presentation of the epistemic and root use of the three modals in textbooks, and raises awareness of disciplinary variations in modal use.

1.5 Organisation of the thesis

The thesis is composed of seven chapters. This chapter has presented the background of the study, the aims and research questions, and the significance of the study. Chapter 2 reviews previous literature on English modality, and more specifically,

modality in academic writing. It points to the need to look at the co-textual features of the modals and the importance of identifying distinctive features of modals in academic writing. The chapter continues with a review of modal use in Chinese EFL student writing and in different disciplines since first language and discipline are the two factors examined in the present study. It then concludes with a brief introduction of learner corpus research and Contrastive Interlanguage Analysis (CIA) to illustrate the theoretical background of the research design.

Chapter 3 describes how CIA is applied, starting with the rationale behind selecting the learner and the reference corpus. It is followed by the procedures to finalise the data and the process for selecting and extracting the target modals. Additionally, this chapter elaborates on the overall analytical approach used, mixing the quantitative and the qualitative analysis, with the specifics of each approach described in the method section of their respective chapters.

Chapters 4 and 5 start with a thorough explanation of the methods taken and present the quantitative and qualitative findings respectively. In Chapter 4, profiles of the three modals in Chinese EFL student writing are demonstrated in terms of frequency, dispersion, meaning distribution, syntactic features and semantics of the verb collocates. Chapter 5 provides a finer-grained assessment of the modal use in

academic writing by examining a small number of texts, focusing specifically on the textual voice expressed by the modals and their distribution across different parts of a text. Variations between the two student groups and between disciplines are identified in these two chapters. From these findings, a more comprehensive picture of modal use in Chinese EFL academic writing emerges.

In Chapter 6, a discussion of the findings is presented, answering the research questions and exploring the potential reasons for the variations in modal use. Chapter 7 summarises two main arguments proposed in the thesis and points out the theoretical, methodological, and pedagogical contributions. It then draws the thesis to a conclusion with a discussion on limitations and suggestions for future study.

2 LITERATURE REVIEW

2.1 Introduction

This chapter reviews prior literature on several topics, aiming to establish a solid theoretical foundation for the present study and to critically evaluate previous studies to highlight gaps in the current understanding.

As discussed in Section 1.2, the focus of this study is to examine the profiles of *must*, *have to*, and *should* in Chinese EFL students' academic writing. Before examining the specific profiles of these three modals in Section 2.2, a general discussion on the modality in English, including its definition, classifications, dimensions, and expressions, is presented to establish a comprehensive framework for understanding modality. This is crucial for analysing modal use in any English writing, including that of Chinese EFL students. In addition, co-textual features of modals are also presented to illustrate the link between these features and modal use. In particular, main verbs that co-occur with the modals are the focus, and thus I give an overview of the classification of verbs to facilitate the following data analysis. It is then followed by a detailed introduction of *must*, *have to*, and *should* since they are the modals in question.

Following this general discussion, Section 2.3 narrows down the topic, giving a working

definition of modality in academic writing with a focus on textual voice. It is followed by a discussion of empirical studies on this topic to show that firstly, modal use differs in various contexts, and secondly, to give an overview on this topic and identify potential research gaps. Chinese modality and its classifications are introduced in Section 2.4 to identify its similar and different patterns compared to English modality. Following this, studies on Chinese EFL students' use of modals are examined to signal areas of challenges students may encounter, the influence of first language, and pedagogical needs. Section 2.5 presents disciplinary variations in the use of modality, highlighting the disciplinary feature of modal use and potential reasons. This is essential for contextualising the modal choices of Chinese EFL students in specific disciplines that the present study focuses on. The last section begins with an overview of the development of learner corpus research, followed by a description of the Contrastive Interlanguage Analysis, which is the framework that guides the research design.

By reviewing these topics, I attempt to underpin the present study on the profiles of three modals, *must*, *have to*, and *should* in Chinese EFL students' academic writing, focusing on the meanings of these modals and their verb collocates.

2.2 Modality in English

The aim of this sub-section is twofold: 1) to provide an overview of English modality in linguistics, including its definition, classifications, dimensions, and expressions (Sections 2.2.1 to 2.2.3); and 2) to discuss the more fine-grained topics in the present study, such as the co-textual features of the modals with a focus on verb collocates and a specific description of the profiles of *must*, *have to*, and *should* (Sections 2.2.4 and 2.2.5).

2.2.1 Definition of modality

Modality is a well-discussed topic in various disciplines, such as logic, philosophy, and linguistics. Despite the fact that the study of modality can be traced back to the time of Aristotle, it is not until the 20th century that linguists start to investigate it as a language phenomenon. This section will focus on how modality is defined in linguistic studies, providing a basis for further exploration of its implications and applications in language analysis.

Before digging into what modality is, it is necessary to look at a relevant concept, *mood*. Mood and modality are mostly mentioned together in the literature due to their close relationship. As Huddleston and Pullum (2002) conclude, the difference between these two concepts is similar to that between tense and time. Mood is a grammatical feature reflected by verb forms and structures, whereas modality is a semantic category

realised by modal expressions.

Although widely researched, there is no universally accepted linguistic definition of modality. One of the pioneering studies is Joos (1964), which examines reports of a murder trial held in London. Joos notes that modals 'assert a specific relation between an event and the factual world' (p. 149) rather than assert the event itself. He believes that modals are mono-semantic and are independent of contexts. Similarly, Ehrman (1955) argues that there is one basic meaning that can apply to all instances of modals. These mono-semantic views of modality make the definition relatively vague since it needs to cover all the semantic range of modals.

Lyon (1977), on the other hand, gives a more satisfactory definition, relating modality with the expression of the speaker's 'opinion or attitude' (p. 452). This is commented on by Palmer (1990) as a helpful definition, as it reveals the common features shared by different formal systems conveying modality across languages, such as modal auxiliaries in English and modal particles in Chinese. Lyon also distinguishes two main categories of modality, epistemic and deontic. This is further adopted by Palmer (2001), who describes epistemic modality concerning the speaker's judgement of the truth of a proposition, whereas deontic modality expresses 'the speaker's attitude towards a potential future event' (p. 8). Huddleston and Pullum (2002) capture these two

meanings and provide a comprehensive definition, describing modality as being 'centrally concerned with the speaker's attitude towards the factuality or actualisation of the situation expressed by the rest of the clause' (p. 173). The terms *factuality* and *actualisation* refer to the epistemic and non-epistemic use of the modality respectively. Although researchers such as Palmer (1990) also propose a third category of meaning, dynamic modality, this category is rarely included in the definition, and is mostly discussed as an afterthought, which will be discussed in Section 2.2.2.

As you may notice, the definitions mentioned above concern mostly the speakers but not the writers. This is partly because previous literature in most cases examines spoken data when investigating modality, and few of them focus on written material. Even if some researchers, such as Palmer (1990) and Collins (2009), examine both the spoken and written data, they still refer to the ones who use modality as speakers. This omission of the distinction in the use of modality in different modes of communication raises concerns, which will be discussed further in Section 2.3.1, where a working definition of modality in academic writing will also be proposed.

The data used in modality studies has undergone a development from invented examples to authentic language use. Early studies (e.g., Diver, 1964; Halliday, 1970) examine invented examples, and the researchers use the approach of paraphrasing

and introspection for investigation since they are native speakers of English. Invented examples can provide simplified sentences to illustrate, for example, the different meanings of modals, which is beneficial for pedagogical purposes. Despite that, the drawbacks are the tendency to be schematic and the lack of authenticity (Hermerén, 1978). Despite being native speakers of English and skilled linguists, researchers may still overlook certain language uses due to individual differences, such as varieties of English (e.g., American or British English), regional dialects, and personal linguistic styles. In addition, invented examples are mostly short in length and are used in the main clause, which lacks context and thus limits the investigation.

By contrast, with the development of corpus studies and more corpora publicly available, corpus-related studies on modality have the advantage of using authentic and spontaneous texts with more contexts provided. This also helps to progress the modality research from focusing on forms and meanings to the relations between modal use and co-textual features, such as aspect, voice, and main verbs used with the modals (see Section 2.2.4 for discussion).

There are three main threads of studies distinguished by their purposes of using corpora, as observed by Collins (2009): corpus-informed, corpus-based, and corpus-driven. I order them by the depth of interaction and involvement of the corpus data.

Firstly, corpus-informed studies on modality use instances taken in the corpora only for exemplification purposes with no mention of their frequency information. To list a few, these studies include the comprehensive grammar studies of Quirk et al. (1985) and Huddleston and Pullum (2002), and the modality studies of Ehrman (1966) and Palmer (1990). Ehrman is one of the first researchers who uses a corpus for modality research, exploring one-third data of the Brown University Corpus of American English (the Brown corpus; Francis & Kucera, 1979). Palmer (1990) uses the spoken and written material collected in the Survey of English Usage (University College London, n.d.), and he states that the material is only used for 'heuristic and exemplification purposes' (p.29).

Unlike the studies mentioned above, corpus-based studies also include frequency data and examine other variations such as modes of genres, varieties of English, and registers. Corpus is not only used to provide examples but also to inform hypotheses and provide evidence. For example, Coates (1983) proposes the association between the meaning of modals and their syntactic features. Although the spoken material analysed by Coates (1983) and Palmer (1990) both comes from the corpus of the Survey of English Usage (University College London, n.d.), only Coates gives the statistical information of the size of the corpus and meta-information of different categories of texts. Collins (2009) is another example, focusing on the modal use in

different varieties of English, and Biber et al. (1999) distinguish the modal use in different registers. The last type is the corpus-driven study, which uses the corpus to formulate new linguistic theories, such as pattern grammar and construction grammar.

The present study is a corpus-based investigation of modality. Corpus is used not only for illustration but also to provide quantitative information on the frequency and meaning distribution of the modals.

2.2.2 Classification of modality

Not only do the definitions vary, but the classifications of meanings of modality also show differences. As mentioned before, early works such as Joos (1964) and Ehrman (1966) claim that one unitary meaning can be identified for each modal. This mono-semantic perspective, although it acknowledges the indeterminacy of meanings, denies the existence of discrete categories (Coates, 1983). While one meaning can be possibly applied to all uses of a modal, it would be too general and not informative enough because it does not take the context into consideration. By contrast, other researchers, such as Coates (1983) and Palmer (1990), take a poly-semantic view of modality, classifying it into discrete categories and formulating different classification frameworks. In the following paragraphs, I will briefly discuss the various classifications and justify my choice for the present analysis. For a more extensive survey of the categorisation put forth by different researchers, the works of Vincent (2015) and

Depraetere and Reed (2020) provide an insightful summary.

Although the classification of modality varies to some extent, most researchers agree that two types of modality are distinct from each other, epistemic and non-epistemic modality. Epistemic modality includes linguistic expressions that show the attitudes of the speaker in the assessment of the truth of propositions (Coates, 1983), as illustrated in Example 2-1.

> 2-1 There *must* be a lot more to it than that I am sure it wasn't just that because they appear to … get on very well. (example provided in Coates, 1983, p. 41)

In this example, *must* expresses the speaker's confidence in the truth of the proposition that 'there is a lot more to it', and this judgement is based on the explicitly stated reason that 'they get on very well'. In addition, 'I am sure' is used with epistemic *must* to express similar meanings, and their combination can be considered harmonic (see Section 2.2.3 for discussion).

Example 2-1 is identified by Coates (1983) as the core and stereotypical use of epistemic sense, as it reflects the subjective judgement of the speaker. By contrast, a periphery example would be objective in terms of logical inference, as shown in 2-2.

This case does not involve the speaker's judgement. Instead, *must* conveys a pure logical inference based on the known fact that 'if there is an endeavour to x, attention to x follows'. *Certainly* is used to underscore the confidence in the judgement.

> 2-2 Shall we then say with G.F Stout that 'desire and aversion, endeavour to and endeavour from, are modes of attention'? Certainly if there is endeavour to x, there *must* be attention to x. (example provided in Coates, 1983, p. 42)

Two pairs of terms used in the preceding paragraphs that require further explanation are ‘core vs. periphery’ and ‘subjectivity vs. objectivity’. These concepts are interconnected. *Core*, *periphery*, and *skirt* are introduced in the fuzzy set model proposed by Coates (1983). This model helps to describe the indeterminacy in the range of meanings and is applicable to modals expressing both epistemic and non-epistemic meanings. The model is illustrated in Figure 2.1 below.

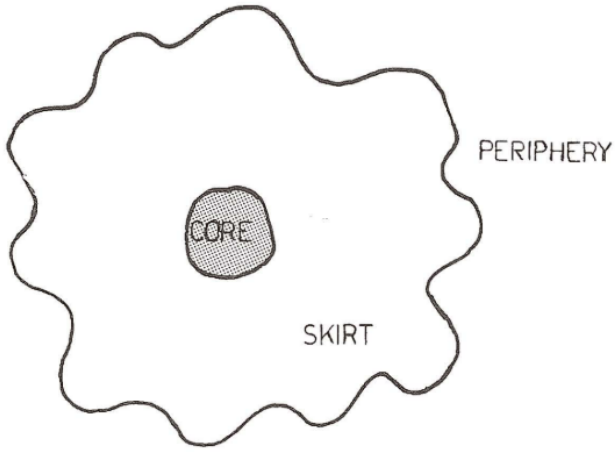


Figure 2.1 A fuzzy set model reproduced from Coates (1983, p. 12)

Core refers to the meaning that is learned first by children, whereas periphery represents the emergent uses that do not conform to stereotypical usage. Skirt includes intermediate instances on the continuum between the two ends: core and periphery. This model is not intended to categorise each instance of modal use into a fixed category but rather to depict a gradient distinction. Coates (1983) notes that although both epistemic and root meanings exhibit fuzziness, it is more prominently observable in root/non-epistemic modality (to be discussed shortly).

Regarding epistemic modality, the core/periphery continuum of meaning is related to the subjectivity/objectivity distinction. As illustrated in 2-1, a core instance conveys the speaker's subjective judgement of a proposition, demonstrating strong confidence. By contrast, a peripheral instance like 2-2 expresses an objective judgement based on logical inference, without the involvement of the speaker. To illustrate the differences,

let us take epistemic *must* as an example, introducing the characteristics identified by Coates (1983, p. 42) in its subjective and core use:

(1) Main predication refers to state or activity in the present or past.

(2) Subject is frequently inanimate.

(3) Verb is usually stative.

(4) Speaker expresses confidence in truth of utterance.

Example 2-1 exhibits all these features. Although *must* does not have a past form, *wasn't* implies that the judgement refers to a state in the past. The subject is existential *there*, which is inanimate, and the verb *be* expresses stative meanings (further discussion on semantic classifications of verbs will be made in Section 2.2.4). The speaker expresses strong confidence in their judgement, as implied by *I'm sure*. In contrast, Example 2-2 does not show the first and the last feature. Epistemic *must* is used in the main clause following a conditional clause, and thus it could be paraphrased with *will* rather than in the present tense. In addition, this example could be paraphrased as 'in the light of what is known, it is necessarily the case that …'but not 'I confidently infer that…' (Coates, 1983, p. 41), and thus it is considered peripheral.

Apart from Coates (1983), other researchers such as Lyon (1977), Huddleston and

Pullum (2002), and Collins (2009) also acknowledge the differences in subjectivity in modal usage, although they may not use the terms *core* and *periphery* to describe these distinctions. Lyon notes that the subjectivity/objectivity distinction is related to the characteristics of the evidence on which the judgement is based. When the speaker's assessment of the truth of a proposition is based on personal knowledge, it is considered subjective. Huddleston and Pullum further argue that modality can extend beyond the speaker's subjective attitudes to include two other cases, aligning with Collins's work, although Collins does not provide examples. One case involves modality used to express mathematical inference, as in 2-3, where *must* signifies purely an objective result of a mathematical problem rather than the speaker's subjective attitude. Example 2-4 illustrates another scenario where the attitude originates from the person referred to in the sentence. Collins adds that this assessment, dissociated from the speaker's attitude, is expressed externally. These instances where the attitude is not that of the speaker, are also related to textual voice, which will be discussed in Section 2.3.

2-3 If x is a prime number between 90 and 100 it *must* be 97. (example provided in Huddleston & Pullum, 2002, p. 173)

2-4 Kim thinks he *must* have written it himself. (example provided in Huddleston & Pullum, 2002, p. 173)

Although the subjective and objective use of modals appears conceptually distinct, in practice, distinguishing them reliably when examining authentic examples proves challenging (Coates, 1983; Gabrielatos, 2010; Lyon, 1977). The first three features proposed by Coates (1983) for subjective epistemic *must* are related to co-text, which Reijnierse et al. (2020) describe as the textual information immediately surrounding a linguistic item. These co-textual features are relatively easy to identify when analysing the examples. However, the last feature, 'speaker expresses confidence in the truth of the utterance', is related to the context, which Yule (1996) discusses as relating to the situation in which the text is used, such as knowledge of the social and physical world, as well as the time and place of the text's usage. In sum, it could be said that co-text is related to information within the text itself, while context concerns factors external to the text.

Lyon (1977) contends that the subjectivity/objectivity distinction is tentative because the linguistic markers (e.g., 'I'm sure' in Example 2-1) that explicitly indicate the speaker's subjective judgement do not necessarily appear in every instance. For instance, if 'I am sure' is omitted from Example 2-1, the example would then read:

2-5 There *must* be a lot more to it than that because they appear to get on very well.

2-5 could have two interpretations: one is similar to 2-1, demonstrating the speaker's confidence in the truth of the proposition. However, another interpretation could be that, based on all known facts, such as relationship status on social media, it is necessarily the case that 'there must be a lot more to it'. In this case, epistemic *must* is used objectively, based on independent evidence that is on record. The discussion above illustrates that the subjectivity/objectivity distinction seems to be unreliable since their interpretation is partly based on the context, which may not always be explicitly stated. However, Halliday and Matthiessen (2004) propose a formal approach to describe the distinction, which will be presented shortly.

Coates (1983) and Collins (2009) both observe that most instances of epistemic modality are subjective, with objective cases being rare. Coates further highlights that the fuzziness of the meanings expressed by modals is more characteristic of non-epistemic modality. Consequently, unlike epistemic modality, non-epistemic modality experiences disagreements regarding terminology and sub-categories, some of which are informed by the distinctions between subjectivity and objectivity.

Researchers such as Coates (1983) and Quirk et al. (1985) argue that non-epistemic modality should be grouped under one term as they view the subjectivity/objectivity distinction in non-epistemic modality as a gradient rather than a categorical distinction. Halliday and Matthiessen (2004), while also dividing modality into two categories, use different terminology and adopt a formal approach to distinguish between subjective and objective uses. Lyon (1977) associates the subjectivity/objectivity distinction with the notion of *source* from which the permission or obligation is derived. This source has been further explored by other researchers, with some (e.g., Huddleston & Pullum, 2002; Palmer, 2001) arguing for the necessity of a deontic/dynamic distinction in the categorisation of modality. This distinction is crucial to differentiate between a possibility or necessity that arises from moral/social norms or from general circumstances. The following paragraphs will discuss each viewpoint in turn.

Let us first look at the bipartite approach in classifying modality. Coates (1983) terms the non-epistemic use under one category, 'root modality', which is hard to characterise since it covers a range of meanings such as weak and strong permission and obligation, willingness, and intention. The fuzzy set model mentioned above (see Figure 2.1) is also applicable to root modality, representing sub-meanings through gradience. This gradience is described along two clines. One cline involves subjectivity and objectivity, which has also been discussed in relation to epistemic modality. The other concerns

the weak-to-strong continuum, which will be elaborated upon in Section 2.2.3 when discussing the strength dimension of modality. Coates (1983, p. 33) identifies four characteristics of a core and subjective example of root *must*:

(1) Subject is animate.

(2) Main verb is activity verb.

(3) Speaker is interested in getting subject to perform the action.

(4) Speaker has authority over subject.

To illustrate these features and their influence on the meaning of root modals, Examples 2-6 and 2-7 each represent a core and a periphery use, respectively.

2-6 ‘You *must* play this ten times over’, Miss Jarrova would say, pointing with relentless fingers to a jumble of crotchets and quavers. (example provided in Coates, 1983, p. 34)

2-7 Clay pots *must* have some protection from severe weather. (example provided in Coates, 1983, p. 35)

Example 2-6 exhibits all four characteristics of a core example. Root *must* is used with an animate subject, *you*, and an activity verb, *play* (further discussion on verb

classification will be in Section 2.2.4). It can be inferred that Miss Jarrova is likely a teacher, and the listener is a student, hence Miss Jarrova both has an interest in obliging the student to practice and holds authority over them. This usage of root *must* is considered subjective, indicating the speaker's involvement and can be paraphrased as ‘I order you to play this ten times over’. In contrast, Example 2-7 does not display any of the characteristics as a core example. Root *must* is used with an inanimate subject, 'clay pots', and a verb *have* that expresses the stative meaning of possessing. There is minimal speaker involvement and no implication of a hierarchical power relationship between the speaker and listener.

However, as Vincent (2015) argues, features (1) and (2) are related to co-text, which is relatively easy to identify, while features (3) and (4) are related to the context, which may not always be explicitly provided. In the case of 2-6, although ‘crotchets and quavers’ suggests that Miss Jarrova might be a teacher imposing the obligation on a student, there is also a possibility that she is merely outlining the rules of a music graded exam. Similarly, in 2-7, the context is not explicitly introduced. The example could depict an obligation imposed by a gardener on an apprentice, or it could be an instruction from a gardening manual.

Coates (1983) concludes that it is difficult to draw a binary categorical distinction

between subjective and objective uses, a view that aligns with Collins (2009) and Smith (2003). Rather than suggesting discrete sub-groups of root modality, this fuzzy set model could better capture the gradual transition from one meaning to another and the indeterminacy. Moreover, Coates (1983) proposes three types of indeterminacy, one of which is *gradience*, which is related to the continuum of meaning. The other two types are *ambiguity* and *merger*, each representing a relationship between the epistemic and root meaning.

Ambiguity refers to the examples where it is impossible to decide between the two meanings, illustrating an 'either/or' relationship, as exemplified in 2-8. *Must* can be interpreted either in the epistemic sense demonstrating confidence in the judgement that he understands what we mean or it can be described as an obligation imposed by an unstated source. It is necessary to decide between the two meanings prior to the interpretation of the example.

> 2-8 He *must* understand that we mean business. (example provided in Coates, 1983, p. 16)

By contrast, mergers indicate that the two meanings can be mutually compatible, demonstrating a 'both/and' relationship, as shown in 2-9. The example could be

interpreted as a demonstration of the speaker's confidence in the beer's quality based on its price, or as an obligation imposed on the maker to provide good beer. Both the epistemic and root reading of this example are applicable, which demonstrates the 'contextual neutralisation' (Coates, 1983, p. 17).

2-9 A: Newcastle Brown is a jolly good beer.

B: is it?

A: Well it *ought* to be at that price. (example provided in Coates, 1983, p. 17)

Quirk et al. (1985) and Biber et al. (1999) agree on this binary distinction of epistemic and non-epistemic, and they name the categories based on the involvement of human control. They classify modality into extrinsic/epistemic (possibility/ability, necessity, or prediction) and intrinsic/root (permission, obligation, or volition). A different type of bipartite approach is taken by Halliday and Matthiessen (2004), who classify modality into *modalization* and *modulation*, and their difference is related to what is exchanged.

Modalization is used in the propositions where the information is exchanged and is divided into two sub-types, probability and usuality. Modulation, on the other hand, exchanges the action between the speaker and the hearer in the proposals and is classified into obligation and inclination. In relation to other frameworks mentioned

above, probability is commonly denoted as epistemic modality while obligation is referred to as deontic modality. Apart from the two main categories mentioned above, Halliday and Matthiessen (2004) identify a further category, ability/potentiality, which is considered to be on the fringe of the modal system.

In addition, Halliday and Matthiessen (2004) introduce the concept of orientation in their exploration of modality. Orientation refers to two aspects, ‘subjective vs. objective’, and ‘explicit vs. implicit’. These two aspects can help to identify the degree of subjectivity of the judgement or obligation proposed, as well as the explicitness of the source of the modality. The researchers mentioned earlier (e.g., Coates, 1983 and Lyon, 1977) base the subjectivity/objectivity distinction on co-textual or contextual features, which Halliday and Matthiessen (2004) describe as free variants for expressing modality. However, they argue that by introducing the additional variations of 'subjective vs. objective' and 'explicit vs. implicit', expressions of modality can be formally positioned within a matrix of four characteristic combinations, as shown in Table 2.1

Table 2.1 Examples of obligation in terms of orientation proposed by Halliday and Matthiessen (2004, p. 620)

	Subjective	Objective
Explicit	I want John to go	It’s expected that John goes
Implicit	John should go	John’s supposed to go

The table illustrates that while the implicit examples, listed in the second row, do not directly state the source of the obligation, the explicit examples do. In the explicit category, 'I want John to go' demonstrates a subjective imposition by the speaker, and 'It's expected that John goes' indicates an obligation deriving objectively from another source. The use of *should* in 'John should go' implies a speaker-imposed obligation and is considered as a subjective use. Conversely, 'John's supposed to go' can serve to objectify the obligation, distancing it from the speaker's personal imposition. This matrix is applicable to both modalization and modulation.

Depraetere and Reed (2020), while adopting the terms proposed by Coates (1983) to distinguish between epistemic and root modality, further segment root modality into sub-categories, namely deontic root, non-deontic root, ability, and volition. They define root modality as the reflection of 'the speaker's judgements about factors influencing the actualization of the situation referred to in the utterance' (p. 211). The latter two sub-categories of root modality, ability and volition, are straightforward to understand and can be exemplified by '*Can* you swim' and 'I *will* support you' respectively. As for the former two sub-groups, they differ in deontic source, which is the one that is responsible for imposing the possibility or necessity. This is related to the subjectivity/objectivity distinction as well. Deontic root modality implies an authority to

impose the permission or obligation, and the authority can range from a person to social norms that are not implicitly presented, as illustrated in 2-10. In this case, John is obliged to go home, which could stem either from the speaker's explicit and subjective imposition or from a social norm. For instance, as a child aged 6, John may be expected to return home before midnight.

2-10 John *must* go home. (example provided in Depraetere & Reed, 2020, p. 211)

In regard to non-deontic root modality, it does not derive from a specific authoritative source but rather results from external circumstances. This use can be paraphrased as 'it is possible/necessary (for) …', which could be considered as objective. One example is shown in 2-11. The non-deontic root reading of it is 'Are there any external circumstances that preventing you from coming tomorrow?'.

2-11 *Can* you come tomorrow? (example provided in Depraetere & Reed, 2020, p. 211)

Researchers mentioned above all take a bipartite approach, despite different terminologies. Among them, Depraetere and Reed (2020) recognise the distinctions in

the deontic source and subdivide root modality into types to highlight these differences. In contrast, researchers such as Palmer (2001) and Collins (2009) go a step further by proposing a third category of modality, dynamic modality, based on the deontic source. This results in a three-fold categorization: epistemic, deontic, and dynamic.

Palmer (1990) points out that the binary distinction leaves no room for the modality expressing ability and volition by *can* and *will* because they are not concerned with the opinions (epistemic) or the attitudes (deontic) of the speakers. Therefore, Palmer proposes the inclusion of a third category, termed ‘dynamic modality’, which he defines as concerning ‘the ability and volition of the subject of the sentence’ (p. 7). He further divides this category into two sub-types: subject-oriented and neutral modality. Subject-oriented modality refers to the ability and volition use of *can* and *will* respectively, which is closely related to the subject of the sentence. Similarly, Huddleston and Pullum (2002) identify a usage where dynamic *must* is used to express an individual's ‘properties/disposition’ (p. 185), as illustrated in 2-12, yet this use appears to have been overlooked by Palmer. In this example, *must* conveys Ed’s internal need to 'poke his nose into other people’s business', rather than an obligation.

> 2-12 Ed’s a guy who *must* always be poking his nose into other people’s business. (Huddleston & Pullum, 2002, p. 185)

Another sub-type of dynamic modality proposed by Palmer (1990) is neutral modality, indicating what is necessary in the circumstances, as shown in 2-13. The obligation to not disclose the name of the leader seems not to be imposed by the speaker or any explicitly identified source within the sentence, but by the force of circumstances, as highlighted by Huddleston and Pullum (2002). It can be inferred that there may be unavoidable and undisclosed rules or regulations preventing the speaker from revealing the leader's name.

> 2-13 Now I lunched the day before yesterday with one of the leaders of the Labour Party whose name *must* obviously be kept quiet – I can't repeat it. (example provided in Palmer, 1990, p. 113)

Neutral modality differs from deontic modality in that deontic modality is concerned with the assignment of the obligation or permission, with a relatively clear demonstration of the source that is responsible for the assignment, whereas neutral modality does not involve a deontic source and is concerned with circumstances in general. However, Coates (1983) holds a different view, noting that this division between neutral and deontic modality omits the unity of root modality and arbitrarily cuts off the cline of root modality from strong to weak sense of obligation despite the fact that they can both be

paraphrased as ‘it is necessary for’. The indeterminacy of this distinction is also admitted by Palmer. The boundary between the two senses seems to be fuzzy and ambiguous if simply based on the source of the obligation (Gabrielatos, 2010; Huddleston & Pullum, 2002). Despite that, Palmer still considers it necessary to have the third category, dynamic modality, mostly because the ability and volition use of *can* and *will* cannot be satisfactorily categorised into the binary framework.

This present study takes a binary approach, adopting the framework of Coates (1983) and classifying *must*, *have to*, and *should* into epistemic and root meanings. This is mainly because the deontic/dynamic distinction mentioned above is mostly concerned with the ability and the volition use of *can* and *will*, but these two modals are not the focus of the present study. In addition, the deontic/dynamic mergers and ambiguous examples are found in the corpora used in the present study, and it is difficult to annotate these instances with one of the meanings (see Section 4.2.1). The distinction between the deontic and dynamic meanings requires a subjective decision from the annotator of meanings and careful interpretation of the context. One example from the learner corpus is shown in 2-14.

2-14 In the 21st century, world is more like a village and managers are all faced with opportunities and challenges in a global market, they *must* deal with economic, political and cultural difference.

In this example, the reading of *must* can be either an obligation proposed by the writer or a recognition of the circumstances that leads to the necessity of the action, and it can be interpreted as the deontic and the dynamic sense respectively. Based on the two reasons mentioned above, the present study focuses on the epistemic-root distinction. In terms of the subjectivity/objectivity distinction, as previously discussed, while it may not always be reliable due to its dependence on co-textual and contextual features which are sometimes not explicitly stated, I concur with Coates's (1983) perspective that views this distinction not as a binary categorisation but rather as a gradience. In my qualitative analysis, I will consider this perspective, referring to Halliday and Matthiessen's (2004) formal approach when necessary to differentiate between relatively subjective and objective examples. However, the subjectivity/objectivity distinction will not be the primary focus of the qualitative analysis, as mentioned in Section 1.3.

2.2.3 Dimensions and expressions of modality

In addition to the kinds of modality we mentioned above, two additional dimensions, *strength* and *degree*, are discussed in this section. This is then followed by a short introduction of the expressions of modality.

Let us look at *strength* first. It refers to the degree of commitment when expressing

epistemic and root meanings. The two types that are widely recognised by researchers (e.g., Coates, 1983; Palmer, 1990) are *possibility* (weak commitment) and *necessity* (strong commitment), and these two types have logical relations in negation. They can be paraphrased as ‘it is possible that/for’ and ‘it is necessary that/for’ respectively when expressing epistemic and root meanings. A representative pair of examples includes *may* for possibility and *must* for necessity. Huddleston and Pullum (2002), on the other hand, propose a third type, medium modality, which includes expressions such as *should*, *ought*, *likely*, etc. These modals show a medium degree of commitment compared to the possibility and necessary modals, and they show little difference between the internal and external negation. In this study, I will follow the more widely accepted terminology and consider *must*, *have to*, and *should* as necessity modals.

The types of strength discussed above are related to the comparison across different modals, and they can also be used to describe different examples of one modal, as pointed out by Coates (1983). For example, root *must* can express both strong and weak obligation based on different co-text and context, as mentioned in Section 2.2.2. In the former case, one example is 2-6, in which the speaker is highly likely to be the teacher of the hearer. It can be paraphrased as ‘it is obligatory for you to play this ten times’. As for the weak obligation, examples like 2-7 can be paraphrased as ‘It is important that clay pots are protected from severe weather.’. The analyses of these

two examples have been presented in the previous sub-section.

2-6 ‘You *must* play this ten times over’, Miss Jarrova would say, pointing with relentless fingers to a jumble of crotchets and quavers. (example provided in Coates, 1983, p. 34)

2-7 Clay pots must have some protection from severe weather. (example provided in Coates, 1983, p. 35)

This difference in strength is not related to the form, but to co-textual features or pragmatic factors, which will be discussed mainly in Section 2.2.4. Nevertheless, Coates (1983) acknowledges this continuum of strength in root modals but not epistemic modals because she characterises the latter on a continuum of more and less confidence in the commitment to the truth of a proposition. In the present study, the strength is used as a general term to describe the features of both epistemic and root modals and both the differences across modals and across different examples of one modal.

Another dimension of modality is *degree*, which is concerned with ‘the extent to which there is a clearly identifiable and separable element of modal meaning’ (Huddleston & Pullum, 2002, p. 179). It measures the notional distance between a modalised expression to its unmodalised counterpart (Gabrielatos, 2010). A low degree of

modality is mostly used in harmonic combinations with other words or phrases and thus it makes little difference when the modal is excluded from the original sentence. Lyons (1977) describes this as ‘modally harmonic’ (p. 807), denoting the pairing of a modal verb with a modal adverb that conveys the same degree of epistemic modality, as exemplified in 2-15. *May* is used with the modal adverb *possibly*, and they both denote epistemic possibility, indicating a coherent level of uncertainty regarding the truth of the proposition.

2-15 He *may* possibly have forgotten. (example provided in Lyon, 1977, p. 807)

Coates (1983) applies this concept in a broader sense, including the combination of a modal with other words or phrases (e.g., *certain* and *surely*) that indicate a similar degree of modality, and this is concluded by Huddleston and Pullum (2002, p. 179) as ‘modal harmony’. The identification of the three dimensions of modality mentioned above, kind, strength and degree, helps to examine the profile of modals systematically.

Halliday and Matthiessen (2004) introduce two other dimensions, *orientation*, which has been discussed in Section 2.2.2, and *value*. Value is classified into two categories, *median* and *outer* (Halliday & Matthiessen, 2004), and their distinction is related to the negative forms. If the negative form on the proposition is the same when it is on the

modality, then it is a median value (probable). If the negative forms shift between high (certain) and low (possible) values when they are attached to the proposition and the modality, it is called the outer value.

As you may notice, the literature mentioned above predominantly examines modal auxiliary verbs to distinguish the meanings of modality and to illustrate the dimensions. This is possibly because modal auxiliary verbs share similar syntactic features and can be concluded as one class. However, since modality is a semantic category, other modal expressions, such as modal adverbs (e.g., *possibly* and *likely*) and semi-modals (e.g., *have to* and *ought to*), are also worth investigating. The following paragraphs will first discuss the syntactic features shared by modal auxiliary verbs, and then move on to present other modal expressions.

Modal auxiliary verbs, as a sub-category of auxiliary verbs, differ from lexical verbs in exhibiting NICE (Negation, Inversion, Code, Emphasis) properties proposed by Huddleston (1976, p. 333). These properties are exemplified as follows:

(1) She *should* not come here. (take negation directly)

(2) *Should* she come? (subject–auxiliary inversion)

(3) She *should* come and I *should* too. (interpret the ellipsis based on the previous context)

(4) She *SHOULD* come. (emphasis by heavy stress)

In addition, modal auxiliary verbs have further criteria that distinguish them from primary verbs, including no person-number agreement, no non-finite forms, and no co-occurrence (Coates, 1983; Palmer, 1990). The properties and criteria discussed above are used to draw a line between central modal auxiliaries and other modal categories. *Can*, *could*, *may*, *might*, *must*, *shall*, *should*, *will*, and *would* are widely accepted as central modal auxiliaries since they meet almost all the seven criteria mentioned above. There are other modals that either only meet some of the criteria or are semantically similar to modal auxiliaries. These modals are classified by researchers under different terms. For example, Quirk et al. (1985) take a rather fine-grained classification, dividing these modals into three groups: marginal modals (e.g., *dare*, *need*, *ought to*), modal idioms (e.g., *had better*, *would rather*) and semi-auxiliaries (e.g., *have to*, *be able to*). This classification is based on the criteria the modal expressions meet on the scale of gradience between modal auxiliary verbs on the one end and full verb forms on the other. Biber et al. (1999), on the other hand, include these modals under one category, semi-modals, although they also divide them further into sub-groups of marginal auxiliaries and multi-word verbs. In addition, they point out that central modal

auxiliaries are more commonly used than semi-modals. Following Biber et al.'s framework, *must* and *should* are recognised as central modals, whereas *have to* is a semi-modal. As mentioned in Section 1.2, all three are referred to as *modals* in the present study.

Apart from modal verbs mentioned above, modality can also be conveyed by other devices. One of the initial studies that extends the investigation beyond modal verbs to include other expressions is the work by Perkins (1983), who argues that modals and their paraphrasing expressions do not have identical meanings, and it is necessary to discuss them separately. What he investigates includes modal adverbs (e.g., *clearly*, *apparently,* and *likely*), modal lexical verbs (e.g., *guess*, *demand,* and *advise*), and others (e.g., *be possible that*, *be sure to,* and *be supposed to*). In addition to these forms, he also briefly discusses how tense, if-clause and questions can be used to express modality. He admits that this is not an exhaustive list of modal expressions, but it helps to build a semantic framework for modality through an exploratory approach. A summary of modal expressions is presented by Huddleston and Pullum (2002), who list a range of linguistic expressions of modality other than modal verbs, such as lexical modals (e.g., *possible*, *certainly*, *insist*, and *permission*), clause type (e.g., imperatives and interrogative types), and subordination (I think she is here.).

2.2.4 Co-textual features of the modals

Modality, as a semantic category, is investigated beyond isolated forms. This section will explore the co-textual features of modals, with a focus on main verbs collocating with them. It will conclude with a discussion on the semantic categorisation of verbs.

Researchers have argued that meanings and interpretations of modal auxiliaries are related to their grammatical features (e.g., Coates, 1983; Collins, 2009; Hermerén, 1978). An initial attempt is made by Hermerén (1978), who investigates the influence of sentence type (e.g., declarative or interrogative) and 'basic sentence units' (p. 71; e.g., subjects and verbs) on the modal use in the Brown corpus. He believes that the meanings of modals emerge from not only the forms but also the grammatical features that co-occur with them. For instance, the epistemic use of modals is found to favour the perfect and progressive aspect, while the interrogative sentence favours the root interpretation. He also points out that the main verbs co-occur with the modal may affect the meaning of that modal, as exemplified by *may*. When *may* is used with an activity verb such as *open*, as in 2-16, it is most likely to be interpreted as permission. However, if it is used with stative verbs like *be*, as in 2-17, *may* mostly expresses the possibility. Similarly, Perkins (1983) acknowledges that syntax motivates the semantics of the modals, but he only covers the syntactic aspects that he recognises as obvious motivations with few explicit justifications. Although Hermerén and Perkins have introduced some new perspectives on modal verbs, their studies are primarily

descriptive and qualitative, lacking statistical information.

2-16 You *may* open the door now.

2-17 You *may* be late.

With the development of corpus linguistics, quantitative analysis is conducted and the coding of co-textual features is more systematic. One example is Coates's (1983) work, investigating 200 instances for each modal auxiliary in both the written and spoken material in the Lancaster corpus (now as the Lancaster-Oslo/Bergen Corpus; Johansson, Leech, & Goodluck, 1978) and the Survey of English usage (University College London, n.d.). The examination focuses on the frequency and meaning distribution, as well as syntactic features, such as voice, main verbs, subject, and negation, which are quantified by percentages. Compared to the qualitative analysis mentioned above, this quantification helps to clarify the association between the modal and the syntactic features.

In regard to *must*, *have to*, and *should*, Coates (1983) finds that perfect and progressive aspect, existential subject, stative verb and inanimate subject are associated with epistemic *must*, whereas root *must* co-occurs frequently with negation, passive voice, agentive verb, second and first-person subject. *Have to* is briefly

discussed in comparison to *must*, emphasising the difference in subjectivity and negation. In the case of *should*, its epistemic use is found to be associated with non-agentive verbs, passive voice, and negation. Root *should* is discussed in terms of tense, aspect, and negation. Some of these co-textual features will be explored further in Section 2.2.5.

The association between the meaning of modals and the subject and main verbs co-occurring with them is also observed by Biber et al. (1999). They examine the Longman Spoken and Written English Corpus (the LSWE Corpus), which consists of four core registers (conversation, fiction, news, and academic prose) and two supplement registers (general prose and non-conversational speech), each containing about five million words. It is reported that extrinsic modals are frequently used with non-human subjects and stative verbs, whereas intrinsic modals are used more often with subjects of human beings and dynamic verbs. In addition, the percentages of marked aspect and voice used with the main verbs that collocate with the modals are also explored. For instance, *must* and *should* show higher percentages to be used with passives in the academic prose compared to other modals and other genres. By contrast, they are used less frequently with the perfect or progressive aspect compared to other modals. However, Biber et al. focus on the difference in registers and thus modal meanings are not distinguished. Both Coates and Biber et al. acknowledge that these associations

between syntactic features and meanings of modals are not absolute, but these findings still support arguments previously made by Hermerén (1978) and Perkins (1983).

Both studies mentioned above seem to focus mainly on a binary distinction in the semantics of the main verbs used with the modals. It is worth noting that Coates (1983) appears not to provide definitions for stative/dynamic or agentive/non-agentive verbs but employs these terms to analyse the examples. I will introduce these binary classifications of verbs here. A more detailed discussion on verb classifications will follow at the end of this section.

In terms of the stative/dynamic distinction, Biber et al. (1999) categorise them as two broad semantic domains of verbs, defining stative verbs as those that denote 'stable states of affairs' (p. 456), such as perception, cognition, and emotion. Conversely, dynamic verbs are described as referring to 'events, acts, or processes with an inherent implication of completion' (Biber et al., 1999, p. 456), such as verbs that describe physical activities. This binary distinction is similarly proposed by Huddleston and Pullum (2002) and Quirk et al. (1985), who explore this classification in relation to situation types. Researchers mentioned above all acknowledge that this distinction is not clear-cut. For instance, the verb *have* is used in both examples below. It expresses

the stative meaning of possession in 2-18, while in 2-19 it can be interpreted as *eat*, which conveys the dynamic meanings associated with the action of eating. This demonstrates that the meanings of a verb can shift from one category to another, and classifications should not be seen as definitive. Instead, they ought to be interpreted based on co-text and context in authentic examples.

2-18 The chair has beautiful carved legs quite frequently. (example provided in Quirk et al., 1985, p. 206)

2-19 We have dinner at Maxim's quite frequently. (example provided in Quirk et al., 1985, p. 206)

In terms of the agentive/non-agentive distinction of verbs, they relate not to the verb's meaning but to the states of the sentence's subject. *Be* in ‘John was thinking.’ (Quirk et al., 1985, p. 207) is considered agentive as it implies that the subject acts as an agent, or a doer of the action concerned. A doer is typically a human, who deliberately initiates the action, and a dynamic meaning of the verb often implies this active doer (Quirk et al., 1985). In contrast, a non-agentive verb, as in ‘The wind is blowing hard.’ (Quirk et al., 1985, p. 207), does not involve an active agent, and *be* expresses the state of the wind.

In summary, the distinction between agentive and non-agentive verbs is related to the emphasis on the agent that takes the action, whereas the difference between stative verbs and dynamic verbs, is associated with the states of the action. Both the stative/dynamic and agentive/non-agentive distinctions are not clear-cut and each verb requires discussion based on specific examples. The agentive/non-agentive distinction is not employed in the subsequent analysis as it is related less to the meanings of verbs and more to the status of the sentence's subject. By contrast, the stative/dynamic distinction will continue to be referenced in the analysis. I adopt the terminology used by previous researchers such as Coates (1983) and Biber et al. (1999) to distinguish between stative and dynamic verbs. However, this does not necessarily imply that if a verb is described as stative, it exclusively belongs to this category. This only means that in that specific example, it is interpreted as stative. In other contexts, the same verb could be described as dynamic. The meaning of verbs depends on the interpretation of each example.

Some specific verbs are also discussed in passing. For instance, Coates observes that root *must* is used with communication verbs and the first-person subject (e.g., 'I *must* say/admit/confess') to express weak obligation, as illustrated in 2-20 below.

2-20 Dai had some quite interesting ideas which surprised me rather. I *must* admit. (example provided in Coates, 1983, p. 34)

In this example, ‘I must admit’ follows a statement expressing surprise, indicating the speaker's initial hesitation to acknowledge that Dai's ideas were interesting, which is later overcome. This combination demonstrates that the obligation is laid on the speaker to acknowledge the statement involved and is fulfilled (Palmer, 1990). De Haan (2012) identifies this use as hedging, which may imply possible previous scepticism or a different expectation regarding how interesting Dai’s ideas were.

Root *have to* can substitute for *must* in this combination with no shift of meaning (Collins, 2009), as exemplified in 2-21, but this is less frequently used. Palmer (1990) adds that this combination, when used with second-person subjects, is to ask the hearer to ‘behave in a similar fashion’ (p. 74), such as ‘you must understand/admit’. A more systematic examination of the semantics of the main verbs used with the modals is necessary, given that the previous discussion did not cover a wide range of uses.

2-21 Although I would love to I have to yes I *have to* confess an often irking thought of am I really really two pounds less than Kate Hamilton (example provided in Collins, 2009, p. 62)

More recently, Gabrielatos and Sarmento (2006) and Collins (2009) also investigate syntactic features of the modals, but the aspects they cover are less comprehensive than Coates (1983). Gabrielatos and Sarmento examine central modals in an aviation corpus consisting of manuals and make comparisons to a reference corpus and across different types of manuals. Although their main focus is on frequency distribution, they also report the distribution of modal verb phrase structures (e.g., 'modal + infinitive', 'modal + be + infinitive' and 'modal + have + past participle') and emphasise the correlation between the voice and the use of modals. In addition, they stress the need to investigate the verb collocates of the modals. Collins, on the other hand, discusses a limited aspect of syntactic features such as negation and person of the subject as his focus is on the regional and stylistic variation of modals.

To date, studies have taken a finer-grained view to investigate the co-textual features of the modals, with more systematic approaches and a focus specifically on individual modals. Deshors (2016) uses a behavioural profile approach to compare the use of *may* and *can* in three varieties of English (English as the native language, English as the interlanguage, and French as the native language). The analysis involves the annotation of three types of co-textual features, namely semantic (e.g., degree of speaker presence and types of modalised lexical verb), syntactic (e.g., negation and sentence type), and morphological (e.g., voice and aspect) variables. The exploration

of lexical verbs used with the modals includes three aspects, the verb use (metaphorically and literally), verb type in terms of time (accomplishment, achievement, process and state), and the semantics of these verbs, which will be discussed in detail shortly.

Similarly, Furmaniak (2020) also manually annotates variables such as semantics and pragmatics, conducting an analysis on the (con)textual properties of *must*, *have to* and *shall* within the framework of integrative grammar. Although Furmaniak's focus is on the use of the modals across discourse modes, he still reveals the relation between modals and the co-occurring main verbs. For example, root *have to* is identified to be used with verbs (e.g., *say*) performatively to make the statement tentative, and it is frequently used to describe contextual information so that the main argument can be led to.

While the two studies mentioned above effectively explored a broad range of co-textual factors associated with modals, the process of manually annotating such information is both time-consuming and requires subjective judgements, which need to be verified by other annotators. Recent works on construction grammar, on the other hand, focus on fewer aspects of co-textual features but take an approach that does not require manual annotation. These studies investigate the co-textual features and collocational

profiles of modals and their influence on modality meaning. One study is conducted by Cappelle et al. (2019), who explore the semantic and pragmatic similarity across four necessity modals, *must*, *have to*, *should*, and *need to* in the British National Corpus (BNC, 2001) through the investigation of n-grams up to five words. It is found that *should* shares the least similarity to the other three modals, among which *must* is the closest one that relates to it. Compared to *must*, *have to* is more related to *need to*. In terms of specific co-textual features, *should* shows a strong association with passive voice. *Must* and *should* share several n-grams when used in the perfect aspect, yet they differ in the meanings they convey. Those instances of *must* mostly express epistemic necessity, which is also observed by Coates (1983). By contrast, *should* is used with the perfect aspect to express a reproach.

Hilpert and Flach (2021) examine BNC as well, but take a different approach, investigating the collocational profile of *must* and *have to* by looking at second-order collocates. It is found that the meaning of *must* is closely related to its collocational profile. Verb collocates of *must* expressing obligation are mostly related to instructions or legal terms. In addition, Hilpert (2016) explores how collocational profiles could be used to discover the diachronic changing of modal meanings in the Corpus of Historical American English (COHA; Davies, 2010) using a different approach. He investigates how meanings of *may* have changed from the 1800s to the 1990s by looking at lexical

verbs in the infinitive used with it and constructing a semantic vector space model to visualise the semantic distribution of these verbs. This approach is used in the present study, and a further explanation and justification of the selection of this approach will be presented in Section 4.2.2.

The studies mentioned above provide a new perspective in investigating profiles of modals. The focus extends from meaning to verb collocates and other co-textual features, which helps us to have a more comprehensive understating of how modals are used and related to co-textual features. The following paragraphs will discuss the core co-textual feature that is investigated in the present study, the main verbs collocating with the modals, since these verbs play an important part in modal use (see Section 1.2) and have not been systematically investigated, as discussed above. First, a working definition of the collocation of a modal will be given, and it is followed by a more detailed description of how verbs can be classified semantically.

Collocation, as a widely investigated notion, is interpreted differently based on the approaches taken. Firth (1957) proposes that ‘You shall know a word by the company it keeps.’ (p. 179), arguing that the meaning of a word should be characterised by the frequent co-occurring words. This is the distributional approach, also called the Firthian approach, and it is frequently applied in corpus-related research. For instance, Sinclair

(1991) follows Firth's definition and identifies collocation as 'the occurrence of two or more words within a short space of each other in a text' (p. 170). Bhalla and Klimcikova (2019) conclude with two additional approaches, the psychological and the phraseological approach. The former considers collocation as word associations that are mentally stored and can thus be processed faster (Wray, 2012). This approach is mostly supported by psycholinguistic research using, for example, controlled and designed tasks. As for the phraseological perspective on collocation, it focuses on the predictability and the compositionality of the word combinations and attempts to divide them into separate groups in different terms. However, the criteria to identify these groups is hard to operationalise to avoid subjectivity, as pointed out by Schmitt (2010). Thus, it is extremely difficult to give rigorous definitions of collocation using this approach.

Despite the differences, the three approaches interpreting collocation overlap to some extent in that the mental storage of word combinations can be supported by evidence of the high frequency of occurrence (the distributional approach) and its non-compositionality (the phraseology approach), as noted by Durrant and Mathews-Aydnl (2011). The present study takes the distributional approach and adopts Sinclair's (1991) notion of collocation, defining it as word combinations that co-occur with each other in a short span. Specifically, this study examines main verbs collocating with *must*, *have*

to, and *should* in a short span of words to the right. Given that the main verbs are extracted manually in the present study, their proximity to the target modal varies; some are directly adjacent, while others may be located further away if used in a coordinate clause, for instance. Consequently, the span of words including these collocates is not strictly defined. A detailed explanation of the operationalisation and extraction of verb collocates of these modals will be provided in Section 4.2.2.

Let us now move on to discuss the semantics of these verbs. Deshors (2016) does not use a pre-existing classification, dividing the semantics of verbs into eight tags: abstract, action general, mental/cognitive/emotional, action motion, communication, copula, action transformation, and perception. Similarly, Furmaniak (2020) manually annotates the verbs that co-occur with the modals using a smaller set of labels, including material, behavioural, verbal, mental, and relational types. The two studies mentioned above take a bottom-up approach, classifying the semantics of verbs manually. Its advantage is that the framework is specific to the dataset. However, similar to the annotation of other co-textual features, it constantly requires subjective decisions and is time-consuming.

By contrast, there are studies using a pre-existing classification of verbs, such as Zou (2018). These classifications of verbs differ among studies, ranging from classification

based on corpus evidence (e.g., Biber et al., 1999) to framework related to the context (e.g., Halliday & Matthiessen, 2004). In addition to the broad semantic domains of stative/dynamic previously discussed, Biber et al. (1999) further classify single lexical verbs into seven semantic domains, including activity verbs, communication verbs, mental verbs, verbs of facilitation or causation, verbs of simple occurrence, verbs of existence or relationship, and aspectual verbs. The descriptions and examples of the seven categories are summarised in Table 2.2 below, reproduced in a reader-friendly form. The classification is based on the core meaning (the first meaning people think of) or the typical use (the most frequently used meaning) of the verb. Similar to the distinction between stative and dynamic verbs, we need to bear in mind that some verbs can belong to more than one semantic domain and thus their classifications depend on the interpretation of authentic examples. For example, while *say* is classified as a communicative verb in the table, and typically dynamic because it involves the action of speaking, it can also convey a relatively stative meaning, as in ‘The law says that everyone must wear a helmet.’.

Table 2.2 Semantic domains of single-word lexical verbs categorised by Biber et al. (1999, pp. 361-364)

Category	Description	Examples
Activity verbs	Denote actions and events that could be associated with choice	bring, buy, carry, come, give, go, leave, move, open, run, take, work, etc.
Communication verbs	A special subcategory of activity verbs that involve communication activities (speaking and writing)	ask, announce, call discuss, explain, say, shout, speak, state, suggest, talk, tell, write, etc.
Mental verbs	Denote activities and states experienced by humans that do not involve physical action and do not necessarily entail volition, including verbs expressing cognitive meanings, emotional meanings (attitudes or desire), perception, and receipt of communication	think, know love, want, see, taste, read, hear, etc.
Causative verbs	Indicate that some person or inanimate entity brings about a new state of affairs	allow, cause, enable, force, help, let, require, permit, etc.
Occurrence verbs	Report events (typically physical events) that occur apart from any volitional activity	become, change, happen, develop, grow, increase, occur, etc.
Existence verbs	Report a state of existence or a relationship between entities	be, seem, appear, contain, include, involve, represent, etc.
Aspectual verbs	Characterise the stage of progress of some other event or activity	begin, continue, finish, keep, start, stop, etc.

Biber et al.'s (1999) classification is form-based, whereas Halliday and Matthiessen (2004) emphasise the multifunctional features of verbs that depend on the context and divide them into six types (three main types and three intermediate types), as shown in Figure 2.2.

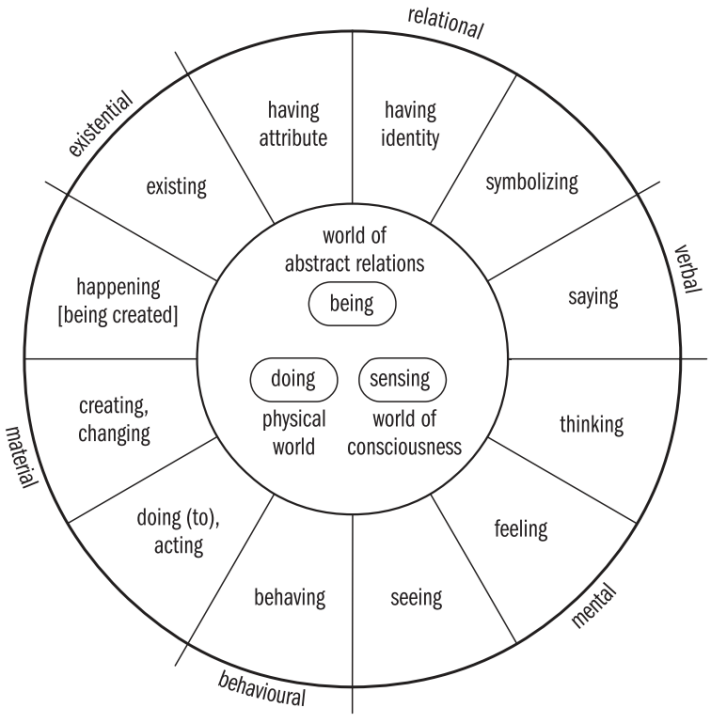


Figure 2.2 Types of verbs categorised by Halliday and Matthiessen (2004, p. 172)

The classification is related to the experience line of meanings and termed as 'types of process'. The three main types are mental, material and relational processes. The former two processes are related to the inner and outer experience respectively. The relational process refers to the verbs of identifying and classifying. There are also verbs

that are located at the boundaries of the three types and are identified as intermediate types, including the behavioural, verbal, and existential processes. These types demonstrate the intermediate meanings of the pairs of processes next to them in the figure.

Another useful tool to distinguish the semantics of verbs is WordNet (Fellbaum, 1998), which provides a specific network where verbs have a pre-decided list of different senses with their semantic relations recorded. In other words, it can be used to retrieve not only the meaning of the verbs but also the semantic relations between the verbs. For example, Perek (2014) employs WordNet to annotate the senses of the verbs used as collexemes in the conative constructions, uncovering the relation between the semantics of these verbs and the constructional meaning. The drawback of using WordNet is that it does not straightforwardly provide general categories that are as fit for the purpose of the present study as those provided by Biber et al. (1999) and Halliday and Matthiessen (2004). In addition, some distinctions of the senses are overly fine-grained, which is hard to distinguish when annotating examples (Perek, 2014). A further discussion on the approach to explore the semantics of verbs will be presented in Section 4.2.2.

2.2.5 *Must*, *have to*, and *should*

So far, we have discussed English modality in general. This section will discuss *must*, *have to*, and *should* in detail, which are the focus of the present study. As mentioned in Section 1.2, these three modals are selected because epistemic necessity and obligation modals are generally under-investigated in modality studies in academic writing (see Sections 2.3.2 and 2.4.2). In addition, among the epistemic necessity and obligation modals, these three modals have a relatively high frequency, as will be shown in Section 3.3, which could help to provide a comprehensive and relatively representative picture of how Chinese students use these modals.

The literature mentioned in the previous sections covers the profiles of *must*, *have to*, and *should* in detail. I will synthesise the findings of those using corpora in the following paragraphs since the examination of the modals in a large corpus is more likely to provide a comprehensive picture of their features compared to those using invented examples and introspection. In what follows, I will introduce each modal separately in terms of meaning distribution, strength of modals, degree of subjectivity, and some of the co-textual features, and make comparisons across the three modals.

Let us look at *must* first. In terms of meaning distribution, epistemic *must* is more frequently used than its root sense in conversation, whereas it shows an opposite trend in academic prose (Biber et al., 1999). Epistemic *must* refers to the speakers'

confidence in the truth of the proposition (Coates, 1983). As mentioned before, epistemic modals are rarely used objectively, and this is also the case for epistemic *must*. When used objectively, it expresses logical certainty and confidence in making the judgement based on known facts, as previously discussed in 2-2. The degree of modality is low because it makes little difference compared to its non-modalised counterpart. The more common use is the subjective epistemic *must*, as in 2-22. The speaker's involvement is implied by the use of the pronoun *you*, and the judgement regarding the truth of the proposition is based on the speaker's knowledge of the listeners' previous working experience over the years.

> 2-2 Shall we then say with G.F Stout that 'desire and aversion, endeavour to and endeavour from, are modes of attention'? Certainly if there is endeavour to x, there *must* be attention to x. (example provided in Coates, 1983, p. 42)
>
> 2-22 With all the bits of work you've done over the years, your CV *must* be pretty full? (example provided in Collins, 2009, p. 39)

Epistemic *must* is sometimes combined with hedges (e.g., 'I think' and 'I suppose') and harmonic combinations (e.g., 'I am sure' and *surely*). The former combination pragmatically weakens its modal strength and is more frequently used than the latter (Collins, 2009). It can be concluded that a semantically strong modal like *must* does not necessarily express a strong sense of obligation pragmatically.

Coates (1983) lists several co-textual features that are strongly correlated with the occurrence of epistemic *must*, especially the subjective ones. These features include the perfect aspect, the progressive aspect, the existential subjective, stative verbs, and inanimate subjects, which have been mentioned in Section 2.2.4. Epistemic *must* is rarely used to refer to activities in the future because 'certainty is an inappropriate feeling to have about the future' (Coates, 1983, p. 45). Most of the instances describe the activities in the present or the past. In addition, epistemic *must* does not have a negated form, and *cannot* is used to express its negation.

With regard to root *must*, according to Palmer (1990) and Collins (2009), its default source of obligation is the speaker. If the speaker's involvement is explicitly presented, or the authority of the speaker over the subject is implied, this will result in a relatively strong sense of obligation. Apart from the presence of speakers, the person of the subject is also an influential factor in the modal strength, as noted by Coates (1983). Generally, examples with a second-person subject express stronger obligation than those with a first- and third-person subject, comparing 2-6 and 2-23.

> 2-6 'You *must* play this ten times over', Miss Jarrova would say, pointing with relentless fingers to a jumble of crotchets and quavers. (example provided in Coates, 1983, p. 34)

2-23 He’s going on the 7.40 tomorrow morning and everything *must* be packed tonight. (example provided in Coates, 1983, p. 35)

In 2-6, the speaker is the one who imposes the obligation, whereas in 2-23, the deontic source is not explicitly presented. However, Collins (2009) highlights that this correlation is merely a tendency and there are exceptions. In addition, the use of passives may weaken the strength of the obligation due to the unspecified subject. For example, if we paraphrase 2-23 into the active voice as in 2-24, the strength of obligation seems to be stronger because it explicitly identifies the person to whom the obligation is assigned.

2-24 He’s going on the 7.40 tomorrow morning and he *must* pack everything tonight.

Have to is usually examined with *must* in that they are used similarly but with notable differences. Epistemic *have to* is rarely used in this sense, and its frequency is lower than that of epistemic *must* (Palmer, 1990). One of the examples is 2-25 below, in which the speaker’s judgement about what is happening to the balloons is grounded in observable phenomena and established knowledge about how molecules behave when heated (i.e., they speed up and move apart, causing the balloon to expand).

Have to is used to suggest the logical outcome based on the given evidence.

> 2-25 And so, the molecules are speeding up, they're getting further and further apart, and taking up more space inside the balloon, so the balloon goes back to its former size and shape. So that *has to* be what's happening to the balloons, that are inside this container here. (example provided in Collins, 2009, p. 63)

Both Coates (1983) and Collins (2009) highlight the connection of this use to the American teenage culture. Coates argues that in British English, epistemic necessity is mostly expressed by epistemic *must*. Collins observes that epistemic *have to* has gained acceptance in both Australian and British English.

Epistemic *have to* seems to be used objectively compared to epistemic *must*, but this distinction is not as clear as their root counterparts. Root *have to* is mostly used without the involvement of a speaker and expresses an objective sense of obligation, as in 2-26.

> 2-26 There is already a great imbalance between what a student *has to* pay if he's in lodgings and what he *has to* pay /…/ if he is in a hall of residence (example provided in Coates, 1983, p. 55)

The obligation in this example does not arise from the speaker's authority or subjective viewpoint but from unspecified source that is impersonal such as university policies and housing market conditions. This is suggested by the use of conditional clauses, indicating that the payment requirements are linked to specific living situations. Coates (1983) states that it is interchangeable with root *must* when the source of obligation is external, but root *must* is rarely used objectively.

Coates (1983) also proposes that root *have to* can be used with habitual activities whereas root *must* cannot, as shown in 2-27. This example suggests that the obligation to wake up at a specific time is imposed by external factors, such as work or school schedules. The routine nature of the action is emphasised by 'every day'. Conversely, replacing *have to* with *must* would imply a personal compulsion, perceived by the speaker, such as a personal goal or health regimen, rather than a routine dictated by external factors.

2-27 I *have to* get up at 7 a.m. every day. (example provided in Coates, 1983, p. 54)

Collins (2009) disagrees with this because he found a counterexample, 2-28, where

root *must* is also used to express habitual actions, as evidenced by *always*. This emphasises that renewing brake shoes in sets of four is a regular, mandatory practice. Collins suggests that this distinction in conveying habitual meanings is not as salient as the subjective-objective distinction between root *have to* and *must*.

2-28 Brake shoes *must* always be renewed in sets of four. (example provided in Collins, 2009, p. 64)

In general, *have to* differs from *must* in terms of negation. While *must* negates the main predication, *have to* negates the modal predication. To put it differently, *must* in negation can be paraphrased as 'it is necessary that … not …', whereas the paraphrase for *have to* would be 'it is not necessary that ... '. In addition, Coates (1983) considers *have to* to be the past tense suppletive form of *must*, but Collins (2009) retains doubts because it is hard to prove suppletion. Palmer (1990), on the other hand, notes that there is no need for *must* to have a past form.

Having discussed the usage of *must* and *have to*, I will now move on to discuss the last modal in question, *should*, which is considered by Leech (2004) as the weaker equivalent of *must*. Epistemic *should* is considered as a tentative form of epistemic necessity by Coates (1983) and Palmer (1990). It expresses a weaker sense of

assumption and confidence compared to *must*, and commonly refers to the future, as shown in 2-29. The use of *should* implies that while the meeting is planned or expected, circumstances could potentially change, and 'later on this afternoon' specifically indicates that the event is scheduled for the future.

2-29 You *should* be meeting those later on this afternoon. (example provided in Palmer, 1990, p. 59)

As for the root use, it can be subjective or objective, but its subjective use seems to be more dominant (Collins, 2009). Similar to root *must*, root *should* also shows weak-to-strong gradience of strength in obligation. What differs is that when using root *should*, the speaker does not expect the hearer to obey the obligation. In other words, it allows for 'non-actualization' (Collins, 2009, p. 45), and this makes its strength weaker than *must* in general. Coates (1983) further notes that the strongest and weakest senses of root *should*, when used subjectively, covey suggestions or advice, as illustrated by 2-30 and 2-31 respectively.

2-30 I think husbands really *should* be made to do the moving actually (example provided in Coates, 1985, p. 59)

2-31 Well perhaps I *should* choose a London map if I'm going to look at Clapham. (example provided in Coates, 1985, p. 59)

In 2-30, ‘I think’ indicates that the obligation is a subjective one, imposed by the speaker on the husbands. The speaker suggests a moral duty for husbands to take a more active role in the moving task, although this is not a demanded action. Similarly, in 2-31, root *should* is used subjectively, as indicated by the pronoun *I*, and its weak strength is implied by *perhaps*. This sentence could be paraphrased as ‘it would be a good idea for me to choose a London map …’.

Root *should* can also be used objectively to describe appropriate procedures (Coates, 1983), as demonstrated in 2-32. The speaker starts by mentioning their firm insistence on using a less formal title, which sets up a contrast to what is universally acknowledged as more appropriate. In addition, the use of *really* emphasises the correctness of the action, indicating that this recommendation is based on the general conventions or norms rather than personal opinions.

> 2-32 I just insisted very firmly on calling her Miss Tillman but one *should* really call her President. (example provided in Coates, 1985, p. 59)

Although Coates (1983) uses *suggestion* to describe solely the subjective use of root *should*, other researchers, such as Collins (2009), employ *suggestion* to describe weak

obligation, irrespective of whether it is root *must* or root *should*. Palmer (1990), on the other hand, associates *suggestion* with the use of possibility modals that express permission, such as *may* and *can*. Despite the varied uses of the term *suggestion*, a common feature among the researchers appears to be that *suggestion* is used to describe a weak sense of obligation, whether it is expressed by root *must* and *should*, or other modals. Thus, in the present study, *suggestion* is used interchangeably to describe weak obligation. It is also important to note that other researchers may use the term differently, as will be illustrated by the empirical studies in Section 2.4.2, among other examples. In the subsequent sections of the literature review, I will refer to their terms as they present them, but for my own analysis, *suggestion* will consistently refer to a weak sense of obligation as defined here.

Similar to *must*, *should* has no past form. To describe the suggestion in the past, root *should* is used in the perfect aspect. Since the past is known, this use sometimes implies that the subject did not take the suggested action in the past and the suggestion is in a hypothetical sense, as in 2-33 below. ‘They should have’ introduces a conditional outcome that did not occur but was possible if the suggested action had been taken, and *perhaps* emphasises the uncertainty of this potential outcome. Collins (2009) notes that this use implies a sense of criticism.

2-33 They *should*'ve left it [=Belfast] completely alone and they'd have got southern Ireland perhaps back into the fold. (example provided in Coates, 1983, p. 62)

The hypothetical use of root *should* is reported by Coates (1983) to be a more frequent use than genuinely perfective use, which is exemplified in 2-34. 'By the age of sixteen' indicates that the action of completing general reading is expected to have been achieved by this specific age. The use of 'should have done' underscores the perfective aspect, emphasising that the reading should not just be underway but fully completed.

2-34 By the age of sixteen, anybody who is going to be an academic *should* have done their general reading. (example provided in Coates, 1983, p. 59)

In addition to the two meanings mentioned above, *should* can also be used as a quasi-subjunctive to express the low degree of modality. *Should* in these low-degree uses expresses negligible modal meaning compared to their unmodalised counterparts. Huddleston and Pullum (2002) list five types of this use: mandative, adversative, purposive, emotive, and conditional. In relation to syntactic features, Collins (2009) adds that conditional *should* is more commonly used with subject-auxiliary inversion than with if-clauses, as illustrated by 2-35 and 2-36 respectively. In 2-35, *should* is used in the implicit conditional clause with subject-auxiliary inversion, and it expresses a

slight doubt compared to the non-modalised sentence containing if-clause.

2-35 I can only hope that I will be able to provide the support, as selflessly as you both have done, to you, *should* you ever require it. (example provided in Collins, 2009, p. 51)

2-36 If you *should* experience any difficulty, please let me know. (example provided in Huddleston & Pullum, 2002, p. 187)

Another minor use of *should* is to supply a first-person variant for hypothetical *would*. As shown in 2-37 below, *should* is used to express genuine hypothetical meaning implied by the use of conditional clause and 'it is possible that'. A more common use is illustrated in 2-38, in which *should* is used to express polite tentativeness pragmatically rather than conveying hypothetical meanings. 'As much as I can' serves to clarify that the speaker's intent to help is sincere, while also acknowledging potential limitations. This use is restricted to first-person subjects (Coates, 1983; Collins, 2009). These two uses of *should* are excluded in the analysis, which will be justified in Section 4.2.1.

2-37 And it is possible that I *should* have met him through Rober Graves / … / if I had not been introduced to him by Sidney. (example provided in Coates, 1983, p. 221)

2-38 I *should* like to help you as much as I can when you come, but […] (example provided in Collins, 2009, p. 52)

So far, we have discussed the three modals separately. Figure 2.3 displays the meaning distribution of the modals, as reproduced from Biber et al.'s (1999) investigation.

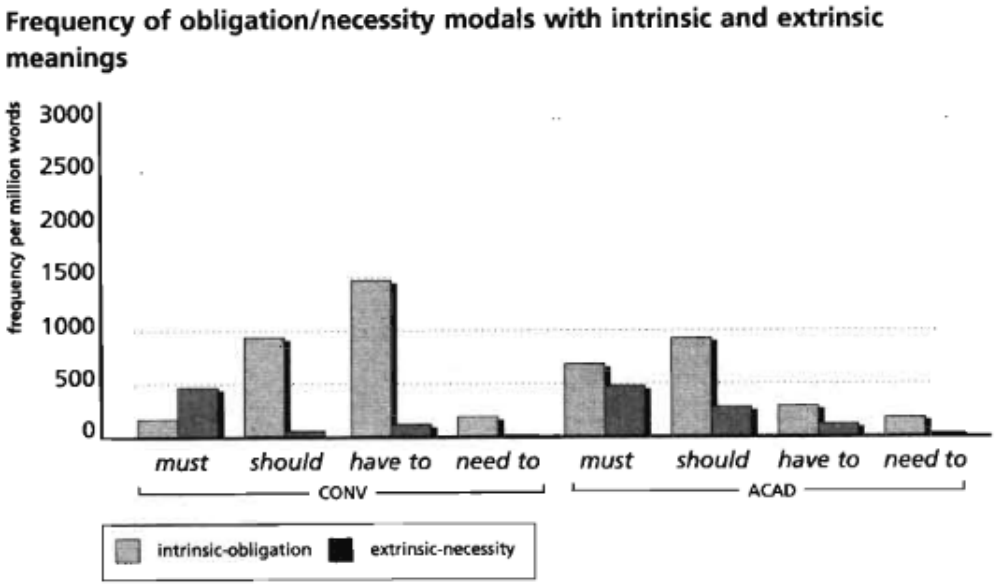


Figure 2.3 Frequency of obligation/necessity modals with intrinsic and extrinsic meanings (Biber et al., 1999, p. 494)

As shown in the figure, the three modals show a similar pattern in the academic prose, with more use in the intrinsic/root sense than the extrinsic/epistemic sense. The variation in meaning frequency is the largest in the use of *should*, followed by *have to* and *must*. However, the pattern is different in conversation. *Should* and *have to* are

used more often in their intrinsic sense, whereas *must* shows an opposite pattern. The difference in the two meanings is the largest in the use of *have to*, followed by *should* and *must*.

In sum, *must*, *have to*, and *should* express two main meanings, epistemic and root sense. *Must* shows the strongest degree of commitment in terms of judgement to the truth of a proposition as well as the actualisation of the suggested action, and it is followed by *have to*, and *should*. However, this simplified identification of strength does not always work when considering pragmatic factors and syntactic features. Depending on the co-text and the context, sometimes *must* shows a weaker strength than *should*. Thus, it is important to, on the one hand, explore the three modals quantitatively to identify recurring patterns and, on the other hand, take a fine-grained view by conducting a qualitative analysis (see Section 3.4 for the overall research design of the study).

2.3 Modality in academic writing

This section will zoom in on what is known about the modality used in academic writing. It will start with a critical evaluation of how the definitions of modality mentioned in Section 2.2.1 can be applied to the context of academic writing with a focus on textual voice. It will then be followed by a discussion on modality studies in academic writing.

2.3.1 Working definition of modality in academic writing

Despite the extensive interpretations pertaining to modals, as Section 2.2.1 indicated, to what extent these understandings can apply to academic writing remains in doubt. Among the researchers who examine modality in general, there appears to be a predominant reliance on spoken material as the primary source of data. Although written texts are also explored, academic writing is included only to a limited extent. Coates (1983) examines fifteen genre categories in the written data (e.g., newspaper editorial, science fiction, and romance and love stories), with no inclusion of academic writing. Palmer (1990) uses the data from the Survey of English Usage (University College London, n.d.), which contains both written and spoken data. However, he does not specifically distinguish the results between the two sets of data because the data is used 'for heuristic and exemplificatory purposes only' (Palmer, 1990, p. 29). The written data in Collins (2009) is from British and Australian components of the International Corpus of English (ICE-GB and ICE-AUS; Greenbaum & Nelson, 1996), and a specially compiled corpus of American English (C-US). These corpora include a small portion of academic writing as well as other registers. What is worth mentioning is that the word number of the spoken data in Collins's study is nearly 1.5 times more than the written data. The studies mentioned above have a shared feature that the data investigated are mostly spoken data, which could potentially lead to a biased definition of modality that is more relevant to its use in spoken data.

Other studies have attempted to investigate modality specifically in the context of academic writing, but few of them give a satisfactory definition of modality. Ewer (1979) includes the notion of the writer when describing the functions of modality in professional articles or papers, stating that 'the modals express the writer/speaker's personal degree of confidence in the truth, likelihood or importance of the propositions they mediate' (p. 6). Simpson (1990) defines modality in literary criticism discourse as a concept concerned with 'degrees of commitment and detachment from propositions', and he highlights the relationship between the modality and the pragmatic aspects of politeness and how information is presented. One oversight shared by these two studies is textual voice identification, that is, whose voice is presented through modality.

Early works such as Sinclair (1988) and Tadros (1993) lay the groundwork for understanding how academic writers navigate the complexities of presenting their ideas while appropriately acknowledging the contributions of others. They divide textual voice into two types, *averral* (the writer's voice) and *attribution* (the voice of a third party, such as antecedent authors). From now on, I will refer to current voices as *writers* and to antecedent voices as *authors*. Subsequent studies, such as those by Groom (2000), explore the relation of the identification of textual voices to two other perspectives, the evaluation of intertextual voices (reporting clauses) and the

positioning of textual voices with a focus on citation practices. A more recent contribution is Abdesslem (2020), who builds upon the dichotomy of textual voice, extending it into clines and relating it to responsibility and writer-author presence.

Sinclair (1988) further proposes a three-fold distinction within attribution based on the degree of reliability to the statements made, namely primary, secondary and tertiary attributions. Primary attribution involves direct quotations or references to other researchers' ideas with direct discussion. Secondary attribution, on the other hand, reports other people's viewpoints implicitly. The writer may not take responsibility for the accuracy or make judgements to other person's statements. Lastly, tertiary attribution refers to the reconstruction or synthesis of other people's viewpoints. This three-fold classification requires a detailed reading of the text and the results may vary across annotators. Another classification, dividing attribution into direct and indirect quotations, is more straightforward and thus is used in the present study (see Section 5.2.2 for explanations).

Academic writing actively involves the evaluation and assessment of other researcher's viewpoints and integrates these with the writer's perspectives, and the use of modals is essential in this process. There are cases where the modals are used by antecedent authors in the literature and the writer reproduces these modals to report

the author's viewpoints. In these cases, writer's attitudes toward the antecedent author's points of view may be explicitly revealed by using a non-neutral reporting verb (e.g., *agree* or *refute*), or implicitly shown in the context. In either case, writers should not twist the meaning that the author originally wants to express. However, O'Brien (1995) finds one instance where the student changes the antecedent author's statement from a tentative one to a blunt one by omitting a hedging verb, *interpreted*. Thus, it is important, especially for student writing, to examine how modals are used to express the writer's and the antecedent author's voices, and how students distinguish between these two voices.

The identification of voices in the use of modality has been mentioned in passing in the previous literature, especially when the modality expresses epistemic meaning. Huddleston (1971) presents examples where the judgement of the truth of a proposition is from the person referred to in the text, as shown in 2-39 . The proposition about the chromatophores being potential structural artifacts is not the speaker's statement but rather Cohen-Bazire's viewpoint, which is paraphrased by the speaker.

> 2-39 Cohen-Bazire thinks that the chromatophores *may* be structural artifacts which have resulted from […]. (example provided in Huddleston, 1971, p. 300)

Thompson (2001) also acknowledges the importance of knowing who is responsible for the uncertainty when using epistemic modality and develops a functional approach for pedagogical reasons to analyse ‘when or why a writer might choose to use a modal auxiliary’ (p. 151). Likewise, Huddleston and Pullum (2002) emphasise that the attitude expressed may be also from the person mentioned in the sentence, but they do not add this in the definition of modality.

In the case of root modals, the textual voice is discussed in relation to the source of obligation, which mostly refers to the subjective and objective distinction, but their relation requires careful consideration. As mentioned before, the default source of obligation is the speaker, and in such cases, modals are mostly used subjectively. This is similar in academic writing, where the ‘default condition’ (Tadros 1993, p. 101) is averral, that is, the writer’s voice. In other cases where the source of obligation is external (e.g., rules, laws, and antecedent authors), the party that is responsible for laying the obligation is objective. The writer’s voice, in this case, can be integrated through, for example, reporting verbs or comments and explanations following and preceding the obligation. The subjective-objective distinction of the source of obligation is questioned by Collins (2009) as lacking consistency since sometimes the source is not specified, and thus it is hard to identify its status. The discussion on textual voice, although seems to be similar to the source of obligation, emphasises different aspects.

The source of obligation deals with the foundation of the writer's proposition, whether it is based on their beliefs, objective evidence, or the viewpoints of a third party. In contrast, textual voice focuses on the engagement with existing literature and the presentation of the writer's insights, which are particularly relevant to academic writing, and thus it is discussed explicitly in the present study.

The research mentioned above demonstrates that the conventional definition of modality cannot be fully applied to those used in academic writing due to the more focus on the spoken data and the omission of textual voice. It is necessary to give a working definition of modality in the context of academic writing. Inspired by Huddleston and Pullum's (2002) definition and the discussion on textual voice mentioned above, I will define modality in academic writing as linguistic expressions concerned with the attitude of the writer or the person referred to in the text to the factuality or actualisation of a proposition.

2.3.2 Modality used in academic writing

This section will discuss the studies of modality in academic writing that use a corpus-based approach as this approach is also used in the present study. It will first briefly present how modality is used differently in academic writing compared to other contexts. It is followed by a brief comment on the general trend of modality investigated in academic writing. In addition, a critical evaluation of the previous studies will be

presented in terms of the topic covered and the potential limitations. The last few paragraphs will examine the studies exploring the epistemic necessity and obligation modals with a focus on *must*, *have to*, and *should* in research articles and student writings.

As discussed in Section 2.2.2, researchers use different classification frameworks of modality for examination, and thus there might be inconsistency in terminology such as root, deontic, and dynamic modality used across studies. I will reproduce the terms used by the researchers when discussing them. No further explanation of their selection of the classification is provided due to space constraints in this thesis.

Modality in academic writing is different from its use in other contexts such as conversations and general written English, as noted by researchers such as Biber et al. (1999) and Parkinson (2020). In the case of the three modals in question, *must* and *should* are used more frequently in academic writing compared to other registers (e.g., conversation, fiction, and news), whereas *have to* shows an opposite trend (Biber et al., 1999). In terms of their meaning distribution, Biber et al. only compare the use in conversation and academic prose since they show clear contrast (see Figure 2.3 in Section 2.2.5). Epistemic *must* shows similar normalised frequency between the two registers, and it is also the case for epistemic *have to*, which is rarely used in this sense.

Epistemic *should*, on the other hand, is more frequently used in academic prose than in conversation. In terms of their root use, root *should* shows similar frequency in the two registers. By contrast, root *must* is used more often in academic prose but root *have to* shows an opposite trend. One limitation of their study is that the data is from the academic prose sub-corpus of the LSWE Corpus, which includes both book extracts and research articles, and they differ greatly in average text length (35,400 words in book extract and 8,050 words in research articles). In addition, a portion of the book extracts (16 out of 75) are written for the lay audience but not audience with a relevant background, which may add noise to the findings.

Other researchers compare the use of the modals in academic writing to general written English and find differences as well. Parkinson (2020) examines modals of obligation and necessity in three science registers: student laboratory reports and student essays in the British Academic Written English (BAWE) corpus (Nesi et al., 2008), and published research articles, each containing approximately 150,000 words. The selected disciplines for examination are Biology, Physics, Engineering, and other applied science disciplines. The findings of modal use are then compared with part of Collins's (2009) study, the written part of ICE-GB (about 400,000 words). This sub-corpus is used as a representative of general written English because it includes not only comparable student and academic writings but also other forms of writing such as

letters, reportage and instructions. It is found that *must*, *have to*, and *should* are used more objectively and more frequently in the epistemic sense in science registers compared to general written English. Deontic use of the three modals, on the other hand, is not as dominant as that is in general written English. The deontic usage is mostly objective, which reduces the authoritativeness of the writer. One example is 'it *must* be noted that…', and it is considered by Parkinson to be nearly formulaic in academic writing.

Warchał (2010) also use Collins's (2009) findings in ICE-GB as a representative of general written English, but the academic writing corpus she chooses for comparison consists of writings of linguistic research articles written by scholars that have the native-like command of English (size of 2.4 million words). Epistemic *have to* is found to rarely appear in both corpora but is slightly used more often in the academic writing corpus. However, epistemic *must* shows an opposite trend, with a larger difference in normalised frequency compared to general English, which partly contradicts Parkinson's (2020) findings. As for the root use, root *must* is slightly over-represented in research articles, whereas root *have to* is markedly less used in this sense compared to the case in general written English. The differences in the findings of Parkinson (2020) and Warchał (2010) are possibly due to the disciplinary variation in the academic writing corpus they select. A further discussion on the disciplinary variation

to the use of modals will be presented in Section 2.5.

The findings above confirm that the use of modality in academic writing differs from that in other contexts in terms of frequency and meaning distribution. This calls for a systematic exploration of modality in academic writing. How modality is investigated in this context is closely related to the views of academic writing. Historically, academic writing has been perceived as predominantly impersonal, prioritising objectivity to maintain scientific impartiality and universality. Nonetheless, over the decades, researchers have started to challenge this perspective, arguing that academic writing involves interaction between writers and readers (Hyland, 2018), and modality plays a crucial role in realising this interaction.

Studies on modality in academic writing have evolved through various stages, reflecting the shifts mentioned above. Early works, as mentioned in previous sections, focus mostly on grammatical and syntactic structures of modality, especially forms such as modal auxiliary verbs. This trend has subsequently transitioned towards a more nuanced understanding of dimensions such as pragmatics, discourse, and sociolinguistics. For example, the investigation of modality shifts to, for instance, how it is used to negotiate interpersonal relationships and reflects writers' identity and positioning within the research community. In addition, recent years also observed an

increase in corpus linguistic studies, where researchers examine modality in large collections of texts to identify patterns. These studies mostly examine research articles due to the availability of the texts to compile corpora, and some of them compare the findings across genres (e.g., Parkinson, 2020) and disciplines (see Section 2.5). There are also other studies that focus on academic writings of EFL students with different first language backgrounds (see Section 2.4.2).

Despite the extensive research on this topic from various viewpoints, there is a noticeable imbalance in the coverage of modalities in academic writing. Epistemic modality, particularly expressions of epistemic possibility, has received considerable attention. However, modal expressions conveying obligation and suggestions remain less explored, with a few exceptions that will be discussed in detail at the end of this section.

In addition, modality in academic writing is discussed mostly in terms of their pragmatic functions. One influential study is the work of Hyland (2005), who explores the stance and engagement markers, which are the sub-types of interactional markers and are used by writers to manage interaction and evaluation in writing. There are different devices to express stance and engagement, among which I will discuss those related to modal expressions. Hedges and boosters, as two of the four elements of stance

markers, are sometimes epistemic possibility modals (e.g., *might*) and epistemic necessity modals (e.g., *must*) respectively. Writers use hedges to make a provisional statement, leaving room for the reader to dispute the writer's interpretations (Hyland, 2005). Boosters, on the other hand, show the writers' certainty in their viewpoints and solidarity with their readers. One of the elements of engagement, the directives, is closely related to modal expressions. Directives can be expressed by obligation modals such as *must* and *should* to ask the reader to perform an action instructed by the writer, as exemplified in 2-40. The writer encourages the reader to agree with their perspective using the pronoun *we* and gives suggestions on what to consider next.

2-40 The next element we *should* focus on is its application.

A certain degree of overlap can be observed between modal expressions and these interactional markers. However, the difference between them is that modality reflects the writers' attitudes towards the content (truth value of their statements and their commitment to the arguments being made), while interactional markers concern the relational and interactive aspects of academic writing, shaping how writers connect with their readers and position themselves. Following Hyland (2005), a group of works emerges in the discussion of interactional markers in academic writing, some of which will be presented below.

As mentioned before, studies of epistemic modality in research articles are extensively studied and cover a wide variety of disciplines and expressions. These investigations will be examined in depth in Section 2.5, particularly focusing on variations across disciplines. While this section will not delve into these studies, a brief discussion will be provided to enlighten the research design of the present study.

Firstly, these studies have different focuses on expressions. The epistemic expressions investigated show a wide variety, ranging from one word such as modal verbs (e.g., *must*) or adverbs (e.g., *certainly*) to multi-word expressions such as 'it is possible that' and 'it seems likely'. For instance, Rozumko (2017) looks at epistemic modal adverbs (e.g., *apparently* and *possibly*) whereas Panocová and Lukačín (2019) examine the combination of the epistemic modal verb and the adverb as emphasiser (e.g., '*must* surely' and 'might actually'). Sameri and Tavangar (2013) examine a wider variety of epistemic devices proposed by Nuyts (2001), including modal auxiliaries, modal adjectives, modal adverbs, and mental state predicates. With the wide variety of expressions, it is difficult to generalise and compare these findings.

Apart from those studies focusing on specific forms, there are others extracting expressions based on their pragmatic functions (e.g., Farrokhi & Emami, 2008; Sameri

& Tavangar, 2013). In this case, an in-depth reading of the texts is essential for annotation since a pre-decided list of them might not be comprehensive. In addition, some of them are polysemous and the researchers need to distinguish their different meanings. The annotation process requires subjectivity, and the results are likely to differ among researchers. The workload of annotation might be heavy and only a limited number of texts can be examined in one study. To overcome these limitations, the present study takes a form-focused approach, extracting instances of *must*, *have to*, and *should* and annotating their meanings (see Section 1.2 for justifications to select these modals). This approach helps to examine a larger number of texts, and it is also predominantly used in learner corpus research, as will be presented in Section 2.4.2. In the meantime, we must admit that the function-based approach has its merits in that it could show a bigger picture of how writers use different devices coherently in the context to express epistemic meanings. Therefore, a qualitative analysis is also conducted for a fine-grained view.

There is only a handful of studies that examine specifically epistemic necessity and obligation modals. Some of them will be discussed below, and those examining disciplinary variation will be presented in Section 2.5.

In her studies, Warchał (2007, 2008) uses a similar research design, exploring the use

of *must* and *should* respectively in a corpus of 200 linguistic-related research articles with 2.4 million words. Both studies examine *must* and *should* in terms of their meaning distribution, clause type, and most importantly, their attributed use, in which a third party's view is involved and it may not be the same as the writer's viewpoints. This aspect has largely been overlooked by other researchers in this field.

According to Warchał (2007), the percentages of attributed use of epistemic and root *must* are similar in main clauses, accounting for 4% and 3% respectively. By contrast, the percentage of epistemic *must* (15%) is twice as high as that of root *must* (8%) in subordinate clauses. In the case of *should*, Warchał (2008) reports that the proportions of attributed use for its two senses are similar in main clauses, accounting for 2%, whereas the percentage in subordinate clauses is 12% for both senses. We can conclude that the attributed uses of *must* and *should* occur more frequently in subordinate clauses than in main clauses. Attributed use can relieve the writer from the responsibility of making arguments in regard to the factuality or actualisation of a proposition. Its frequent use in research articles confirms the necessity to introduce a revised definition of modality in academic writing (see Section 2.3.1).

Additionally, Warchał (2007) reveals that epistemic *must* illustrates the features of its core uses proposed by Coates (1983), such as frequent co-occurrence with stative

verbs and the perfect aspect. Root *must*, on the other hand, deviates from the core use of expressing strong obligation. Its strength of obligation is reduced by syntactic features such as the use with passives or third-person subjects. In regard to root *should*, it is used closer to its core sense, expressing strong and subjective suggestions (Warchał, 2008). However, Warchał (2008) notes that this claim needs to be treated with caution since the examination of the syntactic features of *should* is not as extensive as that of *must* in Warchał (2007).

In specific relation to the subject types associated with the modals, Warchał (2008) does not cover this aspect when investigating *should*. Warchał (2007), on the other hand, examines this feature of *must* and observes similar findings to Coates (1983) (see Section 2.2.4), such as the association between epistemic *must* and the third-person inanimate subject and existential subject *there*. By contrast, root sense of *must* in research articles demonstrates a different preference compared to Coates's findings. It is reported to co-occur more frequently with third-person inanimate subjects and inclusive *we* (refers to both writers and the readers) rather than second- and first-person subjects, possibly due to the different corpora used. Furthermore, Warchał (2007) highlights a pattern where root *must* is used with verbs that denote both speaking and mental process (e.g., *mention*, *note*, or *conclude*), a finding corroborated by Coates (1983) and Palmer (1990) (see Section 2.2.4). However, Warchał (2007)

notes that this combination often appears alongside the inclusive *we*, whereas Coates observes its frequent use with first-person subjects.

A more recent work on necessity and obligation modals is conducted by Oktavianti (2019), but she does not distinguish between the meanings of the modals. *Should* is found to be more frequently used than *must* and *have to* in the academic sub-corpus of the Corpus of Contemporary American English (COCA; Davies, 2008). It is also concluded that 'thinking verbs' (Oktavianti, 2019, p.53) such as *consider* and *understand,* are observed to be the most significant collocates of these modals.

Writers in research articles express necessity and obligation to influence their readers and show their power, but in the meantime, they also need to seek agreement, participation, and solidarity from the readers (Koutsantoni, 2004). Students' academic writing shows a slightly different picture in the modal use. Some of the studies will be discussed below, and those of Chinese EFL academic writings will be illustrated in Section 2.4.2.

Vincent (2020) explores the expression of obligation in proficient student writings in the BAWE corpus, including a wide coverage of disciplines and students from diverse first language backgrounds. He adapts Hyland's (2002b) framework of obligation

expressions and classifies the instances into three main categories: textual, physical and cognitive acts, and the first category has no instances in his data. Two additional orientations of obligation expressions, subjective vs. objective, and explicit vs. implicit, are also discussed. The expression covers beyond modal auxiliary verbs, and the focus is not only on frequency distribution but also on their functions. It is found that verbs used with the obligation expressions that convey cognitive rhetorical acts can be categorised into semantically similar groupings, such as those denoting examining, understanding, and explaining. He adds that some of these verbs, such as *examine* and *understand*, tend to engage the reader more directly in cognitive activities compared to the others (e.g., *explain* and *clarify*). Similar to Vincent, Lee (2010) does not specifically focus on necessity and obligation modals. Instead, she explores the expressions of command strategies, which include *must* and *should*. Upon close examination of twelve undergraduates' essays of EFL learners, she observes that higher-graded essays tend to employ these expressions more cautiously and suitably for the context compared to lower-graded ones, which often demonstrate a strong assertion of authority with little consideration for the institutional positioning.

Parkinson (2022) examines the use of *must*, *should*, *have to,* and *need to* in a corpus consisting of laboratory reports in Biology and Physics written by South African English as a Second Language (ESL) science undergraduates, with a size of 62,071 words.

The findings are further compared to a selection of texts within the BAWE corpus, consisting of writings from native English students in comparable disciplines with 101,434 words. It is found that ESL students use more *must* than *should*, especially when expressing deontic meanings. In addition, the subjective use of the deontic modals is preferred by ESL students. An explanation for the high frequency of *must* is that ESL students tend to use the modals similarly to their usage in general English (Parkinson, 2020). It might also be related to the features of the variety of South African English compared to other varieties of English. This analysis offers the profile of the modals in the laboratory reports, yet it does not address the disciplinary differences between Biology and Physics.

2.4 Modality in Chinese EFL students' writing

This section starts with a brief introduction of Chinese modality system in terms of expressions and meanings, which could provide insights into potential first language influence on the modal use in Chinese students' writings. It is then followed by a close examination of studies on Chinese EFL students' use of English modals. It is important to note that while Chinese can refer to various dialects, such as Cantonese and Shanghainese, this section focuses exclusively on Mandarin, which is the official language of China and the medium of instruction in Chinese universities.

2.4.1 Chinese modality

Chinese modality is a controversial topic regarding its definition and classification, and its studies have evolved over time. Early works are mostly descriptive, documenting the uses of modality without building a theoretical framework. In the 20th century, with the introduction of Western linguistic theories, the investigation of Chinese modality is more systematic. Researchers, as will be mentioned below, start to explore the syntactic structures and semantic properties of modal expressions, leading to a deeper understanding of how modality is constructed and interpreted in Chinese.

A foundational study that systematically applied Western methodologies to the examination of modal auxiliary verbs is the work of Ma (1989). He uses the term *zhudongci [助动词] 'auxiliary verb'* to denote the expressions associated with Chinese modality, and explores four forms: *ke [可] 'may'*, *neng [能] 'can'*, *zu [足] 'be sufficient to'*, and *de [得] 'can'*. Subsequent to this research, Chinese researchers have conducted detailed investigations into this topic, but some of them use other terms to describe Chinese modality, such as *nengyuan verbs [能愿动词] 'modal verbs'* (Liu et al., 1983), *hengci [衡词]* (Chen, 1978), and *nengci [能词]* (Gao, 1986). Among these terms, *zhudongci [助动词] 'auxiliary verb'* proposed by Ma is the most used one, as summarised by Li (2004). One example of the Chinese modal auxiliary verb is presented in 2-41 below. The first line of the example presents the Chinese pinyin alongside the corresponding Chinese words. The second line provides an English

translation for each Chinese word, and the final line offers a paraphrased English sentence. *Yinggai [应该] 'should'* is used to express the confidence in the judgement that he is at home.

2-41 Ta yinggai zai jia. 他应该在家。

He should at home.

He should be at home.

Unlike English modal auxiliaries which share similar morpho-syntactic properties (see Section 2.2.3), Chinese modal auxiliaries primarily differ from other auxiliaries in terms of semantics. Despite the difference in terms used, Chinese modality, similar to English modality, is generally considered as a semantic category that concerns the speaker's attitude toward a situation or an action.

Apart from modal auxiliary verbs (e.g., *yinggai [应该] 'should'*), there are other categories of expressions, such as modal particles (e.g., *ba [吧]*) and modal adverbs (e.g., *yiding [一定] 'definitely'*). Chinese have a rich system of modal articles that can add nuances to a sentence without changing the basic syntax, and most of them are placed at the end of a sentence, as exemplified in 2-42. The modal particle *ba [吧]* is used to reduce the strength of obligation, transforming a command into a suggestion

when included. The sentence remains grammatically correct without *ba [吧]*, but it expresses a stronger sense of obligation. By including *ba [吧]* at the end of a sentence, the speaker indicates that the action is not strictly required but recommended.

2-42 Ni xian zou ba. 你先走吧。

You first go ba.

You go ahead.

Unlike modal auxiliaries and modal adverbs that are exclusively used to express modality, modal particles have other functions such as marking aspect and indicating question type, as shown in 2-43. *Ba [吧]* in this example is used following the aspect marker *le [了]*, which indicates that the action of eating is considered to be completed. The inclusion of *ba [吧]* at the end of the sentence transforms the declarative sentence into a question that seeks confirmation of the truth of the preceding statement.

2-43 Ni chi le ba? 你吃了吧？

You eat le ba?

You've eaten, haven't you?

Modal particles are mostly found in languages such as Chinese or German and are

not prevalently used in English (Palmer, 2001). Thus, Chinese modal particles are not discussed in the following paragraphs since their counterparts in English are not systematically presented.

As for modal adverbs, there is a lack of consensus among researchers regarding the difference between Chinese modal auxiliary verbs and modal adverbs. What is similar between the two parts of speech is that they both occur in the pre-predicate position, as shown in 2-41 and 2-44, respectively. In both cases, the modal expressions are placed before the predicates *zaijia [在家] 'be at home'* and *lai [来] 'come'*.

2-44 Ta yiding lai. 他一定来。

He definitely come.

He will definitely come.

The difference between modal auxiliaries and modal adverbs lies primarily in their verbal properties. Li and Thompson (1981) observe that modal auxiliary verbs are semantically connected to the subject of the verb and can stand alone in responses to yes-no questions. For instance, in response to the question 'Can he come?', one might simply answer with the modal auxiliary verb *neng [能] 'can'*, omitting the lexical verb. In contrast, modal adverbs such as *yiding [一定] 'definitely'* lack the *verblike* (Li &

Thompson, 1981, p. 181) property and cannot be used independently in such an example. The response to the same question would be *yidingneng [一定能] 'definitely can',* which requires the inclusion of another modal auxiliary verb to complete the sentence.

However, the classification of certain modals, such as *bixu [必须] 'must'*, remains controversial. While Li and Thompson (1981), Tang and Tang (1997), and Lin (2012) categorise *bixu [必须] 'must'* as a modal auxiliary verb, it does not entirely conform to the criteria of having the verblike property. Conversely, CKIP (1993) and Cai (2010) argue for its classification as a modal adverb. I would argue that the attempt to rigidly distinguish between modal auxiliaries and modal adverbs in Chinese may be problematic. Unlike English, where modal expressions could be classified based on structural and formal features, Chinese modality appears to lack such overt syntactic and morphological markings (Hsieh, 2005). Consequently, it may be more appropriate to prioritise the semantics of Chinese modals over their structural distinctions when considering their classification.

In terms of the meaning classification of Chinese modality, Hsieh (2005) consolidates multiple perspectives into three distinct models, providing a systematic and detailed representation of divergent viewpoints on this topic. The three models are cross-

language models, individual-language models based on English, and individual-language models based on Chinese. The following paragraphs will briefly describe the three models.

The cross-language models can be applied to any language and in the meantime identify grammatical properties that are exclusive to Chinese. One example is Tsang's (1981) classification, dividing Chinese modality into two types: epistemic and deontic. Six forms are recognised as Chinese modal auxiliary verbs, as listed in Table 2.3 However, this classification excludes what is defined by Palmer (1990) as dynamic modality (e.g., ability and volition), which leads to an incomplete description of Chinese modality. In addition, although cross-language models offer valuable perspectives for comparative research, they might not entirely capture the distinctive aspects of Chinese modality due to their generalist approach.

Table 2.3 Tsang's (1981) classification of Chinese modality

Epistemic		Deontic	
Possibility	Necessity	Possibility	Necessity
hui [会] 'will' neng [能] 'can' nenggou [能够] 'can'	gai [该] 'should'	neng [能] 'can' nenggou [能够] ' can' xu [许] 'permit'	yao [要] 'must' gai [该] 'should'

The individual-language models based on English are the models most frequently used

to investigate Chinese modality. These models are developed from the classification of English modality. For example, Tiee's (1985) classification follows Palmer's (1990) framework, dividing Chinese modality into three types: epistemic, deontic, and dynamic. A more recent and comprehensive study is conducted by Li (2004), who follows Van der Auwera and Plungian's (1998) framework, classifying Chinese modality into two categories: epistemic and non-epistemic. He identifies 16 modal auxiliaries and groups them into semantic categories, as shown in Table 2.4 below. While these models are comprehensive and well-structured, they are inherently influenced by English linguistic frameworks, which may impose limitations when applied to Chinese. The structural variances between the two languages indicate that these models may overlook or misinterpret certain Chinese modality.

Table 2.4 Li's (2004) classification of Chinese modal auxiliary verbs

Epistemic		Non-epistemic			
Uncertainty	Probability	Participant-internal		Participant-external	
		Ability	Need	Permission	Obligation
keneng [可能] 'may'	gai [该] 'should'	neng [能] 'can'	yao [要] 'need'	neng [能] 'can'	yao [要] 'must'
hui [会] 'may, can'	yinggai [应该] 'should'	nenggou [能够] 'can'	xuyao [需要] 'need'	nenggou [能够] 'can'	dei [得] 'must'
neng [能] 'can'	yao [要] 'will'	hui [会] 'can'	dei [得] 'must'	ke (yi) [可(以)] 'may'	yinggai [应该] 'should'
nenggou [能够] 'can'	dei [得] 'must'	ke [可] 'can'		de [得] 'can'	gai [该] 'should'
de [得] 'can'		keyi [可以] 'can'			yingdang, ding, dang [应当，定，当] 'should'
ke (yi) [可(以)] 'can'					

The last framework model does not simply apply cross-linguistic or English-centric frameworks but rather considers the specific linguistic features of Chinese modal expressions. For instance, CKIP (1993) categorises Chinese modality into two semantic categories, epistemic and deontic, and further identifies three categories with different grammatical properties: modal auxiliaries, modal adverbs, and modal verbs.

Hsieh (2005), on the other hand, introduces the concept of 'source involvement' as a critical factor in understanding Chinese modal expressions. This semantic-based approach provides a more effective way of capturing the full range of meanings expressed by Chinese modals compared to frameworks like CKIP (1993), which focus primarily on grammatical features. These features are less distinctive in Chinese modality compared to English modality, as previously discussed. Hsieh's approach categorises Chinese modality into four types: epistemic, deontic, dynamic, and evaluative, each defined by whether they inherently include the source of opinion or attitude. For instance, under epistemic modality, modals like *keneng [*可能*] 'may'* mostly include the speaker as the source of opinion, whereas in deontic modality, a modal such as *keyi [*可以*] 'can* ' may express permission granted by an unspecified authority rather than the speaker. Evaluative modality, illustrated by *xingkui [*幸亏*] 'fortunately'*, indicates the speaker's positive attitude towards the occurrence of an event, mostly

reflecting the speaker's viewpoint.

In summary, there is a divergence in the terminology and categories for studying Chinese modality. The semantic criteria for characterising the expressions of modality also vary among researchers. Some of them only focus on forms expressing epistemic and deontic modality and exclude the uses describing ability and volition (e.g., Tsang, 1981), while others go beyond these meanings and extend modality into forms conveying evaluation on known facts (e.g., Hsieh, 2005). Despite the controversial issues mentioned above, it remains evident that Chinese modality is recognised as a discrete category, exhibiting clear distinctions from other verbs. In addition, some of the forms are polysemous, as illustrated in Tables 2.3 and 2.4. These are the similarities between the Chinese and English modality.

As for the differences, they are also worth discussing since they could be the potential influence on the use of the English modals by Chinese EFL students. One notable difference is the variety of modal expressions. As mentioned above, modal particles are used in Chinese to express modality, whereas English modality does not have equivalent forms. In addition, although most modal meanings can find comparable forms in both languages, their encoding system is different. A single English modal can correspond to several variants in Chinese, reflecting the nuanced expressiveness of

the language. For instance, as shown in Table 2.4, there is a list of Chinese modals that can be translated into *should*, including *yingdang [应当], ding [定], dang [当], yinggai [应该],* and *gai [该]*. Both *yingdang [应当]* and *yinggai [应该]* can be translated as root *should*, but they differ in the strength conveyed. *Yingdang [应当]* typically implies a stronger sense of obligation and is more commonly used in formal or legal contexts, as shown in 2-45, while *yinggai [应该]* is often employed to express suggestions in everyday communication, as in 2-46.

2-45 Ni yingdang zunshou falv. 你应当遵守法律。

You should obey law.

You should obey the law.

2-46 Ni yinggai zao dian shuijiao. 你应该早点睡觉。

You should early bit sleep.

You should go to bed earlier.

Chinese students may face challenges in learning English modals due to the multifunctionality and different expressions of modals used in their native language, potentially leading to negative first language transfer. However, this process may also offer opportunities, if facilitated properly, to enhance cognitive flexibility and cross-linguistic skills through critical engagement and comparison between languages.

Another difference is related to the multi-modal structures. Modal auxiliary verbs in English cannot be used consecutively unless used with semi-modals such as ‘might have to’ (Palmer, 1990). In the case of Chinese modality, not only can the modals co-occur with each other, but the number of modals in a multi-modal structure can be as many as four (Peters & Bembridge, 2016), as exemplified in 2-47. Four Chinese modals are used in this example, including *yinggai [应该] ‘should’*, *keneng [可能] ‘may’*, *hui [会] ‘will’*, and *nenggou [能够] ‘can’*. Although this usage is not commonly used in Chinese, as noted by Lin (2012), it is grammatically correct.

2-47 Zhangsan yinggai keneng hui nenggou lai. 张三应该可能会能够来。

Zhangsan *should may will can* come

‘It should be the case that it is likely that Zhangsan will be able to come.’

(Example provided in Lin, 2012, p. 152)

In addition, modals used in these structures can express similar as well as different meanings, with ordering restrictions. Huang (2009) examines the multiple modal structures and highlights the ordering restrictions among the modals. One such restriction is that an epistemic necessity modal (e.g., *yinggai [应该] ‘should’*) needs to be placed before an epistemic possibility modal *(*e.g., *keneng [可能] ‘may’*), and not the

reverse, as illustrated in 2-48. The epistemic necessity modal *yinggai [应该] 'should'* precedes the epistemic possibility modal *keneng [可能] 'may'*, demonstrating that the speaker initially expresses a stronger confidence that he is at home but subsequently expresses a more cautious possibility by adding the epistemic possibility modal.

2-48 ta yinggai keneng zai jia. 他应该可能在家。

he *should* may at home.

'It *should* be the case that he is likely to be at home.' (Example provided in Huang, 2009, p. 537)

Additional differences between Chinese and English modals include the absence of contracted forms in Chinese modals, syntactic differences in interrogative sentences, and negations that involve modals. These differences, as well as the similarities, help to shed light on the discussion on the influence of the first language on the use of English modals by Chinese EFL students.

2.4.2 Chinese EFL students' use of modality in writing

The exploration of modals used by Chinese EFL students has gained interest in applied linguistics over the decades. The investigation has shifted from a primarily error-focused perspective to a more holistic view. Researchers incorporate the interlanguage theory and use corpora to uncover the profiles of modals in Chinese students' writing.

With the investigation of large datasets of learner language, patterns revealed are more likely to be generalised and thus can contribute to teaching implications. An overview of learner corpus research and analysis framework will be illustrated in Section 2.6. This section focuses on examining these corpus-based studies, comparing their finding, discussing potential reasons leading to the different use of modals by Chinese students and native English speakers, and providing teaching implications.

Before jumping into the discussion on the studies, I would like to first define two terms that will frequently appear in the following paragraphs, ‘English learner’ and ‘native English speaker’. ‘English learner’ generally refers to those whose first languages are not any variety of English, which can be further categorised based on the environment as ESL and EFL learners (Gass & Selinker, 2008). Learners in regions where English is the primary language fall into the ESL category. In these regions, English serves not only as a subject but also as the main language of social interaction and education. By contrast, EFL learners, such as the Chinese EFL students discussed in this study, are taught English as a distinct subject rather than as a medium of everyday communication. In the discussion that follows, *learners* refers to EFL learners unless otherwise stated.

Most of the studies mentioned below term Chinese students as *learners*, and I will follow what these researchers use when reviewing the literature. However, in my analysis, I refer to them as 'Chinese students' rather than 'Chinese learners'. As Selinker (2014) argues, non-native speakers who use complex syntactical features should not merely be viewed as language learners, even though they may still exhibit certain linguistic divergences from native speakers. The learner corpus I use consists of writings of Chinese EFL undergraduates' writings. These students demonstrate grammatical sophistication (as will be shown in Chapters 4 and 5) and engage in cognitive and academic activities in higher education settings, and thus I identify them equally to the British students as 'Chinese students'.

'Native English speakers' refers to those whose first language is English and who acquire it at an early stage through exposure to the environment. As Cook (1997) summarised, some researchers argue that the goal of language learners should be to achieve competence equivalent to that of native speakers. By contrast, with the advocacy for learners' language as a separate system (see Section 2.6 for a discussion on interlanguage), another body of research advocates against comparing learner language to native language. I take a neutral stance, considering comparisons to native language useful for identifying characteristics of learners' English, but not as the sole target or guideline for learners. This perspective is consistent with Gilquin (2022),

who argues that the native language should be considered one of many norms (further discussed in Section 2.6), serving as a point of reference rather than the perfect standard. Correspondingly, the corpus consisting of native speakers' writings is referred to as the 'reference corpus' in the present study.

A commonly used framework to examine Chinese EFL students' use of modals is Contrastive Interlanguage Analysis (CIA; Granger, 1996), which will be discussed in detail in Section 2.6. In short, researchers would compare the use of modals in at least two corpora, consisting of writings of Chinese EFL students and native English speakers. The frequently used Chinese learner corpora are the Spoken and Written English Corpus of Chinese Learners (SWECCL; Wen et al., 2009) and the Chinese Learner English Corpus (CLEC; Gui & Yang, 2003). SWECCL consists of two sub-corpora, containing spoken and written material (each with one million words) respectively. The data in the spoken sub-corpora is collected from oral English tests taken by English major undergraduates. The written sub-corpora contain students' essays on various topics from Years one to four at Chinese universities, and the average length of the texts is approximately 400 words. CLEC comprises solely written material with one million words in total. It includes argumentative essays on a variety of topics such as society, education, and technology. Each essay must be at least 150 words in length but typically does not exceed 300 words. These students show wider

educational levels compared to those in SWECCL, ranging from senior high school students to university students in English and non-English majors (altogether five levels). A shared feature of these two corpora is that they contain short argumentative essays on various topics written by mainly university students. However, there is a lack of research comparing longer and discipline-specific academic writing, except Yang (2018) (to be discussed shortly). The lack of variety in the writings is related to the accessibility of the corpus. The open-accessed English corpora of Chinese learners are limited and are not quite up-to-date. Researchers either use these pre-compiled corpora directly, acknowledging their potential limitations, or they compile their own. However, these self-compiled corpora are generally not open-accessed, requiring researchers to obtain permission from the compilers to access the data.

As for the native speaker's corpus used for comparison, the most selected one is the Louvain Corpus of Native English Essays (LOCNESS; Granger, 1998) since it includes writings of American and British university students' essays that share similar educational levels of the students in most of the Chinese learner corpus. Another option is the Michigan Corpus of Upper-level Student Papers (MICUSP, 2009), which consists of A-grade papers written by university students with different first language backgrounds (around 2.6 million words). A less satisfactory reference corpus is the BNC, which includes texts from a range of genres (e.g., spoken, magazines, fiction,

and academic) with about 100 million words. It is used when the learner corpus includes both spoken and written material, as in Xiao's (2017) analysis of SWECCL, or when comparisons are related to spoken material, as seen in Li's (2020) study of junior high school English textbooks (will be discussed shortly). However, the educational level of the writer and the topics of the texts in the learner and the reference corpus differ in these two studies, which may add noise and variables to the analysis.

Yang (2018), unlike other studies, compiles a corpus consisting of Chinese second-year undergraduates' research reports for the coursework of International Business and Trade, containing 246 files with 1,637,722 tokens. The professional corpus for comparison consists of published articles in the Journal of Business Research written by researchers with various first language backgrounds, containing 206 texts with 1,637,460 tokens. Although these two corpora are more comparable than the corpora we mentioned above since they share similar topics, we must admit that there are still differences in genres and other parameters of writing such as time for writing and the use of other resources while writing. However, compromises have to be made due to the accessibility of the corpora available. What we could do, like most studies that will be mentioned below, is to find the most comparable corpora and bear in mind their limitations and incomparability. The selection of the corpora used in the present study will be presented in Section 3.2.

Most of the previous studies examine the frequency distribution of modals, with a particular focus on central modals (e.g., Bai, 2015; Long, 2013). There are only a few exceptions such as Liang (2008) and Tang (2013) which distinguish the meanings of the modals. In addition, these studies mostly examine how first languages or proficiency levels influence the use of modals, and this is largely due to the fact that the available corpora consist of writings at different academic levels.

Table 2.5 below lists the findings of the frequency distribution of modals in Chinese EFL students' writing and their comparison with texts written by native English speakers. The corpora used in these studies are the ones previously mentioned. Other findings of these studies (e.g., the impact of proficiency level) and studies that distinguish the modal meanings will be discussed separately in the subsequent paragraphs.

Table 2.5 Frequency comparison of modals used by Chinese and native speakers in the previous literature with three target modals highlighted in grey (*O* refers to over-representation, *U* refers to under-representation, and dash indicates similar normalised frequency distribution)

	Yang (2018)	Bai (2015)	Tang (2013)	Long (2013)	Liang (2008)	Cheng & Qiu (2007)	Ma & Lu (2007)
Can	**O**	**O**	**O**	**O**	**O**	**O**	**O**
Could	**O**	—	U	U	U	U	—
May	U	—	U	**O**	—	—	—
Might	—	U	U	U	—	U	—
Will	**O**	**O**	**O**	**O**	**O**	**O**	—
Would	**O**	U	U	U	U	U	—
Shall	—	**O**	**O**		—	**O**	U
Should	—	—	**O**	**O**	**O**	**O**	**O**
Must	—	—	U	**O**	**O**	**O**	—
Have to			**O**				**O**
Used to			**O**		—		
Had better			**O**				—
Be going to			—				U
Be supposed to			—				
Dare			—		—		**O**
Need (to)			—	**O**	—		—
Ought (to)			—	**O**	—		—
(Have) got to			—				—

In the table, the columns are ordered according to the year of publication of the research article. The latest ones are placed on the left. The cells of *must*, *have to*, and *should* are in the grey background since they are the target modals in the present study. The line with a bold border separates the central modals and the others. *O* refers to the over-representation of the modal in the students' writing compared to that of native speakers and is shown in bold font. *U* refers to the under-representation, and dashes are used to indicate that there is no marked difference in the normalised frequency of the modals between the learner and the reference corpus. The over- and under-representation are used to describe the difference in frequency (Granger, 2015), and they should not be used as an indication of the (in)sufficient knowledge of modality used by Chinese learners. The selection of these terms will be further discussed in Section 2.6.

As shown in the table, researchers mostly examine central modals, with a few studies expanding the investigation to semi-modals. Chinese learners generally use modals more frequently than native speakers in writing (Cheng & Qiu, 2007; Liang, 2008; Ma & Lu, 2007; Tang, 2013; Yang, 2018). To be more specific, *can, will,* and *should* are mostly over-represented in learners' writing. By contrast, *could, might,* and *would* are mainly recognised as under-represented modals. The case for *may, shall,* and *must* is

complicated because the findings differ among studies. I will discuss the finding of *must*, *have to*, and *should* in more detail since they are the target modals in the present study.

Among the nine studies listed, *have to* is the least investigated modals among the three target modals, and both Tang (2013) and Ma and Lu (2007) identify it as an over-represented modal in the learner corpus. *Should* is found to be over-represented by Chinese EFL students in five out of the seven studies, with the remaining studies finding no difference between the two groups of students. In addition, Gao (2023) also identifies an over-representation of *should*, and this pattern occurs in the Chinese writings of all seven proficiency levels investigated. Essays examined in Gao's study are written by middle school and university Chinese students, and each text is graded from one to seven based on the Chinese Standards of English Language Ability (CSE; Ministry of Education of China, 2018). It can be concluded that Chinese learners use *should* and *have to* more frequently than their native-speaker counterparts in most of the previous studies.

Unlike *should* and *have to*, *must* is widely investigated but the findings of its frequency distribution compared to the native speakers' data differ among studies. Cheng and Qiu (2007), Liang (2008), and Long (2013) conclude that *must* is over-represented in Chinese learners' writings. Tang (2013), however, identifies an opposite trend. As for

Bai (2015), Ma and Lu (2007), and Yang (2018), they do not find frequency difference in the use of *must* between the two groups of students. The inconsistency in the findings of frequency might result from the different proficiency levels of the students examined. Despite using the same corpus, CLEC, Liang (2008) and Ma and Lu (2007) analyse different sub-corpora: Liang examines writings across all five levels, while Ma and Lu focus exclusively on the writings from the highest level. In addition, Gao (2023) observes that *must* is over-represented in writings graded at low levels (CSE 1-5), and it appears less frequently in writings assessed at higher levels (CSE 6-7).

Another reason could be the topics covered by the corpora. For example, the corpus examined by Yang (2018), as mentioned before, consists of undergraduates' academic writings on the topic of business and trade. By contrast, Bai (2015) uses CLEC, which contains writings of both senior high school and university students on topics such as Chinese festivals and friendship. The variation in topics may lead to differences in the usage of modals, as noted by Yang et al. (2005). Similarly, Hinkel (2009) also finds that Chinese students tend to use more obligation and necessity modals compared to native English speakers when discussing topics such as the role of parents and the importance of receiving higher grades because these topics reflect the cultural norms of duty and responsibility. The learner corpus that Hinkel uses includes 455 essays of Asian ESL students (Chinese, Japanese, Korean, Indonesian and Vietnamese)

collected in five years, and the native speaker corpus consists of essays on similar topics of the learner corpus written by 280 first-year university students raised in midwestern states of America.

Among the seven studies illustrated in the table, two of them, Liang (2008) and Tang (2013), make distinctions between modal meanings. These two studies follow a similar research design, comparing CLEC with LOCNESS. It is found that root modals are over-represented in the learner corpus compared to the native speaker corpus, whereas epistemic modals show an opposite trend. However, the distinction of the meanings is not based on manual annotation. Instead, they list modal sequences based on the previous literature and link these sequences to the meanings of modals. For example, two root modal sequences are identified, 'modal + dynamic verbs' and 'agentive subject + modal'. I would argue that using these syntactic features to identify the meaning of a modal is not reliable since the association is not definite, as pointed out by Coates (1983), and exceptions can be identified. For example, in 2-49, the subject is agentive but *should* expresses epistemic meaning.

2-49 She *should* be here any minute.

The differentiation in the meanings of modals does not depend solely on their syntactic

features but is also related to a wider variety of co-textual features, as demonstrated in Section 2.2.4. Kecskes and Kirner-Ludwigs (2017) compare two approaches to annotate modal meaning, based on semantics and syntax respectively. The results of the annotation show deviations, and they argue that the syntax-based approach is not reliable enough on its own. They further reveal that the deviation between the two annotation results shows the highest degree in the native speaker's use of *should* because native speakers use it less homogenously and consistently compared to the language learners. Thus, it seems necessary to use the semantic-based approach or combine different approaches to distinguish the meanings of modals (refer to Section 4.2.1 for the approaches used in the present study).

Although the results of Liang (2008) and Tang (2013) in terms of meaning distribution are questionable, they do offer some insights into the syntactic features co-occurring with the modals. Liang notes that root *should* and *must* are less frequently used in the passives by Chinese students compared to their native-speaker counterparts. One reason pointed out by him is that Chinese students tend to feel responsible for giving suggestions and thus frequently use first-person subjects with the obligation modals as in 'we *should* pay attention to', whereas native speaker students describe the suggestion more objectively and tend to omit who needs to take the suggestion. Tang adds that Chinese learners frequently use 'modal + infinitive' over other sequences

such as modals in perfect aspect or passives, possibly because learners perceive this combination as more likely to be used correctly.

There are also studies that only look at epistemic modality in Chinese students' writing, which will be discussed shortly. However, to the best of my knowledge, research specifically addressing the use of obligation modals remains unexplored, with only a few studies presented at the end of Section 2.3.2 but focusing on students with a different first language background.

In regard to the studies of epistemic modality, Chen (2012) explores the epistemic stance used in CLEC at different levels and compares it with LOCNESS. It is found that epistemic *must* and *should* are more frequently used by Chinese undergraduates compared to native students, expressing a relatively stronger assertation. Likewise, Hu and Li (2015) identify the over-representation of epistemic *must* when comparing the two components of the International Corpus Network of Asian Learners of English (ICNALE; Ishikawa, 2013): the Chinese and the native English speaker sub-corpus. Both studies note that with the progress of proficiency level, students' ability to qualify the propositions appropriately improves. Hu and Li (2015) add that Chinese students with a higher proficiency level tend to be more tentative and their use of epistemic expressions is closer to that of native speakers.

Thus far, we have presented mainly the statistical findings in the use of modals by Chinese EFL students. The subsequent paragraphs will discuss the underlying reasons for the differences observed between the modal use of Chinese students and that of native English speakers. They can be summarised into three aspects: first language influence, knowledge of the modals, and cultural differences.

Firstly, Selinker (1972) identifies language transfer as one of the central processes in language learning, suggesting that learners' production could be traced back to influences from their native language. According to Cheng and Qiu (2007) and Gao (2023), first language influence could explain the over-representation of some modals (e.g., *can* and *should*) in Chinese students writing. Cheng and Qiu identify the misuse of *can* in Example 2-50. In the given context, w*ill* is more appropriate than *can* to refer to the future action. However, the Chinese translation of 'you will find' is *ni hui fa xian [你会发现]*, in which *hui [会]* is often translated in English as *can*, leading to this misuse and potential over-representation of *can* in similar examples.

> 2-50 First, you may feel difficult. Later you *can* find you needn't see the hole of the needle, you can cross the thread. (example provided in Cheng & Qiu, 2007, p. 12)

Another example is root *must*, which is found by Li (2020) to be pragmatically different to its Chinese translation. The Chinese translation of root *must* does not imply that the writer is in the authoritative position and thus its strength is relatively weaker than its English equivalent. Chinese translations of root *should*, *yinggai [应该]* and *gai [该]*, on the other hand, show little pragmatic difference from their English equivalent. In addition, the two Chinese translations of root *should* can apply to both epistemic and root meaning. By contrast, *yiding [一定]* and *bixu [必须]* express the epistemic and root sense of *must* respectively. Since root *should*, in comparison to root *must*, appears to be more directly translatable and applicable between Chinese and English, Chinese EFL students may feel more familiar and comfortable using root *should*, leading to its over-representation in their writing compared to that of native English speakers.

Differences between student groups may also result from how modals are taught in China, a process in language learning Selinker (1972) refers to as 'transfer of training,' which is relevant to both textbooks' and teachers' presentation of the modals. Cheng and Qiu (2007) find that EFL teachers in China sometimes simplify the use of the modals without providing enough context, and unequally emphasise one of the meanings. As a result, students tend to rely more on the modals and the meanings that the teachers focus on. This can also be observed in textbooks. Yang (2018) points out

that the coursebooks in China sometimes only present Chinese translations of these modals with no explanation of how they are used in the context, and this could contribute to the misuse of modals such as *should* and make the learner sound relatively bossy and rude. Other researchers explore English textbooks for specific grades in China. Before presenting their findings, I will briefly outline the Chinese educational system and the use of English textbooks across different grades.

The Chinese educational system is organised through various stages, beginning with non-compulsory pre-school education for children aged 3 to 6. This is followed by compulsory education which includes six years of primary education (ages 6-12) and three years of junior high school (ages 12-15). After completing compulsory education, students can choose to take the Zhongkao, an entrance examination for senior high school, enrol in vocational school, or enter the workforce directly (Liu et al., 2024). Those who continue in senior high schools often prepare for the National College Entrance Examination (Gaokao), which determines their eligibility for higher education. Universities in China typically offer four-year bachelor's degree and three-year master's degree programs.

English language education generally begins in the third grade of primary school and continues as a compulsory subject throughout both primary, junior, and senior high

school. During these stages, English textbooks are mainly published and approved by the government to ensure standardisation across the educational system and across provinces, which will be further explained in Section 6.2.1.

In the junior high school English textbooks, both Li (2020) and Sun (2018) find that more emphasis is placed on the root use of *must*, *have to* and *should* rather than on their epistemic use. Additionally, Li (2020) notes that while epistemic use of *must* and *should* is introduced in Grade nine EFL junior high school textbooks in China, students in that grade do not use them in their writings, possibly due to the lack of detailed presentation and discussion. As mentioned above, Grade nine is the final year of this stage, crucial for students as they prepare for the senior high school entrance examination (Ye, 2021). In Grade nine, most students have studied English for six to nine years, and they are expected to have a basic to intermediate English proficiency, approximately aligning with the A2 level on the Common European Framework of Reference for Languages (CEFR; Council of Europe, 2020). This level indicates that students should be able to understand basic sentences and expressions related to immediate needs, communicate in simple tasks, and describe their background and environment in straightforward terms (Council of Europe, 2020).

An additional factor for the differences between the student groups could be the

cognitive readiness of learners to use epistemic necessity. Papafragou (1998) proposes that language learners tend to acquire the epistemic meaning of the modals later than their root meaning, indicating a developmental sequence in modal acquisition. This observation is particularly relevant under Pienemann's (1989) Teachability Hypothesis, which asserts that effective language acquisition is closely linked to cognitive development. According to Pienemann, students are only able to effectively use complex linguistic items, such as epistemic modals, once they have reached the necessary level of cognitive maturity.

Furthermore, learners tend to use modals that are first taught (Bai, 2015; Ma & Lu, 2007), leading to, for example, the over-representation of *can*. This is related to what Hasselgren (1994) described as the 'lexical teddy bear' (p. 237). Language learners rely on the devices that they are familiar with such as those learned early, which makes the learners feel safe. Tense and aspect in Chinese are expressed through particles instead of verb inflection. Similarly, modal verbs also have no inflected forms. As a result, learners may feel safe and comfortable when using modal verbs, potentially leading to an over-representation.

Cultural differences are also considered to play a role in explaining the different uses of modals by Chinese students and native speakers. An early work by Hinkel (1995)

finds that cultural and social value contribute to the root use of the modals in Asian students' writing. Kecskes and Kirner-Ludwig (2017) use different learner corpora, ICNALE and Treebank of Learner English (TLE), and affirm Hinkel's findings. Likewise, Bu (2011) finds the relation between the cultural perception of suggestions to the use of direct or indirect suggestion strategies. Chinese culture is considered to value harmony and collectivism due to the influence of Confucianism, and thus Chinese tend to have a stronger sense of responsibility and accountability to their nations (Bu, 2011; Li, 2016). This may lead to more use of direct suggestion strategies than native speakers, as observed by Bu. Cultural value could also explain the epistemic use of the modals. Yang (2018) points out that Chinese learners tend to make more confident and assertive statements compared to professional writers because certainty is considered in Chinese culture as a symbol of strength and hedging as a signal of weakness.

Based on the reasons proposed above, researchers have suggested several implications for pedagogical practices. In terms of first language influence, Yang (2018) recommends introducing students to the differences in modal use between their first and second languages to minimise first language interference. Li (2020) advocates for adding information such as potential errors caused by pragmatic differences between English modals and their Chinese equivalents in textbooks. As for the knowledge of

modals, it is suggested that additional devices for expressing modality be introduced to students, beyond just modals (Chen, 2012; Li, 2016; Yang, 2018). Chen (2012) also gives advice specific to epistemic modality, proposing that teachers should raise their awareness of the presentation of epistemic modality, and more exposure to this meaning in the textbooks is needed (Li, 2020). As for the last aspect, the cultural difference, Hinkel (1995) and Chen (2012) both point out that students should be introduced to the differences in cultures and how this may influence the use of modals.

2.5 Disciplinary variation in the use of modality

Before considering how disciplines influence the use of modality, it is necessary to first define and classify disciplines. Turner (2000) focuses on the organisational aspect, defining discipline as ‘kinds of collectivities that include a large proportion of persons holding degrees with the same differentiating specialization name, which are organized in part into degree-granting units that in part give degree-granting positions and powers to persons holding these degrees‘ (p. 47). Whitley (2000) offers a similar definition but with less emphasis on the degree-granting, viewing discipline as ‘units of labour market control which trained knowledge producers in particular skills that monopolized contributions to particular intellectual goals’ (p.81). Hammarfelt (2019) provides a comprehensive overview of the concept of discipline within academic contexts and presents different qualities of a discipline such as shared norms, international

recognition, and control over the dissemination of knowledge.

In terms of the classification of disciplines, Becher and Trowler (2001) propose four basic dimensions along two axes: soft-hard and pure-applied. These dimensions differ in objectives, methodologies, and engagements with knowledge creation and application. Soft disciplines, typically including Social Sciences and Humanities, often rely on qualitative methodologies and interpretations of human activities to deal with complicated variables. By contrast, hard disciplines, such as Physics, tend to establish principles based on empirical data collected primarily through quantitative and experimental approaches. Similarly, while disciplines such as Mathematics are predominantly categorised as pure, focusing mainly on theoretical explorations to generate new knowledge without a direct focus on practical applications, applied disciplines such as Medicine are oriented towards solving real-world problems. As Squires (2005) notes, applied disciplines tend to prioritise 'acting rather than knowing' (p. 130) compared to pure disciplines.

The classification, however, is not definite. One discipline may cover features of two distinct dimensions and are placed in the middle ground (Becher & Trowler, 2001). One example is Economics, which integrates both the hard and soft elements across its subfields. For instance, while macroeconomics often aligns more with the hard

sciences due to its quantitative approaches, development economics tends to incorporate soft science elements by addressing social and cultural factors. Along the pure-applied continuum, Economics is often seen as a predominantly pure discipline because it deals with abstract mathematical theories. However, it also includes applied aspects, such as applications to industrial organisations (Becher & Trowler, 2001). This blending highlights the complexity of categorising disciplines strictly as either soft or hard, and applied or pure, suggesting that such distinctions should be viewed more as a spectrum than a binary and that identifying these nuances requires thorough, case-by-case analysis.

The following paragraphs review disciplinary variations in modality studies, where researchers often classify certain disciplines along two axes of the four dimensions mentioned above. This classification does not imply that the disciplines mentioned in any study are definitively placed within those mentioned dimensions. Rather, it indicates that in that specific study, based on the texts analysed, the discipline exhibits more features characteristic of one side of the dimension than the other. Similarly, in my analysis, while I argue that Business and Management and English Literature could both be categorised as soft disciplines, this does not mean they exhibit no features of hard disciplines. Instead, it indicates that after a thorough examination, the predominant features in the learner texts align more with those of soft disciplines. A

more detailed discussion will be presented in Section 3.2. Additionally, qualitative analysis is conducted to provide an in-depth examination of disciplinary features and their contributions to the use of modals.

There is extensive research into disciplinary variations of modality in research articles. However, research into modality in academic writings of Chinese EFL students has been confined primarily to either single discipline or argumentative essays on general topics (see Section 2.4.2). The focus has largely been on comparing the learner corpus with the reference corpus within similar disciplines, rather than investigating variations across different disciplines. Therefore, this section will explore the case in the research articles in general, hoping to provide some insights into the discussion regarding the disciplinary differences in Chinese students' academic writings.

Studies of modality in research articles across disciplines have a preference to look at the epistemic sense of modality, with a particular emphasis on their pragmatic roles, which will be addressed initially. There is only a handful of studies that examine the root use of the modals, which will be reviewed towards the end of the sub-section.

The number of disciplines examined for the use of epistemic modals in research articles spans from one (e.g., Vičič & Petek, 2016; Yang et al., 2015) to as many as

twelve (e.g., Peacock, 2014). The disciplines selected mostly make distinctions between the soft and hard disciplines. Additionally, some research incorporates one more category that combines characteristics of both soft and hard disciplines (e.g., Rizomilioti, 2006; Vázquez & Giner, 2008). This preference in the comparison between soft and hard disciplines may stem from the fact that they are largely distinct and thus more differences in modality can be identified.

Sameri and Tavangar (2013) suggest that epistemic modality appears to be used more frequently in disciplines typically categorised as soft, such as Philosophy and Applied Linguistics, compared to those often classified as hard, like Physics and Chemistry. This is also confirmed in Farrokhi and Emami (2008), in which the modal use is compared between Applied Linguistics and Electrical Engineering. Two sub-types of epistemic modality, hedges and boosters, are examined separately in these two studies. It is found that hedges tend to be over-represented in soft disciplines compared to the hard ones, which is also confirmed by Vázquez and Giner (2008) who investigate research articles in Marketing, Biology, and Mechanical Engineering. The distribution of boosters in different disciplines, on the other hand, shows differences among studies. Sameri and Tavangar find an over-representation of boosters in Physics and Chemistry, whereas Farrokhi and Emami's observation shows an opposite pattern, with more boosters used in Applied Linguistics texts they investigated.

As mentioned above, most of the studies compare disciplines that are distant in dimensions, but few studies examine disciplines that are both categorised as, for example, soft disciplines. To fill in this gap, the present study focuses on two soft disciplines, Business and Management and English literature, which are characterised by Nesi and Gardner (2012) as the disciplinary groups of Social Science and Arts and Humanities respectively (see Section 3.2 for discussion). I will first discuss the use of modals across these two disciplinary groups in the following paragraphs and move on to review studies that examine similar disciplines to what the present study focuses on.

Takimoto (2015) finds that both hedges and boosters are used more frequently in disciplines categorised as Humanities (Linguistics and Philosophy) and Social Science (Marketing and Sociology) than in those classified as Natural Science (Physics, Electrical Engineering, Mechanical Engineering, and Chemistry) in this study. This indicates that the former two disciplinary groups tend to share similar features in the use of epistemic modality. However, when taking a closer look at this study, boosters are used more often in Linguistics and Philosophy than in Marketing and Sociology, whereas the distribution of hedges is similar in these disciplines. This is similar to Rizomilioti's (2006) findings that writers in Literary Criticism (represented as Humanities) use the highest frequency of boosters, followed by those in the fields of

Archaeology (represented as the combination of Social Science and Humanities) and Biology (represented as Science).

As for the reasons for the different usage of epistemic modality across disciplines, Panocová and Lukačín (2019), who examine the discipline of Medicine and Humanities in the academic sub-corpora of COCA, argue that it might result from the aims and the analytical approaches of the disciplines. Disciplines in Natural Science mostly aim to present the results objectively and impersonally, and thus writers use fewer epistemic devices to express personal judgements. By contrast, writers in Humanities and Social Science prefer to interpret the material by persuasion and tend to be more speculative and less abstract than their counterparts in Natural Science. This is corroborated by Takimoto (2015) and Vázquez and Giner (2008), with the latter additionally arguing that the nature of the data also plays a role. Vázquez and Giner exemplify that writers in Marketing tend to rely on abstract data such as human behaviours, whereas those in Mechanical Engineering mostly base their arguments on precise data, and possibly require less use of hedges. Peacock (2014) adds another reason, that is the language convention of research articles, such as the kinds of arguments the discipline values, what the reader's previous knowledge is, and how they would like to be persuaded. Apart from the constraints specifically for research articles, individual differences such as writers' personality, writing style, and experience could also have an impact

(Takimoto, 2015).

As for the disciplinary variation in the use of obligation modals, there are only a handful of studies, and Giltrow (2005) is one of them. She examines the use of deontic modals and other expressions in three disciplines, Forestry, Social Psychology and Urban Geography. One distinctive feature of Urban Geography is quoting sources outside research contexts such as organisations or spokesmen. Rather than giving suggestions directly, the writers use this quoting technique to demonstrate the endorsement of the recommendation. Giltrow further notes that writers in Social Psychology describe obligation less frequently compared to the other two disciplines partly due to the missing of organisations or institutions that could take the suggestions. A more related study is conducted by Dontcheva-Navratilova et al. (2020), who examine persuasion strategies in Economics and Linguistics, categorising these two disciplines as Social Science and Humanities, respectively, both of which are considered predominantly soft disciplines. Writings in Economics are observed to be more objective and include a higher frequency of root modals compared to those in Linguistics possibly due to the difference in discipline-specific knowledge production.

The paragraphs above have reviewed the general findings of modals used across disciplines, particularly focusing on those that express epistemic necessity and

obligation. I will now move on to look at the studies in which the disciplines examined are similar to my focus, Business and Management and English Literature. Let us start with business-related disciplines. Peacock (2014) identifies a higher frequency of obligation and necessity modals (*should*, *must*, *need to*, and *have to*) in Business compared to the other four disciplines investigated, including Language and Linguistics, Public and Social Administration, Economics, and Psychology. Likewise, Millán (2008) explores a different set of disciplines and finds that boosters are found to be more frequently used in Business Management than in Food Technology and Urology. Peacock proposes one reason, suggesting that writers in Business tend to make stronger arguments and show greater commitment to their arguments, a tendency that is associated with the topics they discuss.

As for the modality used in English literature, there is no equivalent discipline that is investigated in the previous literature. The most approximate discipline is Literary Criticism. An early work is conducted by Simpson (1990), who reports that writings of the literary critic, F. R. Leavis, are highly assertive since some bold statements are written without modalised devices, as in 'The great English novelists are Jane Austen, George Eliot, Henry James and Joseph Conrad—to stop for the moment at that comparatively safe point in history.' (p. 75). Although Simpson's analysis is limited to a single text from one critic, it provides insights into the distinctive uses of modals in

Literary Criticism. Following his study, Piqué-Angordans et al. (2001, 2002) examine more texts in this discipline in terms of modal use and make comparisons with research articles in other disciplines. Both studies find that epistemic and deontic sense of the modals are combined to be used in Literary Criticism, whereas writers in Medicine, Biology or Health Science favour the epistemic sense. Piqué-Angordans et al. (2002) ascribe this variation to the disciplinary convention of persuading readers. While writers in Science disciplines primarily concentrate on the data to make arguments, writers in Literary Criticism tend to combine the presentation of data and examples in the literary works. Another study conducted by Rizomilioti (2006), as mentioned before, concludes that writers in Literary Criticism use the fewest downtowners and the most boosters compared to those in Archaeology and Biology. Although literacy criticism is different from my target discipline, English literature, in that it might require more critical analysis, this discussion can still provide some meaningful insights.

As you may notice from the studies discussed above, the disciplines investigated overlap to some extent, and we encounter some disagreements in the findings of similar disciplines. One reason for the inconsistency between studies is the different selection of the specific disciplines. Even if the journal articles are within the same discipline, the type of research (e.g., based on data or theory) might also contribute to the difference in the use of modals. Results in one particular discipline cannot

necessarily be generalised to the whole disciplinary group (Rizomilioti, 2006). Nevertheless, the more disciplines in the continuum are investigated, the more comprehensive our knowledge about modal use in academic writing becomes.

The extensive studies of the disciplinary variation in the use of modals in research articles could give us some insight into how student's academic writing might differ in various disciplines. However, research articles and students' academic writing show differences in several aspects. Research articles are targeted at the community members in the discipline, which require the researcher to build solid arguments and persuade the readers to agree with them (Hyland, 2004). By contrast, the academic writings of undergraduate students, who are the target students in the present study, aim to meet the criteria and get a high grade. Although students are also a part of the disciplinary community, they are not expected to reach the same analytical level as the researchers who publish journal articles. The findings of the disciplinary variation across disciplines in the research articles cannot be assumed to be the results we will find in students' academic writing.

2.6 Contrastive Interlanguage Analysis

Corpus linguistics has received attention in second language research. The emergency of learner corpus research in the early 1990s attempts to link these two

fields, applying corpus linguistics methods to the study of language learning and teaching. Learner corpora are defined by Granger et al. (2015, p. 1) as ‘electronic collections of natural or near-natural data produced by foreign or second language learners and assembled according to explicit design criteria’.

Granger (2012) operationalises the naturalness of data by applying Ellis's (1994) three-fold classification of second language acquisition data: natural language, clinical, and experimental, and argues that learner corpus data generally falls between the first two types. According to Granger, natural language data refer to the naturally occurring data produced by learners for authentic communicative purposes, and they are difficult to collect especially in EFL settings. Thus, researchers tend to explore the clinical data, such as written compositions or descriptions of a picture. The naturalness of learner corpus data, as Granger (2012) suggests, should be viewed on a continuum, not as a dichotomy. The learner corpus in the present study comprises undergraduate dissertations written by Chinese EFL students (see Section 3.2.1 for details). Compared to the International Corpus of Learner English (ICLE; Granger, 2020) which primarily includes texts for language practice, it could be argued that the dissertations in this corpus align more closely with natural language data, as they are authentic assignments produced for assessment purposes. The motivation for Chinese EFL students in producing these texts is to achieve high marks and meet academic

requirements. Their writings showcase the application of language within an educational framework.

Granger (2015) also emphasises the importance of design criteria in her definition of learner corpora and concludes that these criteria are related to three main types of variables: the environment, the task, and the learner's profile. These variables are suggested to be thoroughly evaluated before the compilation of a learner corpus to ensure the relevance and utility for the intended research objectives.

One of the earliest and most influential learner corpora is the International Corpus of Learner English (ICLE; Granger, 2020), developed by Granger and her colleagues at the University of Louvain over the last thirty years. It contains essays written by learners at upper intermediate and advanced levels and is published in three versions. The latest version was released in 2020 containing over 5.5 million words with writings collected from 25 first language backgrounds.

Following this project, more learner corpora are compiled with a wider variety of languages other than English and more registers. To name a few, Arabic Learner Corpus consists of written and spoken materials from learners of Arabic from 66 different first language backgrounds (Alfaifi et al., 2014). The Jinan Chinese Learner

Corpus is a collection of 9,000 Chinese texts produced by L2 learners, and these texts are taken from exams and assignments (Wang et al., 2015). A comprehensive list of learner corpora is provided by the Centre for English Corpus Linguistics (2024) at the University of Louvain. In addition, tools are developed due to the growing availability of corpora, such as ColloCaid (Lew et al.,2018), which is a real-time text-editing tool to provide suggestions on collocations for writers of academic English. Its collocational data is from academic English and academic vocabulary lists. More recently, researchers have started to emphasise the importance of standardising the compilation of learner corpus (e.g., Frey et al., 2020) and promoting interoperability among different corpus tools and platforms so that the findings can be generalised and compared effectively. For example, Larsson et al. (2020) and Granger et al. (2022) both highlight the importance of the transparency of instruments used in the learner corpus research. Larsson et al. recommend guidelines for documenting these instruments, including the coding scheme for linguistic features, while Granger et al. publish the Louvain error tagging manual, providing detailed explanations of error tags with authentic examples.

With learner corpus data in hand, the framework to utilise the data to achieve research goals undergoes development, one of which is Contrastive Interlanguage Analysis (CIA; Granger, 1996). This is the fundamental approach used in the present study.

Before moving on to discuss CIA, we shall look at Contrastive Analysis (CA), which is an approach that CIA is derived from. CA is mostly used in translation studies to examine bilingual corpora. It includes two types of research: one is to compare the original texts and their translations, and the other is to compare the parallel original texts, such as newspapers published in more than one language (Granger, 1996). Emerging from the late 1980s, learner corpora provide an extensive dataset applicable for research in interlanguage studies. CA, while acknowledged as a method for data analysis, has certain limitations. The approach focuses on how the same meaning is expressed in different languages, yet it neglects the crucial aspect of contrasting the native language with the interlanguage. Selinker (1989) points out this oversight and calls for a new-developed CA.

In response to the criticism of CA for interlanguage studies, Granger (1996) proposes CIA, a theoretical framework that facilitates the comparison between the native and learner variety of the same language. This approach inherently adopts the concept of interlanguage as a distinct linguistic system from the target language system (Selinker, 1972). Selinker (2014, p. 223) gives a working definition of interlanguage as the ‘linguistic/cognitive space that exists between the native language and the language that one is learning’. In addition, Selinker (1972) proposes that the learners activate a latent psychological structure when producing a second language, which consists of

five central cognitive processes: language transfer, transfer of training, learning strategies, communication strategies, and overgeneralisation of target language linguistic material. The first two processes have been briefly discussed in Section 2.4.2. Regarding learning and communication strategies, the former involves methods used by learners to master the target language, while the latter addresses ways to effectively communicate despite gaps in language knowledge. Overgeneralisation of target language is related to the inappropriate extension of language rules to sentences where they do not apply, such as adding the past tense marker *-ed* to all verbs, regardless of exceptions.

These processes contribute to the ongoing development and modification of the interlanguage. However, Selinker (1972) also introduces the concept of 'fossilization' (p. 215), suggesting that certain features of learners' interlanguage may become fixed and resist further modification, despite continued exposure to the target language. There are disagreements in terms of the inevitability of fossilization, as pointed out by Han (2004). Selinker acknowledges that a small percentage of learners (approximately 5%) may overcome fossilization and achieve proficiency indistinguishable from native speakers, suggesting that the permanency of interlanguage is not inevitable for all learners.

Selinker's (1972) theory positions interlanguage primarily as an individualised, psychological construct, distinctly varying across learners and shaped by their personal experiences and cognitive processes. Nonetheless, he acknowledges that learners may exhibit common interlanguage patterns, indicative of systematic characteristics that learners with similar backgrounds, such as their first languages, might share. This observation is integral to CIA (Granger, 1996), which primarily focuses on the first language as a significant factor influencing interlanguage. However, with advancements in second language acquisition research and further discussions on CIA, a revised version, CIA^2 (Granger, 2015), has been introduced. This updated framework incorporates task and learner variables, highlighting the necessity of exploring a broader range of factors influencing interlanguage beyond simply the learners' first languages.

Having discussed what interlanguage is, the following paragraphs will explain CIA and its development. We will first explore CIA, which is divided into two types, as illustrated in Figure 2.4.

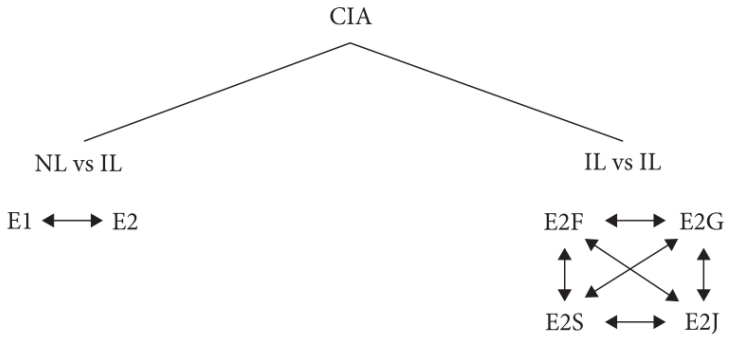


Figure 2.4 Contrastive Interlanguage Analysis (CIA; Granger, 1996, p. 44) (NL = native language; IL = interlanguage; E1 = English as a first language; E2 = English as a foreign language; E2F = English of French learners; E2G = English of German learners; E2S = English of Swedish learners; E2J = English of Japanese learners)

One type of CIA is to compare the native language (NL) and the interlanguage (IL), which helps to identify the different uses of linguistic items and have a deeper understanding of the profiles of the interlanguage. The other type is to compare different varieties of interlanguages, such as English of French learners (E2F) and Japanese learners (E2J). This can help to examine the cross-linguistic transfer.

Granger (1996) and Gilquin (2000/2001) also suggest an integration of CA and CIA because of the interrelationship between bilingual corpus and learner corpus, proposing the Integrated Contrastive Model (ICM). The revised version proposed by Gilquin is shown in Figure 2.5.

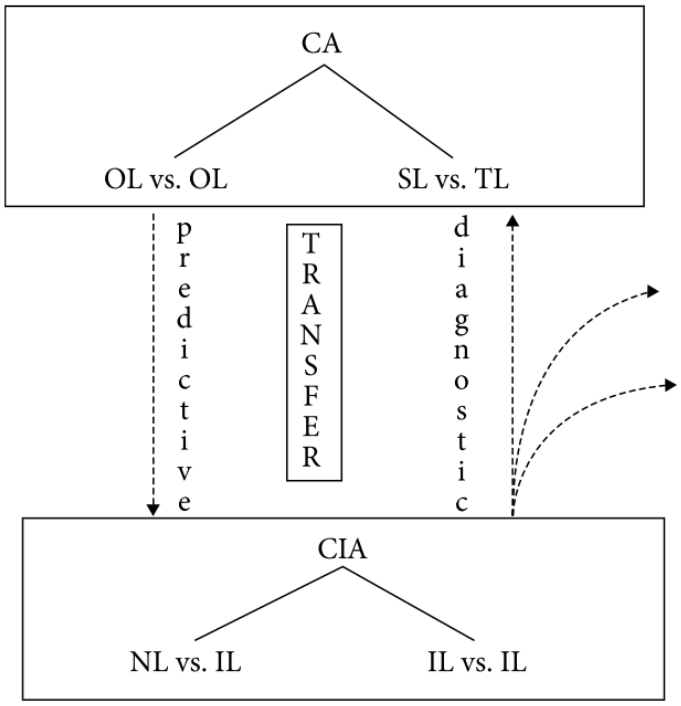


Figure 2.5 The Integrated Contrastive Model (ICM; Gilquin, 2000/2001, p. 100) (based on Granger 1996, p. 47) (CA = Contrastive Analysis; OL = original language; SL = source language; TL = translated language; CIA = Contrastive Interlanguage Analysis; NL = native language; IL = interlanguage)

In this model, the predictions formulated by CA can be checked through the examination of CIA. Conversely, CA can help interpret and support the diagnostic evidence derived from CIA. However, there are also results of CIA that cannot be supported by CA. These findings are represented by arrows pointing out, added by Gilquin (2000/2001), and she also changes Granger's (1996) version of ICM by using broken lines instead of solid ones to show a weaker connection between CA and CIA.

Although widely used in learner corpus research, CIA is mainly criticised from two

perspectives. Firstly, Bley-Vroman (1983) addresses the concern regarding the exclusive use of the target language as a benchmark for evaluating learner language, which might neglect the distinct and systematic features of interlanguage as an independent linguistic system and lead to what is identified as a comparative fallacy. While Paquot (2007) disputes this by advocating for the validity of first and second language comparisons in second language acquisition research, this criticism alerts researchers not to consider interlanguage as a deficient language compared to the native language.

Another criticism of CIA is that it only presents one variety of reference corpus, the native language. Granger (2015) notes that this criticism results from the emergence of studies regarding World English and English as Lingua Franca. In response to that, she proposes the CIA^2, in which the reference variety extends from the native language to others, such as competent second language data. The framework is shown in Figure 2.6.

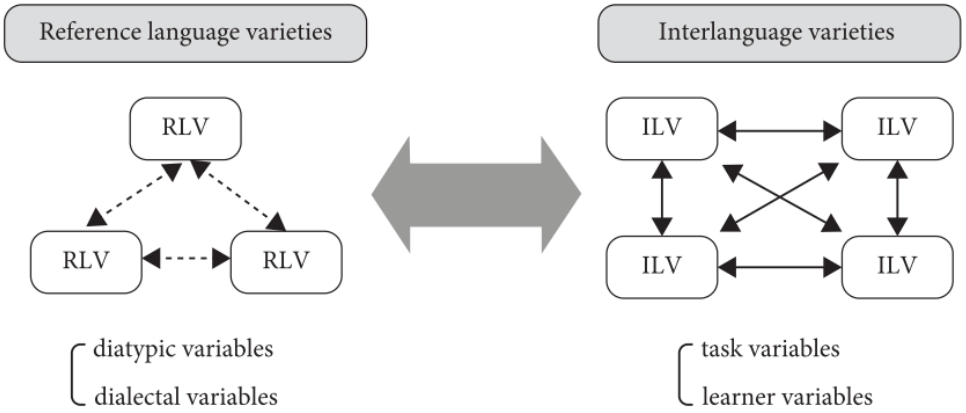


Figure 2.6 CIA[2] (Granger, 2015, p. 17) (RLV = reference language varieties; ILV = interlanguage varieties)

Apart from this change, Granger (2015) also proposes to use *under-represent* and *over-represent* instead of *underuse* and *overuse* to avoid the negative sense of the terms. *Under-represent* and *over-represent* are not to propose the undesirable usage of the learners' language compared to that of native speakers. Instead, it is used neutrally to indicate the frequency differences with no pre-assumed judgement. This corresponds to the principle proposed by Selinker (2014, p. 230) that interlanguage 'must be described in their own terms'. The over- or under-representation of specific forms in the learner corpus should not be considered as errors but as characteristics presented in the interlanguage. In other words, an under-representation of a form in the learner corpus compared to the native speaker corpus, for example, does not necessarily mean that learner's use of this form is deficient and requires further development. This can be totally acceptable and even desirable, according to Granger.

Following the above statement, the present study uses the two terms, *under-representation* and *over-representation*, descriptively with no previous assumption of what should be desirable when compared with the reference corpus. In addition, although the main framework used in the present study is the first version of CIA, I also include the notion proposed in CIA^2, such as the two terms mentioned above. Thus, in this thesis, CIA is used to refer to both CIA and CIA^2. The specific applications and the criteria for selecting the reference corpus will be presented in Section 3.2.

2.7 Summary

This chapter has surveyed literature on five aspects, including modality in general, modality in academic writing, and how modality is used by Chinese students and across disciplines, as well as the approach of CIA. Through the discussion of different classification frameworks of modality, I have justified the selection of the binary framework in the present study, classifying modality into epistemic and root sense. It has been reported that there is a dearth of studies regarding the epistemic necessity and obligation modals in student academic writing, and thus the review of literature was expanded to consider other expressions that share similar pragmatic functions of modality (e.g., boosters and devices of command strategies) and to examine the findings of research articles to provide some insights.

Methodologically, most studies focus on the frequency and meaning distribution of the modals. There are few studies systematically examining the main verbs used with the modals, although these verbs could contribute to the meaning of the modal as well as its strength. In addition, the learner corpus research discussed in Section 2.4.2 highlights a lack of corpora available for examining the modality in Chinese EFL students' discipline-specific academic writing. Most of the learner corpus used is comprised of argumentative writing covering various topics. Thus, based on a recently compiled Chinese learner corpus of academic writing with two disciplines covered, this research explores the meaning and the main verbs collocating with *must*, *have to*, and *should*, aiming to offer a more comprehensive profile of modals expressing epistemic necessity and obligation. It also investigates how first language and discipline impact the use of modals in Chinese students' academic writing.

In the chapter that follows, I will describe the methodology used in the present study, covering the selection and compilation of data, as well as providing an overview of the analytical approaches.

3 METHODOLOGY

3.1 Introduction

The previous chapter has discussed how Contrastive Interlanguage Analysis (CIA) can be used to investigate the EFL learner's language. This chapter elaborates on how CIA is applied in the current investigation, detailing the data compilation and providing an overview of the analytical approaches used.

To recap, the study is designed to explore the profiles of *must*, *have to*, and *should* in Chinese EFL students' academic writing and to examine the influence of first language and discipline on modal use. This can be broken down into five research questions:

RQ 1: How frequently are *must*, *have to*, and *should* used in the Chinese EFL learner corpus and the reference corpus?

RQ 2: How are the meanings of the three modals distributed?

RQ 3: What semantic patterns can be identified regarding the main verbs that collocate with the three modals?

RQ 4: Do the profiles of the three modals differ between the two student groups, Chinese and British students?

RQ 5: Do the profiles of the three modals differ between the disciplines?

The profiles of modals are investigated in three main aspects, which are illustrated by the first three research questions, including frequency, meaning distribution and verb collocates. The latter two aspects, often overlooked in previous literature on modal use in Chinese EFL student academic writing, can contribute to a more thorough understanding of how *must*, *have to*, and *should* are used in their writing. The last two research questions focus on the two factors that may influence modal use, first language and discipline, and they are discussed in terms of the three aspects mentioned above.

As mentioned in Section 2.6, learner corpus research links corpus linguistics with the study of second language acquisition, which helps to examine interlanguage quantitatively and reveal unexpected patterns that may omitted by researchers if they only examine a small number of texts. As an increasing number of learner corpora become accessible, Granger (1996) proposes CIA as the framework for data analysis. One of its types, the comparison between the native language and interlanguage, has been widely applied by previous studies in the investigation of modals by Chinese EFL students, and it is also used in the present study. The outcomes can imply areas where the first language accounts for deviations in EFL students' production from that of native speakers.

The structure of this chapter is as follows. In Section 3.2, the selection of the learner and the reference corpus is justified. It is then followed by the procedures of compiling the final data and how the target modals are selected and extracted. Section 3.4 explains the overall analytical approaches, and the last section summarises the chapter.

3.2 Data

To address the research questions using CIA, a learner and a reference corpus are used for examination. The texts for the learner corpus are sourced from the Chinese Advanced English Learner Corpus of Academic Written English (CAEL-CAWE; Zou, 2018), while those for the reference corpus are drawn from the British Academic Written English (BAWE; Nesi et al., 2008) corpus. This section explains the rationale behind choosing these two corpora and describes the extraction of comparable texts to compile the learner and reference corpora.

3.2.1 The learner corpus

To identify an appropriate learner corpus, two primary strategies are employed: consulting archival websites and reviewing prior literature. Centre for English Corpus Linguistics (2024) at the University of Louvain lists a comprehensive collection of learner corpora around the world, including information such as target language

(mostly English), first language, medium (spoken or written), text type/task type (e.g., dissertations and argumentative essays), proficiency level, size in words, project director and availability. This list allows researchers to filter various categories to align with the objectives of their studies. In my case, the filtering process undergoes two steps. An initial selection was made according to categories including target language (English), first language (including Chinese), medium (written), and text type (academic essays or dissertations written by university students). There are fourteen learner corpora that meet the criteria mentioned above, and their descriptions were copied from the website in a spreadsheet.

Following this, each learner corpus was further searched for other information that is not provided on the website, such as the year of compilation, the site where the data is collected from, and the metadata provided. These were added in the spreadsheet as well as my comments, and three out of the fourteen corpora are shown in Table 3.1 below for illustration. The first ten columns were copied from the website, except for the second column in which I assigned a score ranging from 1 to 5 based on their relevance to the present study. The remaining five columns to the right were added based on the results of the second round of searching, and I also commented on their suitability.

Table 3.1 Learner corpora candidates and their information (three examples provided for illustration)

Corpus	Relevance	Target language	First language	Medium	Text type / task type	Proficiency level	Size in words	Project director	Availability	Year	Notes	Data collected from	Metadata	Extra information
Chinese Learner English Corpus (CLEC)	3	English	Chinese	Written	Daily writing assignments and writing examinations completed	Chinese senior high school students and university students	c. 1.2 m.	Gui Shichun and Yang Huizhong	Book and CQPWeb in BFSU	2003	not up-to-date	Universities in Mainland China	Sub-corpora: St2 to St6	30-minute time span, error tags
The International Corpus Network of Asian Learners of English (ICNALE)	4	English	ESL regions (Hong Kong, Pakistan, the Philippines, and Singapore) and in EFL regions (China, Indonesia, Japan, Korea, Taiwan, and Thailand)	Written and spoken	Colledge Student essays-200-to-300-words essays about two ICNALE common topics ("a part-time job for college students" and "non-smoking at restaurants"). Learners were given 20-40 minutes to write one essay.	A2, B1_1 (B1 low), B1_2 (B1 high), and B2+	c. 1,8 m.	Shin'ichiro Ishikawa	http://language.sakura.ne.jp/icnale/	2010-2019	All metadata is well-documented. There are reference corpora. It is possible to compare different registers (written and spoken) within the same topics.	Universities in Asian	e.g. Basic Attributes, Motivation, L2 learning background	Apart from written corpus, there are also spoken monologue corpus, dialogue corpus and edited written corpus with the same topics
The Spoken and Written English Corpus of Chinese Learners (SWECCL) 1.0 and 2.0	3	English	Chinese	Written (WECCL) and spoken (SECCL)	Written: argumentative and narrative essays	Chinese university students	c. 2 m.	Wei Qiufang	Book with CD-rom and CQPWeb in BFSU	2008	not up-to-date	Universities in Mainland China	e.g. Proficiency levels	

These Chinese learner corpora can primarily be classified into two types: corpora specific to Chinese learners and learner corpora that encompass components of Chinese EFL writing. Those only consist of Chinese students' academic writing, such as CLEC and SWECCL, have been widely investigated (see Section 2.4.2). However, since these corpora consist of argumentative essays on general topics, their examination does not allow for the identification of disciplinary variation in modal usage. This non-discipline-specific feature can also be observed in other learner corpora in which writers have different first language backgrounds and Chinese student writing is one of the sub-corpora. For example, ICNALE, as briefly introduced in Section 2.4.2, consists of essays of Asian EFL/ ESL students' writing on common topics such as part-time jobs and smoking in restaurants. ETS Corpus of Non-Native Written English (Blanchard et al., 2014) comprises writings from the Test of English as a Foreign Language, covering eight topics by speakers of 11 different native languages other than English.

There are other corpora that collect university essays in specific disciplines, such as MICUSP and BAWE. However, the number of texts written by Chinese students is limited. For instance, MICUSP includes less than ten texts per discipline written by non-native speakers. This is because these corpora were compiled not to examine EFL student's writing but for different purposes, such as exploring disciplinary and genre

differences in academic writing.

Given that none of the fourteen learner corpora listed in the spreadsheet satisfactorily provide the data needed to answer the research questions, an alternative approach is taken, searching corpora used in previous literature to investigate Chinese EFL student academic writing. For example, the corpus used by Yang (2018) is not included in the list mentioned above, and it consists of Chinese EFL undergraduates' writings in one discipline, Business and Trade. Most Chinese learner corpora lack writings from various disciplines, primarily because in Mainland China, many fields of study do not require the submission of assignments in English. However, there are exceptions, such as programs in English studies or other disciplines that require English communication in professional practice, such as Business and Trade mentioned above. CAEL-CAWE, which includes two such disciplines, was compiled by Zou (2018) to analyse the academic writing of Chinese university students. This corpus was chosen for the present study as the source of the learner corpus. The subsequent paragraphs will provide an overview of this corpus and justify the selection of texts for analysis.

CAEL-CAWE consists of 456 texts and 4,193,413 tokens. The texts are dissertations written in English by Chinese undergraduate and postgraduate students in two disciplines, Business and Management (BM) and English Literature (EL), submitted as

part of their degree completion requirements. Correspondingly, four sub-corpora are included, namely undergraduate English Literature (UGEL), postgraduate taught English Literature (PGEL), undergraduate Business and Management (UGBM), and postgraduate taught Business and Management (PGBM). Writings were collected in Mainland Chinese universities by Zou (2018) and her colleagues. I got her permission to use the corpus and received the clean-up texts via email. These texts contain the main body of the dissertations and exclude information such as abstracts, footnotes, and titles. According to Zou, quotes from a third party (e.g., words from the literary works that are analysed in the dissertation) in EL were also deleted. This is to exclude other people's writing from students' dissertations. These long quotes were identified and replaced with '[quote]'. Table 3.2 is the meta-information of CAEL-CAWE reproduced from Zou's thesis, and I added a column showing the average tokens per text.

Table 3.2 Meta-information of CEAL-CAWE (before text selection for the learner corpus) (Zou, 2018, p. 65)

	Number of texts	Total tokens	Average tokens per text	Total types
UGBM	**168**	**991,185**	**5,900**	**23,301**
UGEL	**137**	**767,648**	**5,603**	**24,796**
PGBM	73	1,073,130	14,700	21,890
PGEL	78	1,361,450	17,454	37,818
Total	456	4,193,413	9,196	107,805

Despite writings in both educational levels being dissertations, the length of postgraduates' texts is nearly three times longer than those of undergraduates, which may influence the absolute frequency of the modals. Not only does the length of texts differ, but other aspects such as dissertation requirements may also show variations. In addition, educational level is potentially related to proficiency level, which is mentioned in Section 2.4.2 to be an influential factor in modal use (e.g., Gao, 2023; Hu & Li, 2020). Since this factor is not the focus of the present study, I only examine texts at one of the education levels, the undergraduate writing, which is decided due to the consideration of comparability to the reference corpus (see Section 3.2.2 for discussion).

Therefore, the learner corpus (LC) used in the present study comprises two sub-corpora, UGBM and UGEL. For clarification and consistency with the names of the reference sub-corpora later on, I renamed them as LC-BM and LC-EL respectively. The texts were uploaded to Sketch Engine (Kilgarriff et. al., 2014) for analysis (refer to Section 3.3 for justification of using this platform). What is worth noting is that the token size displayed in Sketch Engine differs from that reported by Zou (2018) in Table 3.2, possibly due to different definitions of tokens. Since Zou did not clarify how she defined a token, this study will use the token count in Sketch Engine as the reference, aligning

with the practice for the reference corpus. Table 3.3 shows the information of the learner corpus, including its number of texts, average tokens per text, and the token size of the sub-corpora. There are slightly more texts in LC-BM than in LC-EL, but their average tokens per text are similar.

Table 3.3 Number of texts and tokens per sub-corpus of the learner corpus (after text selection from CAEL-CAWE and based on the token counts provided in Sketch Engine)

	Number of texts	Average tokens per text	Total tokens
LC-BM	168	6,475	1,087,833
LC-EL	137	6,045	828,184
Total	305	6282	1,916,017

3.2.2 The reference corpus

In order to follow CIA, a comparable reference corpus to the learner corpus is required. BAWE is selected as the source of the reference corpus, which was created from 2004 to 2007 and contains 2,897 texts of proficient students' writings collected from the Universities of Warwick, Oxford Brookes, and Reading. These writings are distributed across four broad disciplinary groups (Life Sciences, Physical Sciences, Arts and Humanities, and Social Sciences) and four academic levels (undergraduate Years 1-3 and postgraduate taught levels), and genres of the writings are also documented. This section will justify the selection of the reference corpus and describe the extraction of comparable texts.

Gilquin (2022) concludes two criteria that guide the selection of a reference corpus: the fit to the research purposes and the comparability to the learner corpus. Since one of the focuses of the present study is to explore how first language influences the modal use in academic writing, a reference corpus containing native English speaker's writing would be appropriate. In addition, the aim is mostly pedagogical and thus the reference corpus should reflect what the students need or aim to approximate. BAWE contains writings that could be the achievable goal of Chinese EFL students and most of the texts are written by native English speakers.

The second criterion, the comparability to the learner corpus, is related to various factors, and Gilquin (2022) discusses three of them, regional variety, level of literacy, and text types. These factors can be comparable between the learner corpus and BAWE, provided that appropriate text selection is carried out. They will be discussed in the paragraphs below as well as other more nuanced factors (educational background, educational level, and macrostructure).

The regional variety of English for the reference corpus is considered in two aspects: the language learners are exposed to and the aim of the learners. One of the formal sources of language exposure is textbooks. Junior high school English textbooks in

China have eight editions used in different provinces (see Section 6.2.1 for details), and they focus on either British or American English since the Chinese Ministry of Education (2020) does not specify which variety of English is preferred. As students may also have informal input such as social media and television series, I would argue that generally Chinese students are exposed to a mix of British and American English. As for the learners' expectations and aims, Wei (2016) points out that although Chinese university students acknowledge the diversity of English, their recognition remains somewhat narrow. Specifically, they consider native English varieties, particularly those from Britain and America, as the criterion for 'good English' (Wei, 2016, p. 106). This preference influences several facets of English language education in China, including assessment and learning standards.

Given the above facts about English teaching and learning in China, British and American English are chosen to be the target varieties of English for reference. BAWE was compiled from UK universities, suggesting that students whose first language is English in the corpus are likely to be the users of British English. What is worth noting is that the first language is reported by the students in BAWE and the variety of English is not explicitly stated. Consequently, an additional criterion is needed, focusing on the educational background. Only texts from students who identify English as their first language and have completed their secondary education entirely in the UK are

considered. This ensures that these students share similar language and educational backgrounds, and their writings are what the Chinese students aspire to achieve.

In terms of the level of literacy, the reference corpus could be chosen from novice writing or expert writing, which varies in the writer's experience in producing texts similar to those in the learner corpus (Gilquin, 2022). One example of the novice writing corpora is LOCNESS, which is discussed in Section 2.4.2 to be frequently used as the reference corpus. LOCNESS consists of essays written by native English speakers on various topics and is designed to be comparable to ICLE. Its advantage being the reference is that the students have native knowledge of English and are at a similar cognitive level to Chinese undergraduates as they are at a similar age. However, their writing is not used for assessment, and thus the motivation behind these texts might not aim at showcasing the highest level of writing skill, possibly affecting the overall literacy quality. By contrast, BAWE consists of assessed writings that receive merit or distinction grades, which serve as a better reference in terms of level of literacy and are more equivalent to the texts in the learner corpus. The dissertations in the learner texts, as mentioned before, are written for completing the undergraduate degree. They are considered by Zou (2018) as fairly good quality since poor delivery may result in failing to receive the degree.

Some studies use expert writing as the reference corpus, such as Yang (2018) who compiled journal articles. While expert writing has its advantage in that the writers are mostly academics and are more familiar with conventions and styles in academic writing, Hyland and Milton (1997) comment it as an 'unrealistic standard' (p. 184). Dudley-Evans (1999) is one of the first researchers who calls for more investigation into student writing rather than expert writing to inform EAP teaching. BAWE provides a more achievable goal for Chinese students compared to journal articles, which can be placed in the middle of the novice and expert writing continuum and used for pedagogical purposes.

More specifically, levels of literacy are also associated with educational levels, and I will discuss how texts in BAWE are selected for inclusion in the reference corpus in relation to this aspect. In BAWE, there are four levels of study from undergraduate (Years 1-3) to taught masters (Year 4). As for CAEL-CAWE, the two classified levels are undergraduates and postgraduates taught. The postgraduate level in both corpora is excluded from analysis because these students may have pursued different undergraduate majors, leading to variations in their disciplinary knowledge, but this detail is not recorded in the data. In addition, a master's degree normally requires one year of study in UK universities, but it is three years in China. Since the Chinese students' dissertations are written in their last year of study, this means the age and

cognitive development of British and Chinese postgraduates differ to some extent. Therefore, writings in Year 4 (taught masters) in BAWE and postgraduate taught level in CAEL-CAWE were excluded.

Moreover, Year 1 in BAWE was also excluded from the comparison because the students at this level have not yet acquired sufficient disciplinary knowledge, nor have they fully understood how to write essays. Nesi and Gardner (2006) highlight that lecturers for first-year students might adjust the essay titles or requirements to better reflect A-level writing, thereby easing the transition for students. In addition, the learner texts are submitted in their last year (undergraduate Year 4 in China), which aligns more closely with the writing levels of British students in their Years 2-3. In sum, texts written by British undergraduates from Years 2 to 3 are included in the reference corpus.

The last factor to discuss that influences the comparability between corpora is the text type. Gilquin (2022) concludes that text type can be discussed from two perspectives. At a macro level, we can examine the mode of communication, discipline, and genre. At the micro level, factors such as the length of each file, topic, and condition of the writing process (e.g., access to the reference material and time limit) are taken into consideration. The most used reference corpus mentioned in Section 2.4.2, LOCNESS and BNC, can be ruled out in the present study in that they are not discipline-specific.

Although MICUSP contains student writings in specific disciplines, the texts corresponding to the disciplines of BM and EL in the learner corpus are notably scarce, comprising three texts in Economics and 48 in English. The same limitation can be observed in BAWE, and compromises are made to select the disciplinary groups that contain comparable disciplines. However, this is hard to achieve in MICUSP because only 16 disciplines are included and they are not pre-categorised into disciplinary groups.

In the case of BAWE, it includes 35 disciplines, which are divided into four disciplinary groups: Arts and Humanities, Life Sciences, Physical Sciences, and Social Sciences. To find the closest approximate disciplines to BM and EL, I reviewed the names of the 35 disciplines to refine the search, as similar disciplines may be named differently across various institutions. Texts in potentially comparable disciplines were closely read and contrasted with those in the learner corpus. At last, two disciplines in BAWE, Business and English, were identified as the most comparable ones in terms of topic, approaches, and knowledge production.

However, after extracting texts that meet the criteria previously mentioned, as well as additional ones related to genre and macrostructure (to be discussed shortly), only five texts in Business and 47 in English were found to fit these criteria. Given the limited

number of texts, I decided to include texts in broader disciplinary groups rather than specific disciplines in the reference corpus (RC). The selected disciplinary groups are Social Sciences (SS) and Arts and Humanities (AH), which include Business and English respectively. SS comprises Anthropology, Business, Economics, Law, Politics, Publishing, Sociology, and others. As for the group of AH, disciplines such as Archaeology, Classics, Comparative American Studies, English, History, Linguistics, Philosophy, and others are included. Based on what has been discussed in Section 2.5 about the four dimensions of disciplines, SS can be classified as soft-applied and AH as soft-pure, and it is expected that disciplines in the same disciplinary group share similar features. This strategy of extracting texts in disciplinary groups not only mitigates the challenge of limited texts in specific disciplines but also ensures a comprehensive and relevant dataset for analysis. We still need to bear in mind how this compromise may impact the findings as each discipline in the disciplinary groups shows a different level of similarity compared to BM and EL in the learner corpus. For example, the discipline of Business is more similar to BM than Sociology. Keeping in mind this limitation, texts from SS and AH are regarded as comparable and are to be included in the reference corpus.

After discussing the selection of disciplines in the reference corpus, we will now proceed to look at another aspect of text type, genre. According to Hyon (1996), genre

is explored from three schools: English for Specific Purposes (focuses on its formal properties and communicative purposes in social contexts), North American New Rhetoric (emphasises its functions and institutional contexts), and the Australian school (applies Halliday's systemic-functional linguistics). The present study is concerned with English for academic/specific purposes, and thus this perspective is preferred. In addition, Lee (2001) and Nesi and Gardner (2012) both discuss genre in corpus studies and identify it as a category of socially and culturally conventionalised writing. Nesi and Gardner further specify that factors such as 'purpose, audience, writer role and context' (p. 25) are crucial for genre identification. They also highlight the importance of text parts, which they refer to as 'key stages' (Nesi and Gardner, 2012, p. 26), in this process.

Based on these descriptions, dissertations in the learner corpus can be characterised as a genre written by the students to demonstrate their understanding of the disciplinary topic and to showcase the development of critical thinking and reasoning skills to their lecturers in order to obtain the undergraduate degree. Additionally, after a detailed examination of the texts in the learner corpus, these texts can be divided into three generic parts: an introduction, an analysis (i.e., presentation of arguments with supporting evidence), and a conclusion (see Section 5.2.2 for a more detailed division of text parts).

Argumentative essays on general topics in LOCNESS and journal articles differ markedly from the learner texts in terms of the purposes and readers respectively, and thus they are not considered as the reference corpus. By contrast, BAWE consists of university student academic writings, which are generally similar to the learner texts in terms of context and readers. Specifically, both the learner corpus and BAWE include assessed writings produced by undergraduates, intended for evaluation as part of formal education. The primary readers of these texts are lecturers who assess the quality and depth of the writing and assign grades accordingly. These similarities make BAWE a suitable candidate as the source of the reference corpus. When taking a closer look at the varieties of genres in BAWE, they show a wide range and are well-classified. To find the comparable genre to Chinese students' dissertations, we need to understand how they are classified first.

The classification of genres in BAWE is not based on information provided by the module documentation or the staff. According to Nesi and Gardner (2012), students and staff have an unclear conception of genre and their answers to the type of writing are mostly unreliable. Thus, genres are identified through a close examination of student writing, resulting in the distinction of 13 genre families based on their social purposes and text parts (see Nesi & Gardner, 2012 for detailed procedures). Specific

genres within these families resemble each other in language features, functions, and text parts. Table 3.4 below is an illustration of these genre families reproduced from Nesi and Gardner's (2018) article.

Table 3.4 Genre families, their social purposes and examples in BAWE (Nesi & Gardner, 2018, p. 53)

Social purposes	Genre family	Examples of genres
Demonstrating knowledge and understanding	Exercise	calculations; data analysis; calculations+short answers; short answers; statistics exercise
	Explanation	legislation overview; instrument description; methodology explanation; site/ environment report; species / breed description; account of a natural phenomenon
Developing powers of independent reasoning	Critique	academic paper review; interpretation of results; legislation evaluation; policy evaluation; programme evaluation; project evaluation; review of a book/ film/ play/ website
	Essay	challenge; commentary; consequential; discussion; exposition; factorial
Building research skills	Literature Survey	annotated bibliography; anthology; literature review; review article
	Methodology Recount	data analysis report; experimental report; field report; forensic report; lab report; materials selection report
	Research Report	research article; research project; topic-based dissertation
Preparing for professional practice	Case Study	business start-up; company report; organisation analysis; patient report
	Design Specification	building design; game design; product design; website design
	Problem Question	law problem question; logistics simulation; business scenario
	Proposal	book proposal; building proposal; business plan; catering plan; marketing plan; policy proposal; research proposal
Writing for oneself and others	Empathy Writing	expert advice to industry; expert advice to lay person; information leaflet; job application; letter; newspaper article
	Narrative Recount	accident report; account of literature or website search; biography; creative writing: short story; plot synopsis; reflective recount

The last six genre families were excluded from the reference corpus because they serve different social purposes (preparing for professional practice and writing for oneself and others) compared to the learner texts. Following this, I conducted a close reading of the remaining genre families in BAWE, as well as the learner texts, to identify the most comparable genre. The learner texts show characteristics such as making coherent arguments, critically evaluating propositions, and providing evidence on topics, which align with the features of the essay genre described by Nesi and Gardner (2012). In addition, essays in BAWE are recognised to include three generic text parts: 'introduction, series of arguments, conclusion' (Nesi & Gardner, 2012, p. 38). This structure mirrors the text parts in the learner texts discussed previously.

I must admit that essays in BAWE and dissertations in the learner corpus show differences in aspects such as a more detailed division of text parts (to be discussed in Section 5.2.2) and the length of the texts. Learner texts are approximately 1.7 times longer than reference texts. With these differences in mind, the present study takes a broad definition of a genre, focusing on its social purpose, context and generic text parts. Therefore, essay is selected as the comparison genre in BAWE.

While most assignments in BAWE are considered as one single genre, there are some

exceptions and Gardner and Holmes (2010) propose three types of macrostructures to distinguish them, namely simple, compound, and complex macrostructure. Compound assignments may be a collection of several reports followed by a reflection, whereas complex assignments contain sets of texts that can be seen as a coherent piece, such as assignments structured by questions. These two types of assignments show large differences compared to the learner texts, and thus only simple assignments are included in the reference corpus.

The previous paragraphs justify the extraction of comparable texts from BAWE to be included in the reference corpus, discussing factors including variety of English, educational background, level of literacy, educational level, and text types (discipline, genre, and macrostructure). A flowchart of the extraction process is shown in Figure 3.1 below. The sequence of each step is flexible, provided that all the previously mentioned factors are considered.

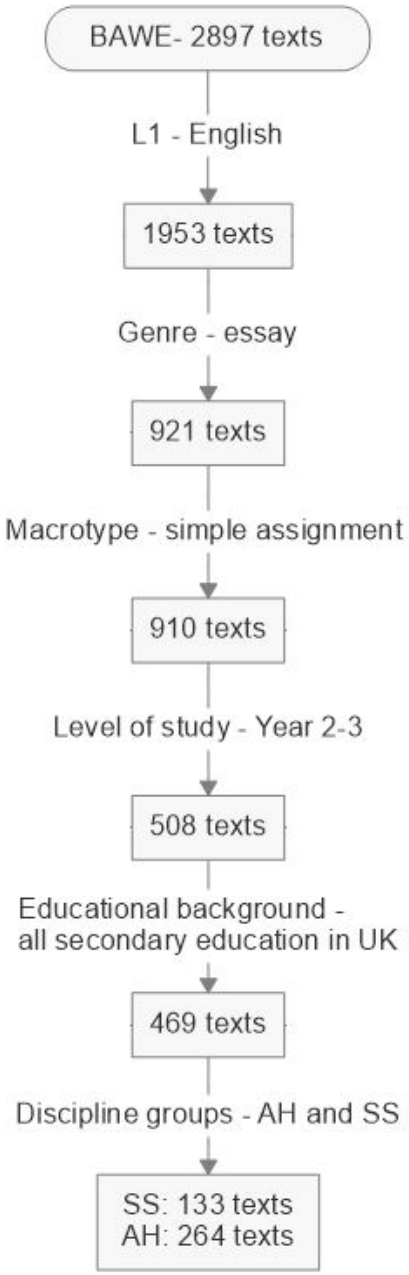


Figure 3.1 Extracting comparable texts from BAWE for the reference corpus

As shown in the figure, the reference corpus (RC) consists of two sub-corpora, containing 133 texts in Social Science (RC-SS) and 264 texts in Arts and Humanities (RC-AH). Table 3.5 below shows the number of texts and tokens in each sub-corpus.

There are about two times more texts in RC-AH than in RC-SS, and their average token number per text is similar.

Table 3.5 Number of texts and tokens per sub-corpus of the reference corpus

	Number of texts	Average tokens per text	Total tokens
RC-SS	133	3,578	475,938
RC-AH	264	3,414	901,331
Total	456	3,020	1,377,269

To conclude, Table 3.6 illustrates the comparability of some of the factors in the learner and the reference corpus. The two corpora show a significant degree of similarity, along with some differences as previously discussed. Given the absence of an entirely comparable reference, it is imperative to navigate a balance between the ideal and what is practically available. The reference corpus is not used as the standard, but rather as the reference for comparison, making the use of BAWE a feasible option.

Table 3.6 Comparison of some factors of the learner and the reference corpus

Factors	LC	RC
First language	Chinese	English
Level of study	Year 4 Undergraduates	Year 2 and 3 Undergraduates
Genre	Dissertations	Essays
Size	1,916,017	1,377,269
Average tokens per text	6,282	3,020
Disciplines	BM and EL	SS and AH

Table 3.7 reproduces the meta-information of the learner and the reference corpus, presenting it in one table.

Table 3.7 Text numbers and tokens of the learner and the reference corpus

	Number of texts	Average tokens per text	Total tokens
LC-BM	168	6,475	1,087,833
LC-EL	137	6,045	828,184
RC-SS	133	3,578	475,938
RC-AH	264	3,414	901,331

As shown in the table, the number of texts in LC-BM and RC-SS are similar, whereas in RC-AH, the number is two times compared to that in LC-EL. In terms of tokens per text, the two sub-corpora in the same corpus show similarity, but texts in the learner corpus are almost two times longer than those in the reference corpus. Given the difference in text size across the four sub-corpora, normalised frequency rather than absolute frequency will be used for comparison, and the normalised base is set at one million unless explicitly stated otherwise.

3.3 Procedure

The previous sub-section justifies the selection of the learner and the reference corpus and describes the data. The discussion will now proceed to the procedures employed in preparing the corpora for analysis.

The initial step involves selecting a platform for storing and processing the data. BAWE is accessible through Sketch Engine, and its original files are available in the Oxford Text Archive. As for CAEL-CAWE, I uploaded the files in Sketch Engine rather than other software such as AntConc (Anthony, 2020) because firstly, managing the learner and the reference corpus in the same platform makes it easier for comparisons and ensures that relevant parameters are the same, such as token definition. Another reason is the functional requirement of the analysis. As will be discussed in Section 4.2.1, I need to annotate the target modals with meanings, which is achievable through the annotation mode in Sketch Engine. Alternative software, such as AntConc is not able to perform this function. Lastly, the support team of Sketch Engine responds to enquiries instantly, thereby saving considerable time otherwise spent on using functions that require customisation for the study. Thus, Sketch Engine is chosen as the main platform to store the data and conduct analyses.

The next step is to upload the data and compile the sub-corpora. For the learner corpus, the initial encoding of the texts is ANSI, which was then transformed to UTF-8 in the software Notepad ++ (Ho, 2018). As the texts are stored in separate folders based on educational level and discipline, it is relatively easy to compile the sub-corpora, LC-BM and LC-EL, and upload them in Sketch Engine. As for the reference corpus, the

compilation of RC-SS and RC-AH presents challenges due to the absence of certain parameters in Sketch Engine's filtering options for BAWE, such as educational background and macrostructure, and these factors were filtered manually.

At this stage, the sub-corpora in the learner and the reference corpus are ready for investigation in Sketch Engine. The following step is to conduct a pilot study to decide the target modals. Section 1.2 has demonstrated the rationale for adopting a form-based approach and selecting a limited number of modals for discussion. Drawing from the literature mentioned in Section 2.2, a list of epistemic necessity and obligation modals is compiled, including *must, need, need to, have to, have got to, should*, and *ought to*. This list is not exhaustive but covers most of the frequently investigated and used ones.

When loading the files in the learner corpus, texts were automatically annotated with part of speech, and this facilitates the Corpus Query Language (CQL) searching for the modals. These search terms could exclude some non-modal uses and include all the forms of the semi-modals, such as *has to* and *had to*. The search terms for each modal are listed below, in which MD is the tag for modal (verbs), and their frequency information is presented in Table 3.8.

- Must: [lemma = "must" & tag = "MD"]
- Need: [lemma = "need" & tag = "MD"]
- Need to: [lemma = "need"] [lemma = "to"]
- Have to: [lemma = "have"] [lemma = "to"]
- Have got to: [lemma = "have"] [lemma = "get"] [lemma = "to"]
- Should: [lemma = "should" & tag = "MD"]
- Ought: [lemma = "ought"] [lemma = "to"]

Table 3.8 Absolute frequency of epistemic necessity and obligation modals in the learner corpus

Modals	LC-BM	LC-EL
Must	**504**	**288**
Need	13	6
Need to	414	136
Have to	**416**	**323**
Have got to	0	4
Should	**1,704**	**704**
Ought to	23	33

As shown in the table, the three modals in bold, *must*, *have to*, and *should*, exhibit markedly higher absolute frequency compared to the other modals. This high frequency has the potential to deepen the discussion and uncover patterns for

generalisation. In addition, the three modals are commonly discussed in combination due to their similarities and differences, and it has been argued that Chinese EFL students tend to have trouble using them (see Section 2.4.2). While *need to* exhibits a higher absolute frequency compared to *need*, *have got to*, and *ought to*, it is excluded from the analysis in that its frequency is lower than that of *must*, *have to*, and *should*. In addition, *need to*, as shown in Section 2.4.2 Table 2.5, does not demonstrate significant differences in use between Chinese EFL students and native English speakers. This suggests that Chinese EFL students likely encounter fewer challenges in using *need to* compared to must, *have to*, and *should*. Therefore, must, *have to*, and *should* are selected as the representatives of the epistemic necessity and obligation modals for investigation.

To extract the three modals, as mentioned before, the CQL query tool in Sketch Engine was used. All the instances were further manually checked when annotating the modal meanings (see Section 4.2.1 for details). The process for searching the three modals in the reference corpus is similar to that in the learner corpus, except that the part of speech for modals is VM and VMK instead of MD. Thus, the search term, for example, is [lemma = "must"& tag = "VM"] in the reference corpus.

3.4 Analysis

The present study takes a mixed-methods approach, combining quantitative and qualitative analysis to address the research questions. Detailed accounts of each approach will be presented in Sections 4.2 and 5.2. This section offers an overview of the analytical approaches used in the study.

In line with most studies on modality in EFL writing, this study compares a learner corpus with a reference one to examine the modal use quantitatively. The quantitative analysis examines *must*, *have to*, and *should* in three aspects: frequency, meaning distribution, and co-textual features. Among the co-textual features, the main verbs used with the modals and their semantic similarities are examined in depth as it helps to reveal what action or state is modulated by the modals. The investigation of these aspects helps to provide comprehensive quantitative profiles of *must*, *have to*, and *should*, and patterns are revealed when comparing between student groups and between disciplines.

While the quantitative analysis can provide a broad picture of how *must*, *have to*, and *should* are used in Chinese EFL student academic writing, it has limitations. One such limitation is that, given the volume of data analysed, it is challenging to identify distinctive features of modals used in academic writing. In addition, the disciplines

selected in the learner and the reference corpus are not perfectly comparable, as mentioned in Section 3.2.2. Those in the reference corpus are disciplinary groups, which include a broader variety of disciplines compared to the two specific disciplines in the learner corpus, and this compromise may add noise to the data analysis.

To address these limitations, a qualitative analysis is conducted on 16 texts, with four texts selected from each sub-corpus. These texts were selected through a detailed reading to ensure that they are as comparable as possible in terms of disciplines, topics, and general structures (see Section 5.2.1 for a detailed explanation). In addition, the qualitative analysis offers a more nuanced evaluation of the modal features that are distinctive in academic writing, such as the textual voice expressed by the modals and the modal distribution in different parts of a text.

In sum, the quantitative analysis focuses on the big picture of the modal use, while the qualitative analysis explores the aspects previously examined in the quantitative analysis in a finer-grained view and offers a new lens to explore characteristic features of the modals in academic writing. The two analyses complement each other, forming comprehensive profiles of *must*, *have to*, and *should* in Chinese EFL student academic writing and revealing similarities and differences between student groups and between disciplines.

3.5 Summary

This chapter starts with a detailed explanation of how the learner and the reference corpus are selected for analysis. The learner corpus is compiled from CAEL-CAWE, which originally consists of 456 dissertations written by Chinese undergraduates and postgraduates in the disciplines of BM and EL. It is then narrowed down to two sub-corpora, 168 texts in LC-BM and 137 texts in LC-EL, excluding the postgraduate data. The reference corpus is sourced from BAWE, with texts extracted according to criteria such as first languages, disciplines, and genres. The resulting reference corpus consists of 133 texts in RC-SS and 264 texts in RC-AH.

What follows is the explanation of the procedures of how the final data is stored and processed, as well as a pilot study to select the three epistemic necessity and obligation modals: *must*, *have to*, and *should*. A mixed-methods approach is used to analyse the data, combining the quantitative and qualitative analysis. The quantitative analysis focuses on the frequency and meaning distribution of *must*, *have to*, and *should*, as well as the semantic patterns of their verb collocates. In terms of the qualitative analysis, 16 texts are selected and analysed to provide a finer-grained view in terms of the modal use in academic writing. Comparisons are made across the four sub-corpora to explore the influence of first language and discipline on the profiles of

modals. The mixed use of both approaches provides a more comprehensive view of how the three modals are used in Chinese EFL students' academic writing.

The next chapter discusses one aspect of the analyses, explaining the detailed procedures for conducting the quantitative analysis and presenting the findings.

4 QUANTITATIVE ANALYSIS OF *MUST*, *HAVE TO*, AND *SHOULD* BETWEEN STUDENT GROUPS AND BETWEEN DISCIPLINES

4.1 Introduction

This chapter reports quantitative findings of three modals, *must*, *have to*, and *should,* among the two student groups and among disciplines. The analysis can be broken down into three main aspects in accordance with the first three research questions, as listed below:

RQ 1: How frequently are *must*, *have to*, and *should* used in the Chinese EFL learner corpus and the reference corpus?

RQ 2: How are the meanings of the three modals distributed?

RQ 3: What semantic patterns can be identified regarding the main verbs that collocate with the three modals?

In other words, the profiles of the three modals will be examined in terms of their frequency distribution, meaning distribution, and verb collocates. In addition, comparisons are made between Chinese and British students, and between disciplines, to answer the last two research questions:

RQ 4: Do the profiles of the three modals differ between the two student groups, Chinese and British students?

RQ 5: Do the profiles of the three modals differ between the disciplines?

To recap, there are two corpora, the learner and reference corpus, and each consists of two sub-corpora. The two sub-corpora in the learner corpus are undergraduate writings in Business and Management (LC-BM) and English Literature (LC-EL). The reference corpus consists of writings in Social Science (RC-SS) and Arts and Humanities (RC-AH).

This chapter is organised as follows. Section 4.2 describes the methods used for the quantitative analysis. It is followed by an overview of the modals used in the learner corpus as a whole. In the three sub-sections that follow I report the results on the profiles of *must*, *have to*, and *should* in the four sub-corpora separately. The last sub-section summarises this chapter.

4.2 Method

In Sections 3.2 and 3.4, the selection of the learner and the reference corpus, as well as how they will be used to answer the research questions in general have been discussed. This sub-section will present the detailed procedures to conduct one thread of the analysis, the quantitative analysis. An overview of the steps taken is as follows.

The instances of *must*, *have to*, and *should* are first extracted from the learner and the reference corpus and annotated with meanings based on the classification framework selected in Section 2.2.2. The main verbs collocating with these modals are then extracted manually and placed on the semantic plots through distributional semantics. The co-occurring syntactic features, such as negation, voice, tense, and aspect, are also annotated for each concordance line. The sections below will explain each procedure in detail.

4.2.1 Annotation of meanings

The annotation is to facilitate the analysis of the meaning distribution of *must*, *have to*, and *should*. This procedure involves a detailed reading of each concordance line containing the three modals. Such scrutiny not only helps to differentiate the meanings of the modals, but also allows for excluding instances such as non-modal uses mislabelled by Sketch Engine or those appearing in direct quotations.

As mentioned in Section 3.3, both the learner corpus and the reference corpus were accessed through Sketch Engine, and the instances of the three modals were extracted using CQL searching. Once the search was completed, each instance was annotated with a meaning label in the concordance annotation mode of Sketch Engine. As discussed in Section 2.4.2, relying solely on syntax to annotate modal meanings is insufficiently reliable. Thus, the present study manually annotated the meanings and

invited a second rater to ensure accuracy and reliability, details of which will be presented shortly.

Descriptions of different labels and the corresponding examples are shown in Table 4.1 below. *Must* is used as a representative. The information in the bracket of each example specifies the meaning/labels of the modal, the sub-corpus to which it belongs, and the name of the text file.

Table 4.1 Descriptions and examples of each label for the modals, illustrated by *must*

Labels	Description of the labels	Examples
Epistemic	Assessment of the truth of the proposition	[…] but from the reaction of people in other boxes we have a feeling that this woman *must* have a past that is mysterious as well as regretful. (Epistemic_LC-EL_L 00041)
Root	Lay obligations or give suggestions	We *must* be very careful of our definition of intensity of competition when discussing this topic. (Root_RC-SS_0058g)
Unclear	Difficult to distinguish between the epistemic and root meanings	This awareness *must* include self-requiring, self-understanding and self-developing. (Unclear_LC-EL_L 10408)
Quote	Used in the direct quotations	Paragraph 1.1 outlines a general objective: 'All persons in custody *must* be dealt with expeditiously'. (Quote_RC-SS_0119g)
Other	Non-modal use	Besides, Helen is the only friend that Jane has to support her when she receives the same ostracism as in Gateshead–'locked' to a high stool and everyone else is forbidden to speak to her. (Other_LC-EL_L 02104) Mr. Lee reckons that marketing research is a must. (Other_LC-BM_UGBM 00609) *Should* Dorian Gray's secret double life have been discovered then it would have been considered shocking and no doubt have led to social ruin. (Other_RC-AH_3005d)

The two pre-decided labels are *epistemic* and *root*, representing the two main meanings these modals express. The selection of Coates's (1983) binary classification of modal meanings has been justified in Section 2.2.2, and its applicability to the current dataset has also been demonstrated in the annotation process. Other labels, such as *unclear*, *quote*, and *other*, were not pre-determined and were added in the annotation process.

The label *unclear* was assigned to instances identified by Coates (1983) as ambiguity and merger (refer to Section 2.2.2) due to the difficulty in classifying them under the epistemic and root meanings. These instances were emphasised for clarity during the discussion with the second rater (will be discussed shortly). Instances that were labelled as *quote* are those in direct quotations. These instances were excluded from the quantitative analysis because they are not the words of the students. However, they will be discussed in the qualitative analysis to examine how the students use these quotations with the target modals in them to express their own viewpoints (see further explanation in Section 5.2.2). The *other* label was applied to three scenarios, as exemplified in the table: first, to instances incorrectly identified as modals by the CQL query of Sketch Engine, such as *have* and *to* used separately; second, to cases where *must* is used as a noun, which Jenkins (1972) identifies as marginal because they show little semantic resemblance to the modal use; and third, to instances where

should functions as a quasi-subjunctive or *should* supplies a first-person variant for hypothetical *would* (see Section 2.2.5). In the last case, only the conditional *should* with subject-auxiliary inversion was observed in the data, and these instances were not discussed in the analysis because they express a low degree of modality and the removal of *should* does not markedly alter the meaning of the sentence.

Following the annotation scheme described above, the annotated results in Sketch Engine were saved and downloaded as spreadsheets. To ensure the reliability of the annotation, a second rater was invited to annotate the data. The process was documented below, following Larsson et al.'s (2020) advice on providing information about the measurement of the inter-rater reliability in the learner corpus research. I randomly selected a sample of 497 instances, which represents approximately 10% of the total occurrences of *must*, *have to*, and *should* (5,551 instances) in the two corpora. The number of instances in the sample seems to be a manageable amount for the second rater and, at the same time, could offer insights into the accuracy of the annotations. In addition, the selected instances of each modal are proportional to their frequency in the learner and the reference corpus respectively. This approach ensures that the sample accurately represents the distribution of these modals within the two corpora. A second rater, who is a native speaker of English and a lecturer in Linguistics and Teaching English as a Second Language, worked independently and coded the sample. The results were then compared with the annotations made by me.

Although raw agreement, which measures the percentage of the agreement cases in all cases, is useful in some way, it overlooks the agreement by chance, namely, the agreement that would be achieved by the two raters if they code randomly. The agreement coefficient, on the other hand, could take this into consideration. Gwet (2014) describes two agreement coefficients, Cohen's Kappa (Cohen, 1960) and Gwet's AC_1 (Gwet, 2008), both of which are applicable to my data since they are suitable for analysing agreement on two or more nominal variables between two raters. Gwet's AC_1 is used in the present study because it addresses a limitation of Cohen's Kappa that a low value is sometimes unexpectedly yielded when the raters achieve high agreement (Feinstein & Cicchettis, 1990). The cut-off points of Gwet's AC_1 for agreement and very good agreement are 0.67 and 0.8 respectively according to Krippendorff (2012 [1980]).

The measure shows very good agreement between the raters ($AC_1 = 0.86$, $p < 0.001$). A review of the differences in the annotations between raters found no systematic pattern of disagreement. The two raters also discussed some of the differences and instances that were annotated as *unclear* and *other*, ultimately reaching a consensus. Given the nature of the judgement variable, the amount of agreement was deemed sufficient.

4.2.2 Distributional semantic analysis for verb collocates and concordance line annotations

Apart from the annotation of modal meanings, another crucial aspect, as mentioned in Section 3.4, is to examine the semantics of the main verbs collocating with the three modals. While previous literature has explored other co-textual features of modals, such as syntactic features of voice or aspect, the main verbs and their semantics have largely been overlooked (see Section 2.2.4). Focusing on this aspect could provide a more comprehensive profile of the three modals in Chinese EFL student academic writing. To achieve this, distributional semantics was used to explore the semantic similarity of the verb collocates. In the following paragraphs, I will explain why this method was chosen over alternatives, its application to the present study, and the specific procedures to apply it.

Before introducing distributional semantics, some alternative methods will be discussed. I intend to group the verb collocates based on their semantic similarities, and Section 2.2.4 has introduced and evaluated some approaches. While a bottom-up approach could be employed to categorise them through a detailed examination of concordance lines as demonstrated by Deshors (2016) and Furmaniak (2020), this method remains subjective and is impractical for managing 5,551 concordances in this study. The same challenge arises when considering the use of verb classifications proposed by Biber et al. (1999) and Halliday and Matthiessen (2004).

The framework of Biber et al. (1999) was used as a reference to assist the analysis when needed, yet there remains a necessity for a more objective way to semantically group the verbs, and one option is to query the WordNet (Fellbaum, 1998). WordNet is a digital lexical database of English that organises words into sets of synonyms (also called synsets). Different meanings of a word are listed as separate synsets, and these lists are searchable for conducting semantic annotation. The synsets are interconnected via relations such as super-subordinate relations (e.g., *fruit* and *apple; communication* and *whisper*), part-whole relations (e.g., *backrest* and *chair*), antonymy (e.g., *young* and *old*), etc. The classification and interpretation of word senses are based on two resources: corpus and dictionary, which ensures that the classified senses are representative of actual use and accurately reflect semantic relationships. To illustrate how WordNet could be used for the present study, Figure 4.1 below is a screenshot of the entry for the word *question* in WordNet 3.1.

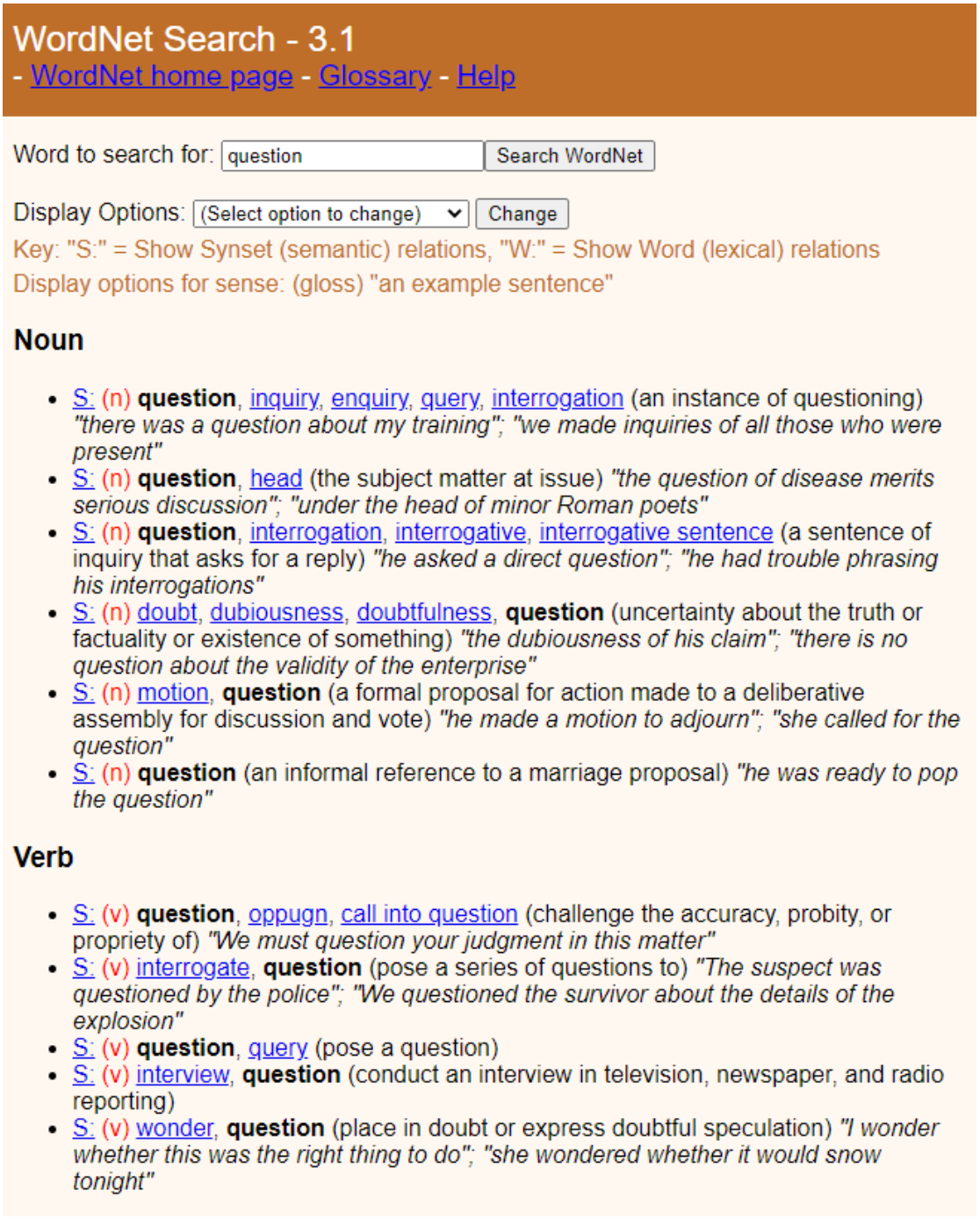


Figure 4.1 Screenshot for the entry of *question* in WordNet 3.1

As shown in the figure, *question* can function as both a noun and a verb, and there are five senses listed for the verb reading of *question* in WordNet 3.1. Each verb sense

entry includes near-synonyms with hyperlinks directed to the corresponding words, interpretations in the brackets, and an example sentence. I have to admit that each sense shows slightly different interpretations and emphases in the use of *question*. However, to what extent are these differences meaningful for the present study remains in doubt. Perek (2015) comments that the distinctions of some word senses are relatively arbitrary and sometimes excessively detailed. I initially considered annotating each verb collocate with specific senses from WordNet and then manually documenting the semantic link between verb collocates. However, this process is rather time-consuming and some of the nuanced senses are not markedly different or applicable in the context. What I am looking for is a method that can automatically and systematically identify the semantic similarity between the main verbs used with the modals, and one such approach is distributional semantics.

Distributional semantics quantifies the meanings of words based on their co-texts of occurrence in large amounts of data, in particular the words that frequently co-occur with them within a specified window in the text. It holds the view that words with shared co-text tend to have similar meanings, or as Firth (1957, p.11) states, 'you shall know a word by the company it keeps'. For example, *eat* co-occurs with food (e.g., *pizza* and *chicken*) or tableware (e.g., *fork* and *knife*) frequently. These collocates also tend to be used with *taste*, but they are unlikely to appear with *increase* or *fly*. This is because *increase* or *fly* collocate with words that are less likely to be shared with *eat* and *taste*.

To measure the semantic similarities numerically, vector space models (Turney & Pantel, 2010; also called distributional semantic models) are used as the implementation of distribution semantics. Words are represented as vectors based on their co-text in a mathematical space (this will be discussed shortly), and the semantic similarity between the words is modelled by the spatial relationship between the vectors. In other words, words that are semantically similar will be placed closely in the space. Before discussing the selected models and detailed procedures of distributional semantic analysis, I will present the applications of this method and its advantages for the present study.

Distributional semantics, while originating in Natural Language Processing, is also applied in linguistic studies. For example, Hilpert and Perek (2015) explore how this method can be used to detect and visualise the diachronic semantic change of constructions using motion charts. Perek (2016) further investigates the syntactic productivity in diachrony. Levshina and Heylen (2014), on the other hand, study the two near-synonymous constructions containing *laten* and *doen* in Dutch. There are also applications in modality studies (e.g., Hilpert, 2016 and Hilpert & Flach, 2021), as mentioned in Section 2.2.4. However, these studies focus mostly on general English or historical English, and no attempt has previously been made to use this approach in the study of modality in EFL student academic writing. The present study aims to fill

in this gap, exploring the semantic patterns of the verbs collocating with the modals in question, and making comparisons between student groups and between disciplines.

One advantage of distributional semantics is that it quantifies the semantic similarities automatically and systematically, rather than relying on human intuition. This could allow for the examination of a relatively large number of words. Unlike WordNet, which requires the researcher to search for the entry for each verb and note down their semantic relations, the visualisation of distributional semantics, as will be shown in the following sub-sections, presents the semantic relations straightforwardly. The distance between verbs represents the degree of their semantic similarities. What is worth noting is that the semantic similarity quantified by distributional semantics presents not only the hyponymy and synonymy but also antonymy. Thus, 'semantic similarity' refers to all the relations mentioned above, indicating to what extent the words can be substituted for each other (Perek, 2016). However, one limitation of distributional semantics, as pointed out by Perek, is that it does not distinguish different senses of a word since it is related to the word form. For example, verbs such as *leave* and *cover* can be interpreted in abstract and concrete meanings depending on the context. Nevertheless, this issue may only be a serious problem for the polysemous verbs with balanced frequencies across different meanings, which constitute only a minor portion of the overall dataset. What we could do is to bear this issue in mind and double-check the concordance lines when analysing the groupings of polysemous verbs.

Let us now discuss the distributional semantic model used for the present study. Several approaches to build the models have been developed for distributional semantics. For example, the 'bag-of-words' approach (Manning et al., 2008) is a foundational technique that treats text as a collection of words and quantifies the co-occurrences of individual words within the text. Specifically, the procedures to build the model include documenting the word and its co-occurrence with other words in a co-occurrence matrix. In this matrix, rows represent target words and columns represent their collocates. Each cell records the frequency of their co-occurrence, and each row forms a word vector. The semantic similarity between words is then measured by calculating the cosine of the angle between these vectors. This approach is useful for preliminary analyses due to its simplicity in focusing primarily on the frequency of words and their individual collocates.

In contrast, the word2vec model (Mikolov et al., 2013) uses a neural network to capture and quantify the relationships between a word and its surrounding words within a specified window. Compared to ‘bag-of-words’, this method uses a more sophisticated computational model, a technical description of which goes beyond the scope of the thesis. It not only identifies which words are used but also how they co-occur in meaningful patterns, capturing both syntactic and semantic relationships. Furthermore, word2vec has been shown to correlate more closely with human judgements of

semantic similarity than the ‘bag-of-words’ model, as demonstrated by Perek (2021).

Thus, among the distributional semantic models for verbs built by Perek (2021), the one trained with word2vec, rather than created by ‘bag-of-words’, is used in the present study. To be more specific, the model was trained with Continuous Skip-gram, one of the neural network architectures of word2vec, on the COCA corpus. The Skip-gram model predicts the co-text given a target word to generate word embeddings. Throughout the training process, the neural network adjusts its weights to improve the prediction of co-textual words, and these refined weights serve as semantic vectors for each word, capturing nuanced linguistic relationships. Words are considered similar if they predict similar co-texts, and these relationships are captured in the weights, forming the distributional semantic model.

As mentioned before, the model is based on the COCA corpus. Since the model mirrors the data fed to it, one might question the extent to which the COCA data accurately represents the semantics of verbs in the learner and reference corpus. Admittedly, a more valid choice would have been to use data from a corpus consisting of academic writing. However, building a model for distributional semantics requires a large amount of data in the corpus to make sure the results are relatively reliable. Most of the available academic writing corpora only include limited texts, which is not sufficient to assess the meanings of the words with a certain degree of reliability. Using an

American English balanced corpus to model the semantic similarities of the verbs in student academic writing is not as problematic as it may seem, because the meanings of words are not likely to differ markedly between the two sets of writings.

The model contains semantics for all lemmatised lexical verbs with a frequency threshold of 100 in COCA. This threshold is set to mitigate the issue of data sparsity. Primary verbs such as *be*, *have* and *do* are excluded from the model, irrespective of their roles as main or auxiliary verbs, due to challenges in accurately and straightforwardly distinguishing their main verb uses from their more frequent auxiliary uses. These verbs generally carry minimal semantic content and serve primarily grammatical functions. Thus, their exclusion helps to reduce noise and enhance the model's ability to delineate meaningful semantic relationships.

There are different word2vec models that vary in parameters such as the window size from which co-text is extracted and the dimensionality of the vectors. As Perek (2021) notes, the performance of these models with different parameters may depend on the task type. After several attempts to perform the same task (capturing meanings of the verb collocates of the target modals) using different models, the word2vec model for verbs based on COCA with a window size of two to the left and right of the target word and a dimensionality of 1,000 of the vectors performed well and was used in the present study.

Having decided on the model to use, I then used a technique called ‘t-SNE’ (Van der Maaten & Hinton, 2008) for visualisation, placing each data point, in my case, verbs, in a two-dimensional semantic plot based on their semantic proximity. The relative positions of the verbs are reflective of how verbs are co-textually and semantically related according to the data the model was trained on. Compared to other visualisation techniques such as multidimensional scaling, one advantage of using t-SNE is that it minimises the possibility of overlapping the data points in the centre of a map, as commented by Van der Maaten and Hinton, and correspondingly presents more meaningful clusters.

The steps above show how the model and the visualisation technique are decided. The following paragraphs will present how to prepare the data to load in the model and generate the semantic plots. As mentioned in Section 4.2.1, concordance lines of *must*, *have to*, and *should* with meaning annotations were downloaded in spreadsheets through Sketch Engine. Main verbs collocating with these modals were identified manually, rather than being automatically extracted using the CQL query based on a predetermined span to the right. The manual extraction allows for an accurate differentiation between the main verb *be* and its role as an auxiliary verb in the passive voice. In instances of passive voice, the main verbs used with the modals, rather than the auxiliary verb *be*, were extracted.

The main verbs were then listed in a separate column in the spreadsheets, as well as co-occurring syntactic features such as negation, tense, aspect and voice. These syntactic features were predetermined for annotation to uncover the modal use in a broader sense. Additional annotations were added in the *Note* column during the process of annotating. Some of them were added because they were observed to have associations with the modals in prior research (see Section 2.2.4), such as the frequent co-occurrence of the modals with pronouns and the existential subject. Others, such as the co-text following main verb *be*, were annotated because *be* is frequently used with epistemic sense of the modals (will be presented shortly), but it was excluded in the distributional semantic model, as mentioned before. Annotating the co-text following *be* helps to address this limitation and provides additional insights. Table 4.2 below is an illustration of how instances of *should* in LC-BM were annotated in the spreadsheet.

Table 4.2 Illustration of annotating concordance lines for the quantitative analysis (taking *should* in LC-BM as an example)

Text id	Left	Kwic	Meaning/ Label	Verb collocate	Syntactic features	Note	Right
UGBM_UTF8/ UGBM 00004.txt	take all the responsibility of this crisis. If the guests come all the way from a distant place to visit this park, they	should	Root	compensate	passive		be compensated by the company. In related laws, the cost to be recovered *should* consist of communication charge,
UGBM_UTF8/ UGBM 01809.txt	products are never on sale. As a representative 'No Discount' brand, the company believes that one beauty of luxuries	should	Epistemic	be		be + noun	be the exorbitant price. So besides the 'No Discount' strategy, the price of products keeps increasing. The long
UGBM_UTF8/ UGBM 00004.txt	culture when they conduct cross-cultural management. When Disney expands to another world to build their brand, they	should	Root	strengthen			strengthen the cooperation with local government or enterprises. They are facing the business environment,

Two columns in the spreadsheet, the meaning/label and the verb collocates, were copied into a new spreadsheet, and the absolute frequency of the co-occurrence of each verb collocate with the modal in the sub-corpus was added in this spreadsheet. The frequency was calculated using the formula COUNTIF in Excel. Before importing these verbs and their frequency information into the selected distributional semantic model, they needed to be checked in terms of the variety of English. The texts chosen in the reference corpus are written by British students, who would mostly use British spelling since their secondary education is all based in the UK. Regarding the learner corpus, the spelling may vary between British and American English. However, the distributional semantic model I use is based on COCA, which includes mostly American English. Therefore, verb collocates need to be transformed from British spelling (e.g., *realise, recognise, characterise, emphasise* and *analyse*) to their American equivalents to avoid missing data points retrieved from the model. It is important to note, however, that when presented as examples in the study, the British spelling in student writing is retained in its original form.

In addition, a pilot study was carried out to decide the need for a frequency threshold and to test different parameters for a reader-friendly visualisation (e.g., the font size of the verbs). A frequency threshold of two was set for the co-occurrence of the verbs

collocating with the root sense of the target modal but not their epistemic use. In other words, verbs that co-occur with the root use of the modals only once were excluded when constructing the semantic plots. This is because, in the pilot study, I plotted all the verb collocates of root sense of the modals with no frequency threshold, and the semantic plots were full of verbs overlapping with each other. It is hard to read and identify verb clusters with similar meanings, and most of the verbs only co-occur with the target modal once. To reduce the noise, a frequency threshold of two is set for the verbs collocating with the root modals. Epistemic use of the modals, due to their low absolute frequency in general, does not have such a problem and thus was not set a frequency threshold. The pilot study also concluded that semantic plots should be constructed separately for the four sub-corpora. This is because plotting the verbs from the entire corpus, which includes two sub-corpora, into a single plot led to considerable overlap due to the wide variety of verb collocates, making it difficult to read.

The revised spreadsheets with three columns (meaning/label, verb collocate, and absolute frequency) were then processed by the R code provided by Perek (2021), and t-SNE was used to generate the semantic plots. A detailed explanation of how to read the plots will be presented in Section 4.4.4.1 where the first plot is analysed.

4.3 Overview of the modals used in the Chinese EFL students' academic writing

Having presented the methods used for the quantitative analysis, I will discuss the frequency and meaning distribution of the three modals in the learner corpus as a whole, offering an overview of how the modals are used in Chinese EFL students' academic writing.

4.3.1 Overall frequency of the three modals in the Chinese EFL students' academic writing

Table 4.3 below shows the overall absolute frequency (AF) of *must*, *have to*, and *should* in the learner and the reference corpus, together with the normalised frequency (NF) per million words and the percentages (%) of each modal in each corpus. Although both the learner and the reference corpus contain two sub-corpora, as mentioned in Section 3.2, normalisation was performed by treating the entire corpus as a single dataset to normalise against to show the overall trend. The results will be discussed when a figurative representation of the normalised frequency is presented later in Figure 4.2.

Table 4.3 Frequency distribution of the three modals in the learner and the reference corpus

Modals	Learner corpus			Reference corpus		
	AF	NF	%	AF	NF	%
Must	708	369.52	19.08	767	556.90	41.68
Have to	689	359.60	18.57	416	302.05	22.61
Should	2,314	1,207.71	62.36	657	477.03	35.71
Total	3,711	1,936.83	100.00	1,840	1,335.98	100.00

The Chi-squared test (written as χ^2) is used to assess the significance of differences in the use of modals between student groups since it is appropriate to use with categorical data (Balakrishnan et al. 2013). Prior to applying this test, there are two assumptions that we need to check. First is the independence of observations. We assume that the use of modals in English is independent of other observations. However, this is not entirely true because linguistic features relate to each other. With this bear in mind, we need to relax this assumption with potential consequences noted, such as falsely significant results (Brezina, 2018). The other assumption is that expected frequencies need to be greater than five, a criterion our data meets as demonstrated later in Table 4.5.

Since the data generally meets these two assumptions, the next step is to use observed frequency and expected frequency to calculate the chi-squared test value. The absolute frequencies of each modal in Table 4.3 are observed frequencies, which

are reproduced in Table 4.4.

Table 4.4 Observed frequencies: modal by corpus type

Modals	Learner corpus	Reference corpus
Must	708	767
Have to	689	416
Should	2,314	657

The expected frequencies were calculated by multiplying the row total by the column total and then dividing by the grand total. The results are shown in Table 4.5.

Table 4.5 Expected frequencies: modal by corpus type

Modals	Learner corpus	Reference corpus
Must	983.72	491.28
Have to	740.52	364.48
Should	1,986.76	984.24

The chi-squared test value was calculated by squaring the difference between the observed and expected frequencies for each modal category, and then dividing these squared differences by the corresponding expected frequencies. Summing these values across all categories yielded a total chi-squared statistic of 409.89. Given the degrees of freedom (df) of 2, the p-value of less than 0.001 strongly confirms a statistically significant association between the first languages students use and their usage of the three modals: *must*, *have to*, and *should*.

To quantify the strength of this association, the effect size measure Cramer's V was calculated as 0.272, with a 95% confidence interval (CI) ranging from 0.245 to 0.298. This Cramer's V value indicates a medium effect size, which suggests a significant yet not overwhelming association (Cohen, 1988). The medium effect size underscores that the influence of first language on the use of the modals is significant and consistent enough to warrant attention but is not dominant across all observations.

Figure 4.2 below is a graphical presentation of the normalised frequency of the three modals in the two corpora. It illustrates that the two student groups have different preferences in using the three modals, and the distributional patterns of the three modals show more differences in the learner corpus than those in the reference corpus.

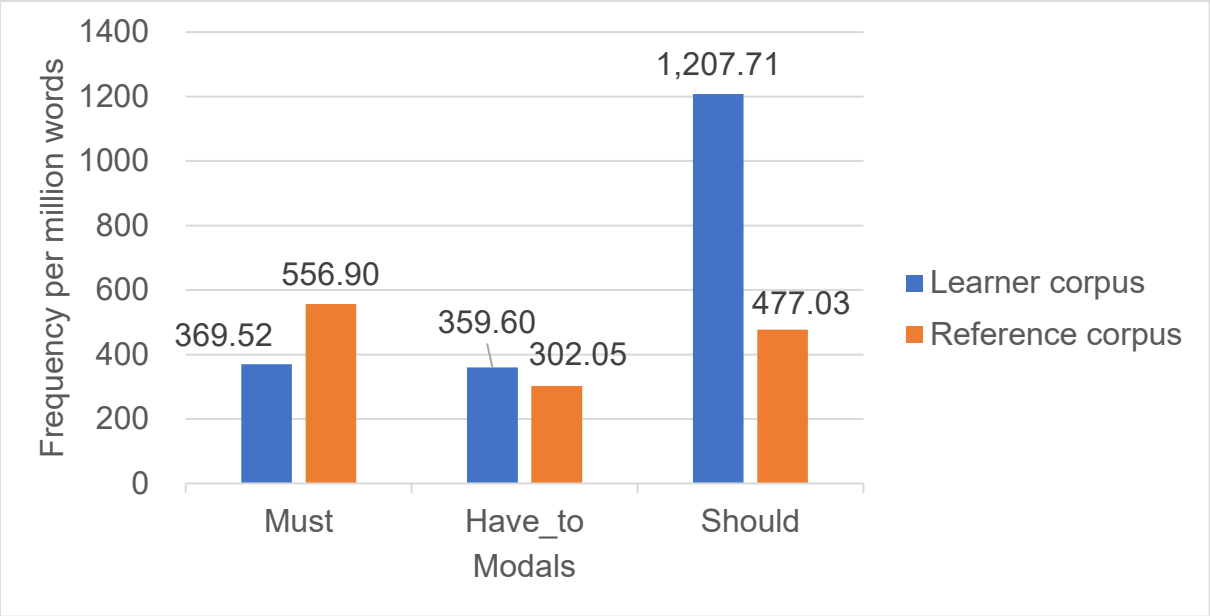


Figure 4.2 Normalised frequency per million words of the three modals in the learner and the reference corpus

The figure shows that *have to* and *should* are over-represented in the learner corpus compared to the reference corpus, with *should* showing markedly larger differences (over two times) between the two student groups than *have to*. *Must*, however, is under-represented by Chinese students, accounting for just over half of the instances used by their British counterparts.

In regard to the proportions of the three modals used by each student group, as shown in Table 4.3, Chinese students use *should* over three times more frequently than *must* and *have to*, whereas the latter two modals are distributed evenly, each accounting for roughly 20% among all three modals. The predominant use of *should* by Chinese

students might be explained by the concept of the 'lexical teddy bear' (Hasselgren, 1994, p. 237) since *should* is the first epistemic necessity and obligation modal introduced to the students in their junior high school English textbooks. By contrast, *must* and *should* show a balanced distribution in British students' writing, accounting for around 40% each among all three modals, whereas the percentage of *have to* is only 20%.

4.3.2 Overall meaning distribution of the three modals in the Chinese EFL students' academic writing

After summarising the overall frequency of the three modals between the two student groups, one might wonder if the differences result from a specific meaning of the modal. This section will present the overall meaning distribution of the three modals in the learner and the reference corpus. Let us start with Table 4.6 below, which displays the absolute frequency (AF) and normalised frequency (NF) per million words of each meaning of the modals in the two corpora. The grey background serves as a visual aid to differentiate across the meanings. Although the numbers of instances labelled *unclear* are also presented, the discussion will primarily focus on the epistemic and root use, as they constitute the majority of the instances.

Table 4.6 Meaning distribution of the three modals in the learner and the reference corpus

Modals	Meanings	Learner corpus		Reference corpus	
		AF	NF	AF	NF
Must	Epistemic	110	57.41	109	79.14
	Root	584	304.80	643	466.87
	Unclear	14	7.31	15	10.89
Have to	Epistemic	10	5.22	11	7.99
	Root	668	348.64	403	292.61
	Unclear	11	5.74	2	1.45
Should	Epistemic	93	48.54	64	46.47
	Root	2,207	1,151.87	581	421.85
	Unclear	14	7.31	12	8.71
Total	Epistemic	213	111.17	184	133.60
	Root	3,459	1,805.31	1,627	1,181.33
	Unclear	39	20.36	29	21.05

The overall meaning distribution of the three modals shows similarities between the two student groups. All three modals are predominantly used in their root sense rather than in the epistemic sense. In addition, *have to* shows the largest difference in the use of the two meanings, followed by *should* and *must*. The predominant use of root *have to* compared to its epistemic sense has also been observed by researchers such as Coates (1983) and Biber et al. (1999), as mentioned in Section 2.2.5.

As for the difference between the two student groups, while the chi-squared test reveals no statistically significant differences in the distribution of epistemic sense of

the three modals between the learner and the reference corpus ($\chi^2 = 3.31$, df = 2. $p > 0.05$; see Section 4.3.1 for calculation procedures), the normalised frequencies suggest a pattern of slight under-representation of epistemic *must* and *have to* in the learner corpus, along with an almost negligible over-representation of epistemic *should*. When examining the proportions of epistemic use of the three modals shown in Figure 4.3, both corpora exhibit a consistent pattern where epistemic *must* accounts for the highest percentage, followed by epistemic *should* and *have to*.

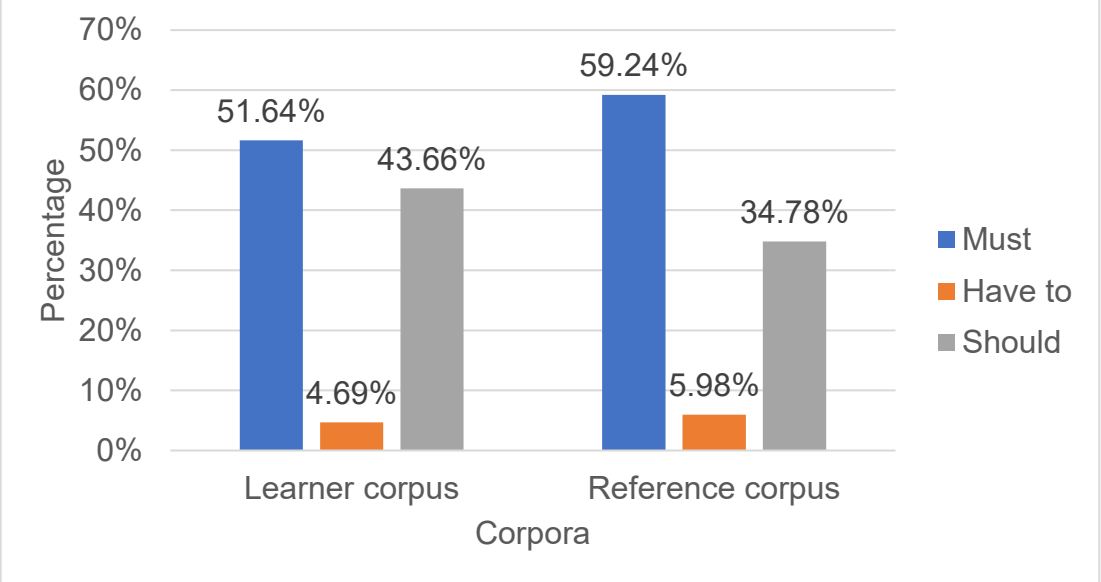


Figure 4.3 Proportion of the epistemic use of the three modals in the learner and the reference corpus

As for the root use, there is a significant association between the first languages students use and root sense of the three modals ($\chi^2 = 410.02$, df = 2, $p < 0.001$). For the effect size, Cramer's V = 0.284, 95% CI [0.256, 0.311] (see Section 4.3.1 for

calculation procedures). This suggests that the association is medium, which indicates that the influence of the first language on the root use of the three modals is marked, yet not overwhelming.

While the total normalised frequency of the root sense of the three modals is over-represented in the learner corpus, this is not the case for all three modals. It is only true for root *have to* and *should*, and the former is only slightly over-represented in the learner corpus. Root *should* shows the largest differences between the student groups, being used over 2.5 times more frequently by Chinese students compared to their British counterparts. In contrast, the normalised frequency of root *must* in the learner corpus is just over half of that in the reference corpus. In addition, Chinese students use root *should* nearly four times more than root *must*, whereas British students use these two modals evenly. Figure 4.4 below shows this pattern, demonstrating the proportion of root use of the three modals in each corpus.

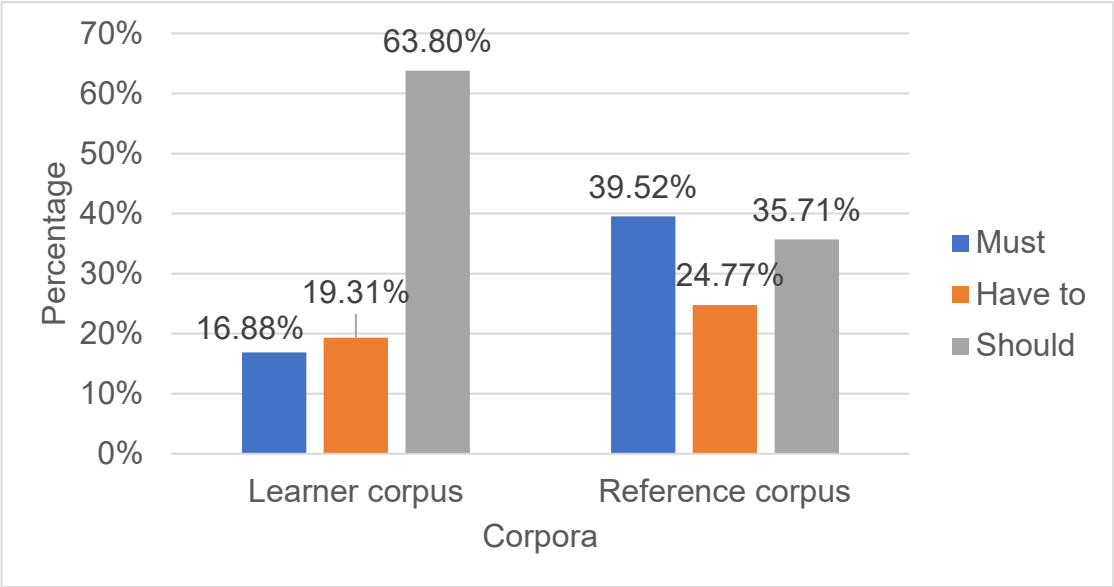


Figure 4.4 Proportion of the root use of the three modals in the learner and the reference corpus

4.4 Profiles of *must* across four sub-corpora

So far, the analysis has examined the frequency and the meaning distribution of *must*, *have to*, and *should* used by Chinese EFL students and made comparisons between the learner and the reference corpus. This sub-section, as well as Sections 4.5 and 4.6, will discuss the profiles of the three modals separately in the four sub-corpora, revealing the modal patterns between student groups and between disciplines.

The structure of these sub-sections is identical, divided into four parts: meaning distribution, dispersion, co-occurrence with syntactic features, and semantics of the verb collocates. Each instance of the modal, as mentioned in Section 4.2.1, is annotated with a meaning/label, and the discussion focuses on those identified as

epistemic and root senses as they comprise the majority of the instances. The exploration of dispersion helps to see if the modals are evenly distributed across the texts. Syntactic features such as negation, aspect, voice, and tense are investigated across the four sub-corpora to identify their association with the modal use. The last aspect, the semantics of the verb collocates of the modals, is studied to show the semantic patterns of the actions modulated by the modals. Discussions on these aspects provide a comprehensive profile of the three modals in Chinese EFL students' academic writing. In each sub-section, a comparison between the two student groups will be made first, followed by an analysis of disciplinary variations. The examination will start with the modals that convey a stronger sense of strength, *must* and *have to*, before progressing to *should,* which denotes a weaker sense.

4.4.1 Meaning distribution of *must* in the four sub-corpora

Table 4.7 shows the absolute frequency of each meaning of *must* in the four sub-corpora. Figure 4.5 below is a graphical presentation of the normalised frequency per million words of the epistemic and root use of *must*, which helps to draw a contrast across the sub-corpora. Sub-corpora representing comparable disciplines (LC-BM and RC-SS, LC-EL and RC-AH) in the learner and the reference corpus are presented next to each other in the table and the figure.

Table 4.7 Absolute frequency of different meanings of *must* in the four sub-corpora

	Epistemic	Root	Unclear	Total
LC-BM	43	432	8	483
RC-SS	21	296	0	317
LC-EL	67	152	6	225
RC-AH	88	347	15	450

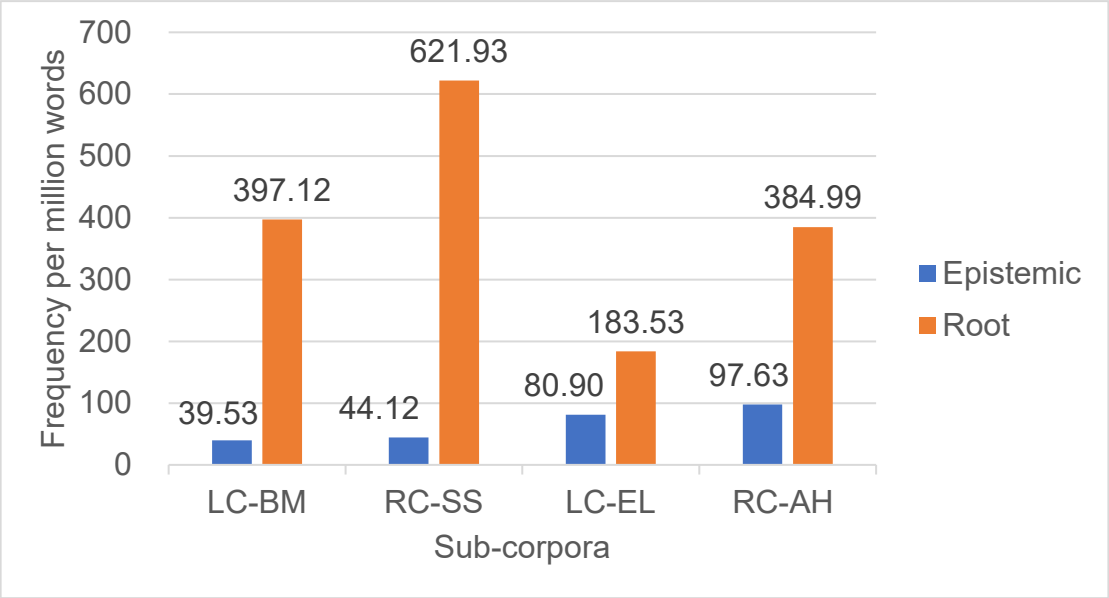


Figure 4.5 Normalised frequency per million words of epistemic and root *must* in the four sub-corpora

Section 4.3.2 has revealed that Chinese students use *must* less frequently in both of its meanings than British students. Figure 4.5 shows that this is true for both disciplines, LC-BM and LC-EL. Chinese students use epistemic *must* slightly less frequently than British students, whereas the difference in root *must* is much more salient.

As for the disciplinary variation, the pattern is similar between the two student groups.

Students use epistemic *must* markedly less frequently in LC-BM and RC-SS compared to LC-EL and RC-AH, while the use of root *must* shows an opposite trend, with considerably higher normalised frequency in LC-BM and RC-SS. This might be due to the convention of the disciplines, which will be elaborated in Section 6.3.2. Students in LC-EL and RC-AH are more accepted to make confident judgements regarding the truth of a proposition compared to those in LC-BM and RC-SS, and thus using more epistemic *must*.

This difference in using epistemic *must* is not as large as that in the root use. Root *must* is used approximately two times more frequently in LC-BM and RC-SS than in LC-EL and RC-AH. One explanation is that applied disciplines such as BM and those in SS tend to focus more on making an impact on current practice compared to pure disciplines such as EL and AH, and thus would involve more use of root modals to lay obligations or give suggestions.

4.4.2 Dispersion of *must* in the four sub-corpora

Thus far, we have looked at the frequency of different meanings of *must* in the four sub-corpora and identified features between student groups and between disciplines. However, to fully describe the profile of *must*, its dispersion is also worth investigating, which helps us to see if *must* is evenly distributed across texts.

To compare the dispersion, two measures are used, percentage range (range%) and Juilland's D. The percentage range calculates the proportion of texts in the sub-corpora where each sense of *must* appear at least once, which is useful for preliminary assessments focusing on the presence rather than the frequency of *must*. On the other hand, Juilland's D provides a more detailed analysis of distribution evenness, understanding not just whether each meaning of *must* appears, but how consistently it is used across different texts in the sub-corpora. Unlike other measures, such as standard deviation, which is influenced by the mean frequency of words and does not normalise for differences in word frequencies (Brezina, 2018), Juilland's D provides normalised values between 0 and 1. This scale facilitates straightforward interpretation, where higher values indicate a more even distribution across texts. Moreover, compared to standard deviation, Juilland's D accounts for the parts of the corpus. Although Biber et al. (2016) argue that the effectiveness of this measure decreases with extensive segmentation of the corpus (e.g., 1,000 parts), this issue is almost negligible in my study due to the relatively small number of texts involved. Thus, Juilland's D remains a robust and insightful measure for analysing distribution. Table 4.8 below presents the range% and Juilland's D of each sense of *must* in the four sub-corpora.

Table 4.8 Dispersion (range% and Juilland's D) of *must* in the four sub-corpora

	Epistemic		Root	
	Range%	Juilland's D	Range%	Juilland's D
LC-BM	18.45%	0.81	69.05%	0.90
RC-SS	11.28%	0.73	72.93%	0.90
LC-EL	32.85%	0.84	48.91%	0.87
RC-AH	21.21%	0.84	50.76%	0.89

As shown in Table 4.8, students in all four sub-corpora use epistemic *must* less evenly than root *must*. Let us compare the usage between student groups first. In the case of epistemic *must*, Chinese students in LC-BM use it more evenly compared to their British counterparts in RC-SS, whereas its distribution across texts is similar between LC-EL and RC-AH, as indicated by the same Juilland's D values of 0.84. As for root *must*, while there is a subtle difference in range%, it shows a similar distribution between the two student groups in both disciplines.

Regarding disciplinary variations, epistemic *must* is less evenly distributed in LC-BM and RC-SS compared to that in LC-EL and RC-AH. Conversely, root *must* shows a more even distribution in LC-BM and RC-SS than the other two sub-corpora. The disciplinary variations in the learner and the reference corpus show a similar dispersion pattern for both meanings of *must.*

4.4.3 Co-occurrence of *must* with syntactic features in the four sub-corpora

This section discusses the profiles of *must* in the four sub-corpora in terms of its co-occurrence with syntactic features, including negation, voice, and aspect.

Must is rarely used in negation in both meanings, and this does not show noteworthy differences between student groups and between disciplines. Among the four sub-corpora, there is only one instance of epistemic *must* that is used in negation, and that is in RC-AH. As for root *must*, ten and eight instances are used in negation in LC-BM and in RC-SS respectively. The number is two in LC-EL and eight in RC-AH.

In regard to voice, Table 4.9 shows the normalised frequency per million words of *must* used with passive voice in the four sub-corpora.

Table 4.9 Normalised frequency per million words of *must* used in the passive voice in the four sub-corpora

	Epistemic	Root
LC-BM	2.76	74.46
RC-SS	4.20	245.83
LC-EL	13.28	32.60
RC-AH	11.09	120.93

Epistemic *must* in the passive voice shows a similar pattern between student groups.

In terms of the disciplinary variation, it is used less frequently in LC-BM and RC-SS than in LC-EL and RC-AH. The normalised frequencies of epistemic *must* are relatively low since passive voice has a much stronger association with its root sense (Coates, 1983).

The co-occurrence of root *must* with passive voice shows marked differences across the four sub-corpora. Chinese students in both disciplines use markedly fewer instances of root *must* with passive voice compared to their British counterparts. It seems that Chinese students prefer to clearly state who needs to follow the obligation, whereas British students tend to prioritise describing the necessary actions without explicitly naming the doer. As to the disciplinary difference, students in LC-BM and RC-SS use more root *must* in the passive voice than those in LC-EL and RC-AH, as shown in the table. However, if we further look at the percentage of these co-occurrences given the total frequency of root *must*, it appears that the difference is not as large as it seems when comparing the normalised frequency. The percentage in the learner corpus is similar between the two disciplines, with 18.75% in LC-BM and 17.76% in LC-EL. British students show a slightly larger difference in the percentages between the two disciplines, with 39.53% in RC-SS and 31.41% in RC-AH.

The variations in using passives with root *must* across the four sub-corpora might be

related to the semantic patterns of their verb collocates. Biber et al. (1999) note that there is a strong association between the lexical factors of the verbs and the voice used. They argue that some verbs are more commonly used in passives in academic prose, such as *measure* and *define*. In addition, Coates (1983) highlights the strong association between root *must* and agentive verbs, which are more likely to be used in passives compared to stative verbs. A detailed analysis of the semantics of the verb collocates of the modals will be discussed in Section 4.4.4.

In terms of aspect, Table 4.10 below shows the normalised frequency per million words of *must* used in the perfect aspect in the four sub-corpora.

Table 4.10 Normalised frequency per million words of *must* used in the perfect aspect in the four sub-corpora

	Epistemic	Root
LC-BM	2.76	0.00
RC-SS	16.81	2.10
LC-EL	16.90	0.00
RC-AH	36.61	3.33

What stands out is the more frequent use of epistemic *must* in perfectives compared to root *must*, which corresponds to Coates's (1983) finding that there is a strong association between the perfect aspect and the epistemic meaning of *must*. Root *must* in perfectives only appears in the reference corpus with a low normalised frequency.

Chinese students use epistemic *must* in the perfect aspect less frequently than British students, with the difference being particularly noticeable in the use of epistemic *must* in LC-BM compared to RC-SS. Specifically, 6.98% of epistemic *must* in LC-BM is used in perfectives, whereas the percentage is 38.10% in RC-SS. As for LC-EL and RC-AH, the percentages are 20.90% and 37.50% respectively.

As mentioned in Section 2.2.5, *must* does not have a past form. Epistemic *must* used in perfectives can be used to refer to activities in the past (Coates, 1983). Thus, its under-representation in the learner corpus might be due to the fact that Chinese students tend to make confident judgements regarding the truth of a proposition relating to states in the present rather than in the past, with only a few exceptions as illustrated in 4-1 and 4-2. 4-3 and 4-4 are two examples taken from the reference corpus. These four examples can be divided into two parts for interpretation. The front part is to describe the certainty of the judgement expressed by epistemic *must*, and it is followed by a proposition. The judgement is made in the present, whereas the actions in the proposition happened in the past.

4-1 What is more, the company promptly and publicly demonstrated that the foreign objects *must* have been planted by the purchaser. (Epistemic_LC-BM_UGBM 02107)

4-2 According to Woolf (1931, 55) in A Room of One's Own: It is inevitable that without the private space for physical survival, women's literary creation *must* have been interrupted because she were bothered and felt anxious in a shared room where […]. (Epistemic_LC-EL_L 00507)

4-3 This is put nicely by Overton, who claims that because prices failed to rise as fast as population in the late 18th century, there *must* have been a significant rise in output to stop prices rising with demand as they had in previous centuries, indicating […]. (Epistemic_RC-SS_0202f)

4-4 Even the most obedient wife and most masterful husband *must* have come across occasions when such formal address would have been inappropriate. (Epistemic_RC-AH_0010d)

As for the disciplinary difference, it can be observed from Table 4.10 that the instances of epistemic *must* in the perfect aspect in LC-EL and RC-AH outnumber those in LC-BM and RC-SS. One explanation is that writings in LC-EL and RC-AH tend to involve more evaluation of the activities in the past, which would contribute to a higher usage of epistemic *must* in the perfect aspect.

There are also instances of the present perfect progressive tense being used with epistemic *must*, but their co-occurrence is rare, with only one instance found in each corpus, as shown in 4-5 and 4-6. These instances emphasise that the assessment of factuality is related to a state that started in the past and continues into the present.

However, the judgement itself is unaffected, as highlighted by Coates (1983). In other words, regardless of the tense and aspect co-occurring with epistemic *must*, the instances can be paraphrased as 'it can be confidently inferred that …'.

4-5 And of course, members in a 'communitas' *must* have been influencing each other. (Epistemic_LC-EL_L 01108)

4-6 However, this was not as far-reaching as the selling off of council houses in the UK since the prospective buyers *must* have been occupying the property for at least 10 years. (Epistemic_RC-SS_0075m)

4.4.4 Verb collocates of *must* in the four sub-corpora

So far, we have explored the syntactic features that co-occur with *must*. This sub-section will focus on another co-textual feature, examining the semantic patterns of the main verbs used with the modals. Comparisons will be made between student groups and between disciplines.

As mentioned in Section 4.2.2, the semantics of verb collocates are examined through distributional semantics, and the naming of some verb clusters references Biber et al.'s (1999) classification (see Section 2.2.4, Table 2.2). Main verb *be*, due to its high co-occurrence with epistemic sense of the three modals (will be shown shortly), is discussed separately because it is excluded in the distributional semantic model. In

the following sub-sections, I will first discuss the verb collocates of epistemic *must*, followed by those of root *must*.

4.4.4.1 Verb collocates of epistemic must

According to Coates (1983), epistemic *must* has a strong association with stative verbs, which is confirmed by my data. Among all the stative verbs, main verb *be* is the most frequent one, accounting for around 50% in the three sub-corpora except in RC-AH. Only 38.64% of instances of epistemic *must* are used with *be* in RC-AH. Students in RC-AH seem to use a wider variety of other verb collocates, which will be shown when comparing the semantic plots shortly.

Main verb *be* is primarily used as a copular verb to link the subject in a sentence with additional information, which means that *be* itself does not carry meaning. Thus, it is worth looking at the co-text following 'must + be' to provide additional insights. Table 4.11 below shows the normalised frequency per million words of the co-text following 'must + be' in the four sub-corpora.

Table 4.11 Normalised frequency per million words of the co-text following epistemic 'must + be' in the four sub-corpora

	Adjective	Noun	Preposition	Clause
LC-BM	0.92	16.55	0.92	0.00
RC-SS	2.10	16.81	2.10	2.10
LC-EL	9.66	24.15	6.04	0.00
RC-AH	15.53	16.64	4.44	1.11

As shown in the table, the combination of 'must + be + noun' occurs more frequently than the others across the four sub-corpora. If we closely examine the concordance lines, it can be found that over 40% of this combination is preceded by *there* in all four sub-corpora (61% in LC-BM, 100% in RC-SS, 40% in LC-EL, and 53% in RC-AH). Although the strong association between existential subject and epistemic *must* is also identified by Coates (1983), she overlooks the co-text following it. All instances of 'there + must + be' in the four sub-corpora are followed by a noun. This combination is frequently used by both groups of students, regardless of disciplines, to make a judgement regarding the factuality of described situations, as exemplified in 4-7 and 4-8.

4-7 But because the author's involvement in PR work is not long enough, there *must* be some limitations in the discussion. (Epistemic_LC-BM_UGBM 01506)

4-8 Although the intention is that knowledge flows from the mentor or coach to the individual concerned, there *must* be some learning achieved on the part of the mentor or coach too. (Epistemic_RC-SS_0320b)

The co-text following ‘must + be’ shows a similar normalised frequency pattern across the four sub-corpora, with nouns ranking first, followed by adjectives, prepositions, and clauses. No marked differences are observed in the co-text following epistemic ‘must + be’ between the two student groups, while disciplinary variations are noticeable in its use with adjectives. Writings in LC-BM and RC-SS use less ‘must + be + adjective’ than those in LC-EL and RC-AH. One reason could be the higher frequency of adjectives in LC-EL and RC-AH, which leads to a greater likelihood of using *must* to modalise them. However, I calculated the normalised frequency of adjectives in each sub-corpus, and an opposite pattern emerged. Ruling out this reason, another explanation might be related to the disciplinary conventions. The expression of certainty regarding a quality expressed by an adjective tends to be subjective, as illustrated by 4-9 and 4-10 below. This style is more accepted by students in LC-EL and RC-AH, whereas students in LC-BM and RC-SS generally use a more impersonal tone.

4-9 With the arrogant idea of manufacturing human beings according to the opinions of controllers themselves, the approaches they use to create people *must* also be extreme and wrong. (Epistemic_LC-EL_ L 05207)

4-10 However it *must* have been very difficult to maintain a purely master/subordinate relationship especially when feelings of love and affection existed in a marriage. (Epistemic_RC-AH_0010d)

Thus far, we have discussed the instances of epistemic *must* used with main verb *be* across the four sub-corpora. The following paragraphs will focus on the other verb collocates and how they are semantically distributed by constructing semantic plots. Figure 4.6 below presents the semantic distribution of these main verbs used with epistemic *must* in each sub-corpus.

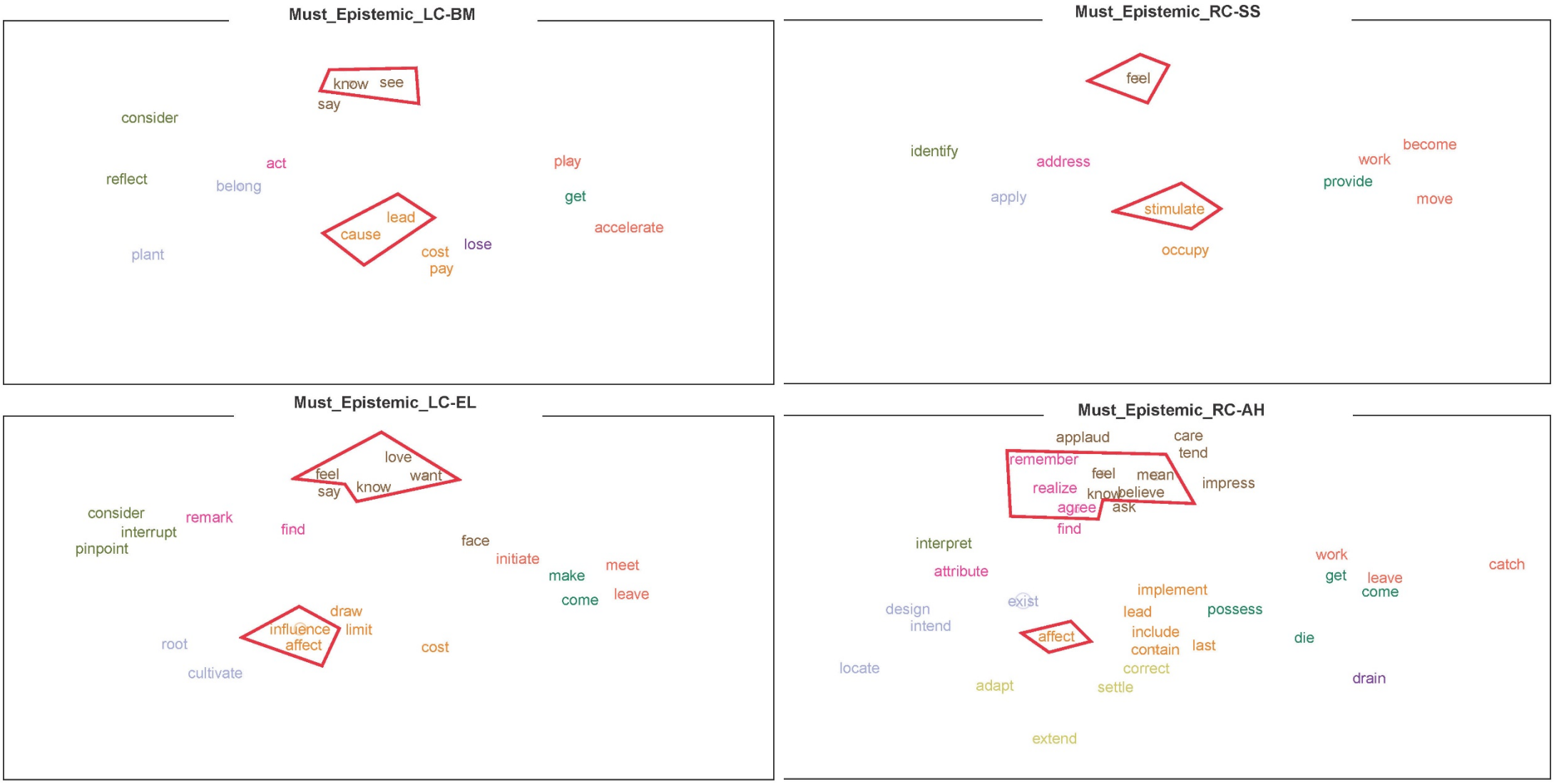


Figure 4.6 Distributional semantic plots of the verb collocates of epistemic *must* in the four sub-corpora

In order to fit on one page, a landscape orientation and a small font size are used to minimise overlapping of the verbs. Although this allows the general trend to be observable, there may still be dense clusters that are difficult to read. The high-resolution versions of all the semantic plots are provided in the Appendices for reference.

Before digging into the comparisons, I will first explain how to interpret the figure. The figure contains four semantic plots, each showing the results of one sub-corpus. The label at the top of each plot presents basic information. For example, 'Must_Epistemic_LC-BM' indicates that verbs in this plot are collocates of epistemic *must* in the sub-corpus of LC-BM. The presence of each verb is indicated by bullets, the diameters of which are proportional to the natural logarithm of the absolute frequency of the target modals used with the verbs in each sub-corpus. The frequently used verb collocates, such as *influence* with five co-occurrences, are represented by larger bullets, whereas those that co-occur only once, such as *accelerate,* have no bullet around them.

The relative positions of the verbs in the plot, as mentioned in Section 4.2.2, are based on what the distributional semantic model learns during training in regard to the co-

textual and semantic similarities of these verbs. The closer two verb collocates are, the more similar they are in meanings. For example, *lead* and *cause* are placed closely in the plot because they share similar meanings. In this study, clusters refer to groups of verbs that are semantically similar, identified by analysing both the results generated by the model and through a manual examination of meaningful groupings. I will explain each of these processes in turn.

Preliminary clusters are visually distinguished by colours that are automatically and inductively assigned by the hierarchical clustering algorithm. However, Perek (2018) notes that these colours are intended to serve as a visual aid and may not always accurately reflect meaningful semantic relationships. To address this issue, Hilpert and Perek (2015) perform a manual and intuitive classification of the target words and observe differences between these categories and the ones generated by the algorithm.

However, this approach is not entirely suitable for the present study. The primary focus here is not to examine all semantic categories of verbs that collocate with *must*, *have to*, and *should*. Instead, the emphasis is on the comparisons between the student groups and the disciplines, as outlined in the last two research questions (see Section 1.3). Consequently, I conducted a manual analysis to identify prominent clusters rather

than categorising all the verbs into semantic groupings. This identification is based on both the distinctiveness of the groupings and their relevance to the research questions. These prominent clusters are circled with red polygons, which were added manually following two procedures.

First, I created a spreadsheet that lists all the verbs used in each sub-corpus across separate columns to facilitate comparisons between the student groups and the disciplines. This was achieved by using the conditional formatting feature in Excel. For instance, when comparing the verbs collocating with epistemic *must* in LC-BM and RC-SS, I compared the relevant columns of verbs using conditional formatting. Verbs exclusively used in one sub-corpus were highlighted in red and marked on the corresponding semantic plots. Areas on these plots where exclusively used verbs were densely distributed in close proximity were then circled. Attention was also directed towards identifying verbs that appeared across different sub-corpora. These circled areas form the preliminary clusters that are worth further discussion.

As noted in Section 4.2.2, one limitation of distributional semantics is that it does not distinguish different senses of a verb. Therefore, a detailed examination of the concordance lines for verbs included in these preliminary clusters was conducted to verify their meanings and ensure that the verbs within each cluster genuinely share

similar meanings. Additionally, the classification of verbs by Biber et al. (1999) was used as a reference to categorise and name the clusters. For example, although *say* was placed next to *know* and *see* on the plot of LC-BM in Figure 4.6 and assigned the same colour, it was excluded from the cluster because, according to Biber et al.'s classification, it is used as a communicative verb in 4-11 rather than a mental verb like *know* and *see*, as in 4-12 and 4-13. Additionally, the usage of *know* and *see* as mental verbs, rather than in other senses, was confirmed by reviewing these examples.

4-11 If he were the Chinese, he *must* have said the following: 'Today I prepared the food especially for your coming. Hope you'll like it.' (Epistemic_LC-BM_UGBM 06705)

4-12 In Case 2, I must confess that the man who translated Sprite into 'xuebi' *must* know Chinese culture as a book. (Epistemic_LC-BM_UGBM 00707)

4-13 If a company wishes its customers to remember its brand name, prints its local name big and bright on the surface of its products. "Coca Cola" can appear in its brand, but "KEKOUKELE" *must* be seen more strikingly. (Epistemic_LC-UGBM 00705)

The manual intervention helps to ensure that the identified clusters for further analysis are not only statistically valid but also meaningful and relevant to the research questions. Moving forward, the discussion will focus on these prominent clusters

marked by red polygons, which highlight the salient groupings and facilitate the reporting of the results. This focus does not undermine the importance of other clusters but is intended to prioritise the discussion on the most distinct ones.

As mentioned before, epistemic *must* is mostly used with main verb *be* and thus the remaining verb collocates show a relatively limited variety. We need to bear in mind that the density of the verb collocates may result from their absolute frequency, which does not necessarily indicate the differences in variety. For instance, the difference in the density of verbs in Figure 4.6 across sub-corpora corresponds to the absolute frequencies shown in Table 4.7. Thus, emphasis should be placed on analysing semantic patterns of the verb collocates.

As shown in Figure 4.6, there is no marked difference in the use of verb collocates of epistemic *must* between the two student groups (comparing the plots horizontally) and between the disciplines (comparing the plots vertically). Most verbs that collocate with epistemic *must* are stative verbs, which is consistent with Biber et al.'s (1999) finding. Among them, two verb clusters stand out across the four sub-corpora, as highlighted by the red polygon. One cluster includes mental verbs such as *know* and *feel*, which denotes cognitive and emotional states, as exemplified in 4-12 and 4-14 respectively.

4-12 In Case 2, I must confess that the man who translated Sprite into ‘xuebi’ *must* know Chinese culture as a book. (Epistemic_LC-BM_UGBM 00707)

4-14 The shock we feel when he dies is the greater because of this and gives us a sense of what Baumer *must* have felt every time he lost a friend, all contributing to Remarque’s way of representing the unrepresentable. (Epistemic_RC-AH_3005e)

The other cluster consists of causative verbs such as *influence* and *affect*, and most of these verbs are only used once, except *influence*. Epistemic *must* is used with *influence* five times in LC-EL, and three of the instances are in the perfect aspect. Chinese students in LC-EL tend to use slightly more of this verb cluster with root *must* to make confident judgements about what is the influential factor that leads to the current situation based on evidence observed such as the characters’ behaviour, as exemplified in 4-15 below.

4-15 And if Ishmael is able to finish his pilgrimage finally, he *must* have been influenced by other members. (Epistemic_LC-EL_ L 01108)

The two clusters discussed above are difficult to be identified in RC-SS due to the low absolute frequency of epistemic *must* in this sub-corpus.

4.4.4.2 Verb collocates of root must

Figure 4.7 below shows the distributional semantic plots of the verb collocates of root *must* in the four sub-corpora. As mentioned in Section 4.2.2, the co-occurrence threshold for the verb collocates of root use of the modals is set at two when constructing plots to reduce noise and overlapping.

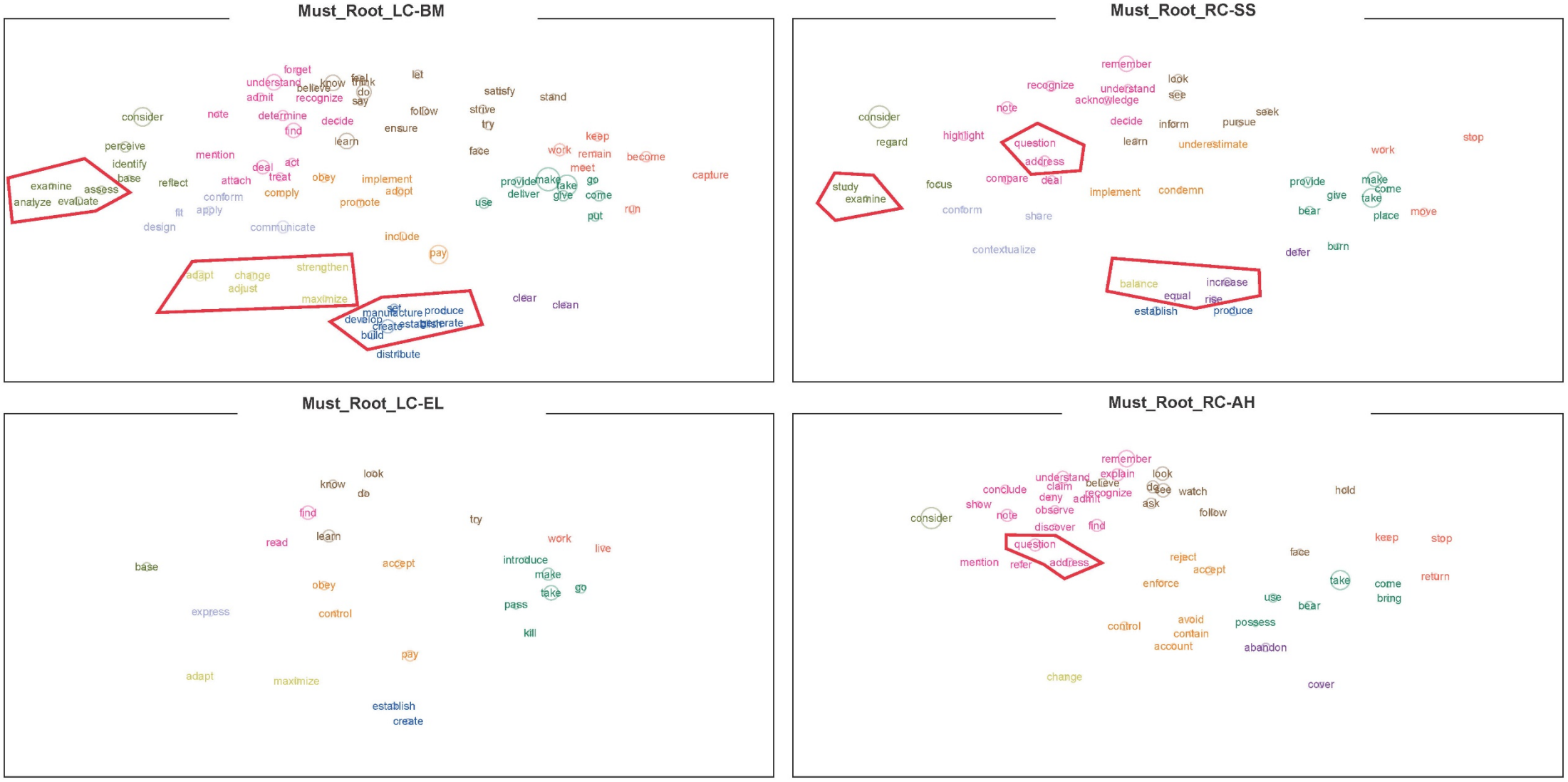


Figure 4.7 Distributional semantic plots of the verb collocates of root *must* in the four sub-corpora

An inspection of the clusters in the figure reveals that verb collocates in LC-BM show the highest degree of homogeneity compared to the other three plots. The verb collocates of root *must* in LC-BM can be classified into several clusters, whereas those in RC-SS are loosely distributed. Chinese students in LC-BM tend to give practical suggestions for business on particular aspects, including exploration and evaluation (e.g., *assess* and *evaluate*), improvement (e.g., *adjust* and *strengthen*), and development (e.g., *create* and *produce*). These uses express a relatively weak sense of obligation, as the writers do not hold authority over the subject (e.g., managers and marketers) to demand that they take the recommended action. These three clusters are circled in red in the plot from left to right, and the examples are illustrated in 4-16, 4-17, and 4-18 respectively.

4-16 At this time, managers *must* assess how dangerous this crisis is, who would be influenced most, and let every staff know the crisis and respond to it. (Root_LC-BM_UGBM 03509)

4-17 Thus L'Oréal *must* adapt its products. (Root_LC-BM_UGBM 05706)

4-18 If the marketers want to be the winners in the cross-cultural marketing they *must* create the marketing mix that meets the consumer's values on a right to their culture. (Root_LC-BM_UGBM 11304)

Only the first cluster mentioned above can also be found in RC-SS but not the latter two. The variations between the two student groups may result from the different cultural and social values, which will be discussed in Section 6.4.2. Another reason could be related to the topics covered in LC-BM. In most cases, as will be shown in Section 5.4.2, Chinese students in LC-BM identify problems in a company or a brand and conduct data analysis or review literature to give suggestions. These identified verb clusters seem to be semantically related to the suggestions provided on this topic. Although Chinese students in LC-BM do not have the power to ask the company to follow these suggestions, the use of root *must* emphasise the urgency or the importance of fixing problems. These examples can be paraphrased as 'it is important to …'.

Unlike Chinese students in LC-BM, British students in RC-SS cover a broader range of topics, which in turn is reflected in a more diverse use of verbs with root *must* to describe suggestions. In addition, there is a distinctive cluster in RC-SS related to statistics that is absent in LC-BM, such as *increase*, *rise* and *equal*, as illustrated in 4-19 and 4-20.

4-19 To restore money market equilibrium the interest rate *must* rise to induce savings. (Root_RC-SS_0058f)

4-20 In order to reduce the level of unemployment, below the ‘natural rate’ (U*, from U* U) real wages *must* be increased, thus providing an incentive for unemployed individuals to enter into work. (Root_RC-SS_0399b)

As for the comparison between writings in LC-EL and RC-AH, Chinese students use fewer mental verbs with root *must* than their British counterparts. This potentially indicates that Chinese students in LC-EL tend not to give suggestions at the cognitive level.

In addition, British students in both disciplines tend to use *question* and *address* with root *must* to underscore the necessity of challenging and critically treating propositions, as exemplified from 4-21 to 4-24, whereas this cluster is not observed in the learner corpus. In addition, three out of five instances of ‘must question’ is used with the pronoun *we* in RC-AH but not in the other three sub-corpora, engaging the readers to agree with the writer’s viewpoints. This is also one of the disciplinary variations, which will be discussed shortly.

4-21 In light of these arguments it *must* be questioned how a meaningful and internally cohesive system of regulation can be devised to reach the aim of balancing the interests of science and morality. (Root_RC-SS_0352a)

4-22 At the more general level of wider health, more broad socio-economic circumstances remain key and thus it is structural problems in society that *must* be addressed. (Root_RC-SS_0252i)

4-23 More importantly we *must* also question, whether it even mattered if the masses were able to comprehend his ideas. (Root_RC-AH_0252i)

4-24 The fact that only the Guardians would appear to be described as 'just' or self motivated to be 'just' will however be addressed. The meaning of 'justice' *must* also be addressed. (Root_RC-AH_3019h)

In terms of disciplinary variation, one cluster including verbs denoting exploration (e.g., *study* and *examine*) is used in LC-BM and RC-SS but not in LC-EL and RC-AH. It seems that students in LC-BM and RC-SS tend to give suggestions on the research itself, as illustrated in 4-25 and 4-26.

4-25 To understand any specific issues of luxury market in China, one *must* examine the total size of the worldwide business and then analyze the size of the Chinese market. (Root_LC-BM_UGBM 09508)

4-26 To answer this question, one *must* examine and compare the reasons given in the articles for the underachievement of black boys. (Root_RC-SS_0408b)

In addition, it is found that root *must* is frequently used with the pronoun *we* in

LC-EL and RC-AH, accounting for 12.50% and 27.95% of all the instances respectively. By contrast, the percentages in LC-BM and RC-SS are 6.71% and 14.19% respectively. This implies that students in LC-EL and RC-AH tend to use a more interactive tone, involving the readers to agree with their suggestions, as illustrated in 4-23.

> 4-23 More importantly we *must* also question, whether it even mattered if the masses were able to comprehend his ideas. (Root_RC-AH_0252i)

4.5 Profiles of *have to* across four sub-corpora

Having discussed the profile of *must* in the four sub-corpora, I will move on to explore the second modal in question, *have to*.

4.5.1 Meaning distribution of *have to* in the four sub-corpora

The absolute frequency of different meanings of *have to* is shown in Table 4.12, and Figure 4.8 below shows their normalised frequency per million words in the four sub-corpora for comparison.

Table 4.12 Absolute frequency of different meanings of *have to* in the four sub-corpora

	Epistemic	Root	Unclear	Total
LC-BM	7	383	7	397
RC-SS	2	103	0	105
LC-EL	3	285	4	292
RC-AH	9	300	2	311

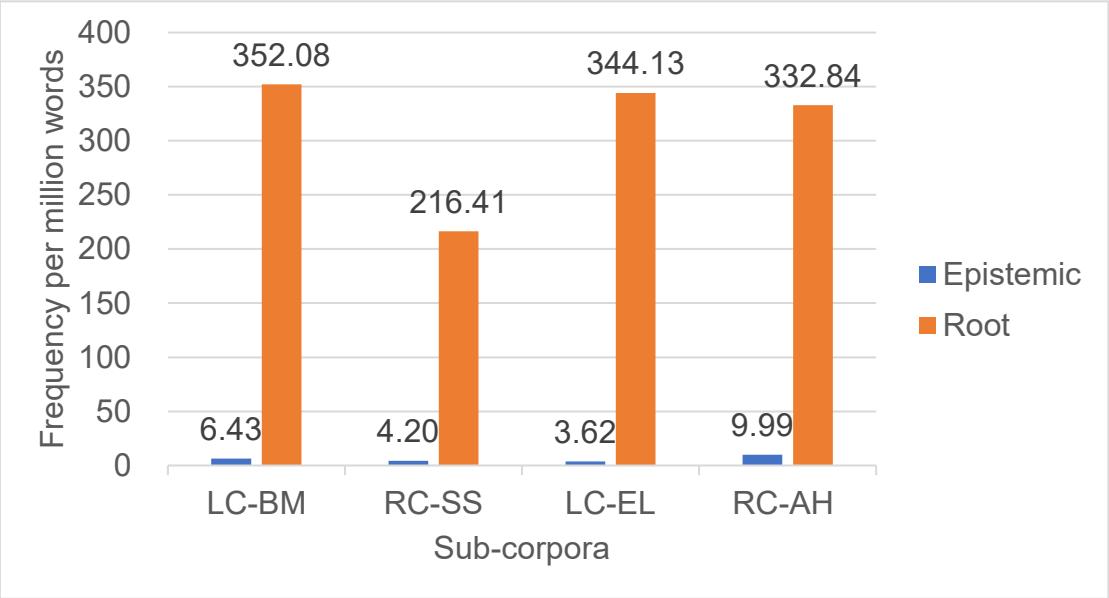


Figure 4.8 Normalised frequency per million words of epistemic and root *have to* in the four sub-corpora

As mentioned in Section 4.3.2, *have to* is predominantly used in the root sense, and it is true for all four sub-corpora. It can hardly make comparisons of epistemic *have to* in the four sub-corpora because their absolute frequency is all lower than ten. As for root *have to*, the normalised frequencies are consistent across sub-corpora, with the exception of RC-SS. Students in the other three sub-corpora

use root *have to* over 1.5 times as frequently as those in RC-SS, which could be attributed to the objectivity expressed by root *have to* (see Section 6.3.2 for discussion).

4.5.2 Dispersion of *have to* in the four sub-corpora

Table 4.13 below shows the range% and Juilland's D of the two senses of *have to* in the four sub-corpora.

Table 4.13 Dispersion (range% and Juilland's D) of *have to* in the four sub-corpora

	Epistemic		Root	
	Range%	Juilland's D	Range%	Juilland's D
LC-BM	3.57%	0.58	77.38%	0.92
RC-SS	1.50%	0.30	43.61%	0.86
LC-EL	2.19%	0.43	75.18%	0.92
RC-AH	3.03%	0.64	48.48%	0.87

As shown in the table, epistemic *have to* only appears in a few texts, with a range% under 4% in all four sub-corpora. As for root *have to*, the range% is much higher compared to its epistemic use. Root *have to* is distributed more evenly in the learner corpus than that in the reference corpus, and it does not show disciplinary variations in terms of dispersion.

4.5.3 Co-occurrence of *have to* with syntactic features in the four sub-corpora

This section will explore how *have to* is used with syntactic features such as negation, voice, tense, and aspect.

Have to is only used in negation in the learner corpus, and all such instances are its root use. Specifically, there are nine occurrences in LC-BM and ten in LC-EL. Unlike negation with root *must*, the negation of root *have to* affects the modal predication rather than the main predication (Coates, 1983), and it can be paraphrased as 'it is not necessary for'. One of the examples is 4-27.

> 4-27 Chinese Americans do not *have to* completely abandon one culture to pursue another one, especially in today's globalization. (Root_LC-EL_L 07707)

With respect to the passive voice, Table 4.14 below displays its usage with *have to* in the four sub-corpora.

Table 4.14 Normalised frequency per million words of *have to* used in the passive voice in the four sub-corpora

	Epistemic	Root
LC-BM	0.92	47.80
RC-SS	0.00	42.02
LC-EL	0.00	18.11
RC-AH	1.11	68.79

Have to is infrequently used in the epistemic sense, and thus its epistemic use is also rarely observed to co-occur with the passive voice. As for root *have to*, its normalised frequency difference between writings in LC-BM and RC-SS is not as large as it is between LC-EL and RC-AH. It seems that British students in RC-AH have a preference for using root *have to* in passives, accounting for 20.67% of all the instances of root *have to*. This percentage is the highest across the four sub-corpora. As for disciplinary differences, Chinese students in LC-BM tend to use more root *have to* in passives than those in LC-EL, whereas the trend is the opposite when comparing British students' writings in RC-SS and RC-AH. These variations between student groups and between disciplines might related to the verbs collocating with *have to*, which will be discussed in Section 4.5.4.

In terms of the tense, as shown in Table 4.15 below, epistemic *have to* is rarely used in the past tense compared to the root sense due to its overall low absolute frequency.

Table 4.15 Normalised frequency per million words of *have to* used in the past tense in the four sub-corpora

	Epistemic	Root
LC-BM	0.92	45.96
RC-SS	0.00	56.73
LC-EL	1.21	99.01
RC-AH	5.55	181.95

As for root *have to*, it is used less frequently in the past tense in the learner corpus than in the reference corpus, with a more pronounced frequency difference between LC-EL and RC-AH than between LC-BM and RC-SS. In terms of disciplinary variations, students in LC-EL and RC-AH use over two times more root *have to* in the past tense than those in LC-BM and RC-SS. This might be because writings in LC-EL and RC-AH mostly discuss literary works or other materials that were produced in the past, as exemplified in 4-28. Writings in LC-BM and RC-SS, on the other hand, are less likely to offer suggestions based on past activities or events, and one of the exceptions is illustrated in 4-29.

4-28 Finally she *had to* immorally marry her sister's lover Frank for his money. (Root _LC-EL_L 00704)

4-29 Wal-Mart and Carrefour *had to* solve lots of problems they had never faced when they first arrived in China. (Root_LC-BM_UGBM 03205)

In regard to aspect, there are only four instances of *have to* used with the perfect aspect across the four sub-corpora, all of which are found in RC-AH. Three instances express the epistemic meaning, while the remaining one conveys the root sense. The examples are illustrated in 4-30 and 4-31 below respectively.

> 4-30 In Kant's thesis he assumes that it follows whatever an event is (e), it *has to* have followed a cause (c), this cause also *has to* have had a cause (c-1), which requires (c-2) and so on ad infinitum; […]. (Epistemic _RC-AH_0407a)
>
> 4-31 What is widely agreed upon is that in order for specialisation to have any real cultural or behavioural meaning it *has to* have been a conscious decision by the hominids. (Root _RC-AH_6033b)

4.5.4 Verb collocates of *have to* in the four sub-corpora

Having discussed the meaning distribution, dispersion, and syntactic features of *have to*, I will explore the verb collocates used with each sense of *have to*. We will start with epistemic *have to*.

4.5.4.1 Verb collocates of epistemic have to

As shown in Table 4.12, *have to* is rarely used in the epistemic sense. Therefore, its verb collocates are limited, and it can hardly make comparisons between

student groups and between disciplines. Figure 4.9 below shows the semantic plots of its verb collocates in the four sub-corpora.

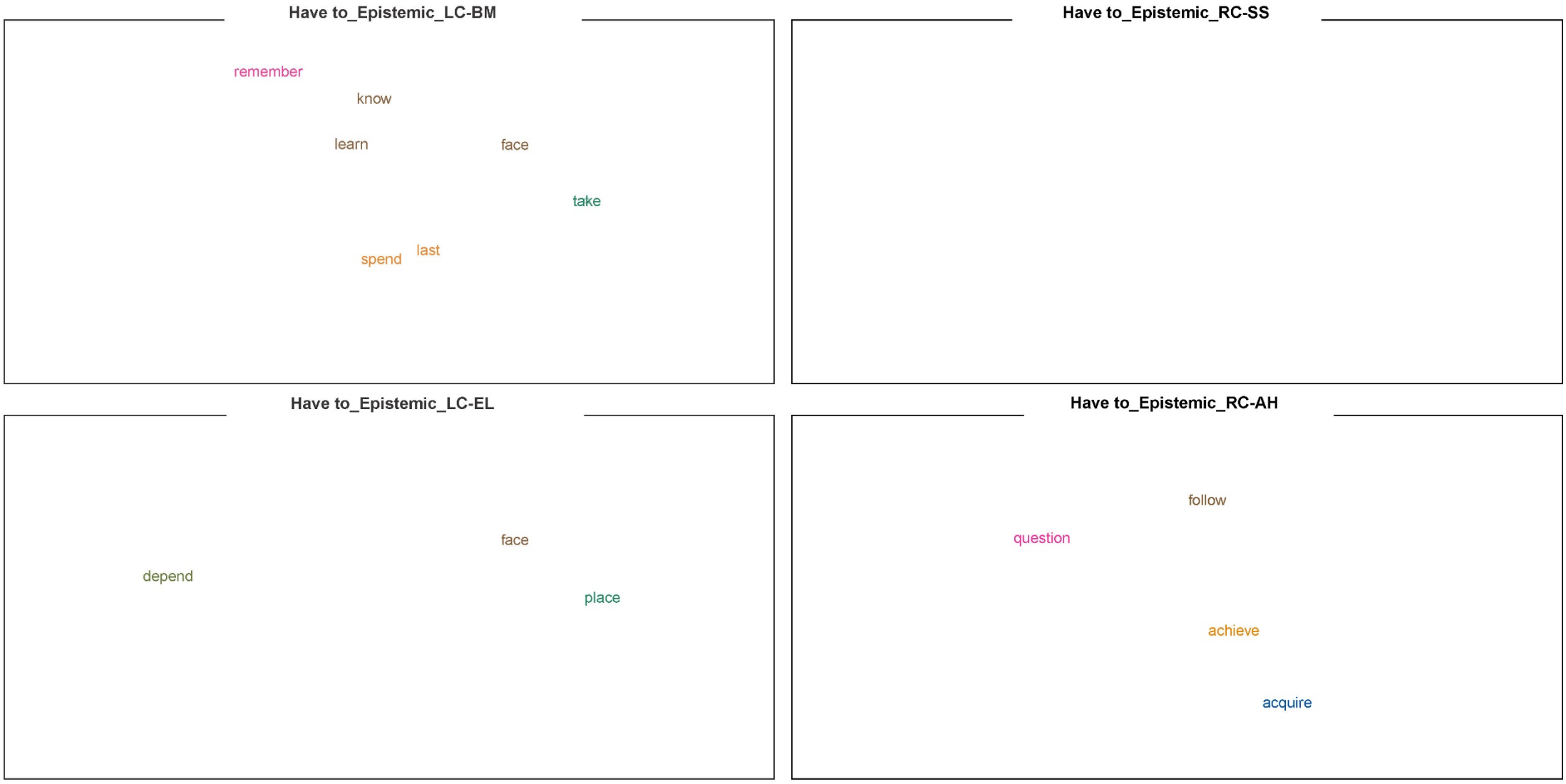


Figure 4.9 Distributional semantic plots of the verb collocates of epistemic *have to* in the four sub-corpora

Unlike epistemic *must*, epistemic *have to* in the learner corpus does not collocate with main verb *be* at all. Verbs collocating with epistemic *have to* in the learner corpus are mostly stative verbs, such as *know* and *remember* in LC-BM and *face* and *depend* in LC-EL, as shown in Figure 4.9. As for the reference corpus, there are only two instances of epistemic *have to* in RC-SS, and they are all used with main verb *be*. Thus, its corresponding semantic plot in the figure is empty because *be* is excluded from the distributional semantic model. Among the nine instances in RC-AH, three of them are used with main verb *be*, and one with *have*. These two verbs are not shown in the plot because of their exclusion from the model (refer to Section 4.2.2). The remaining five instances, four of which are shown in the plot (*achieve*, *acquire*, *follow*, and *question*), while *steamroller* does not meet the frequency threshold of 100 in the COCA data used for the model, hence it is also not displayed. 4-32 and 4-33 are two examples taken from the learner and the reference corpus respectively.

4-32 All that goes to show that a poor woman at that time *had to* depend on a husband financially. (Epistemic_LC-EL_L 01606)

4-33 I feel this could prove the pallake *had to* be of the most beautiful and desirable women as they cost a great deal. (Epistemic_RC-AH_6109b)

In 4-32, the Chinese student makes an epistemic judgement based on the woman's

living condition and historical background described in the previous sentences, assuming that a poor woman needs to depend on her husband for living. The judgement made in 4-33 is based on the evidence of cost. Both of the two examples explicitly present the evidence and support for the judgement regarding the factuality of the proposition.

4.5.4.2 Verb collocates of root have to

Figure 4.10 below shows the semantic plots of verb collocates of root *have to* in the four sub-corpora.

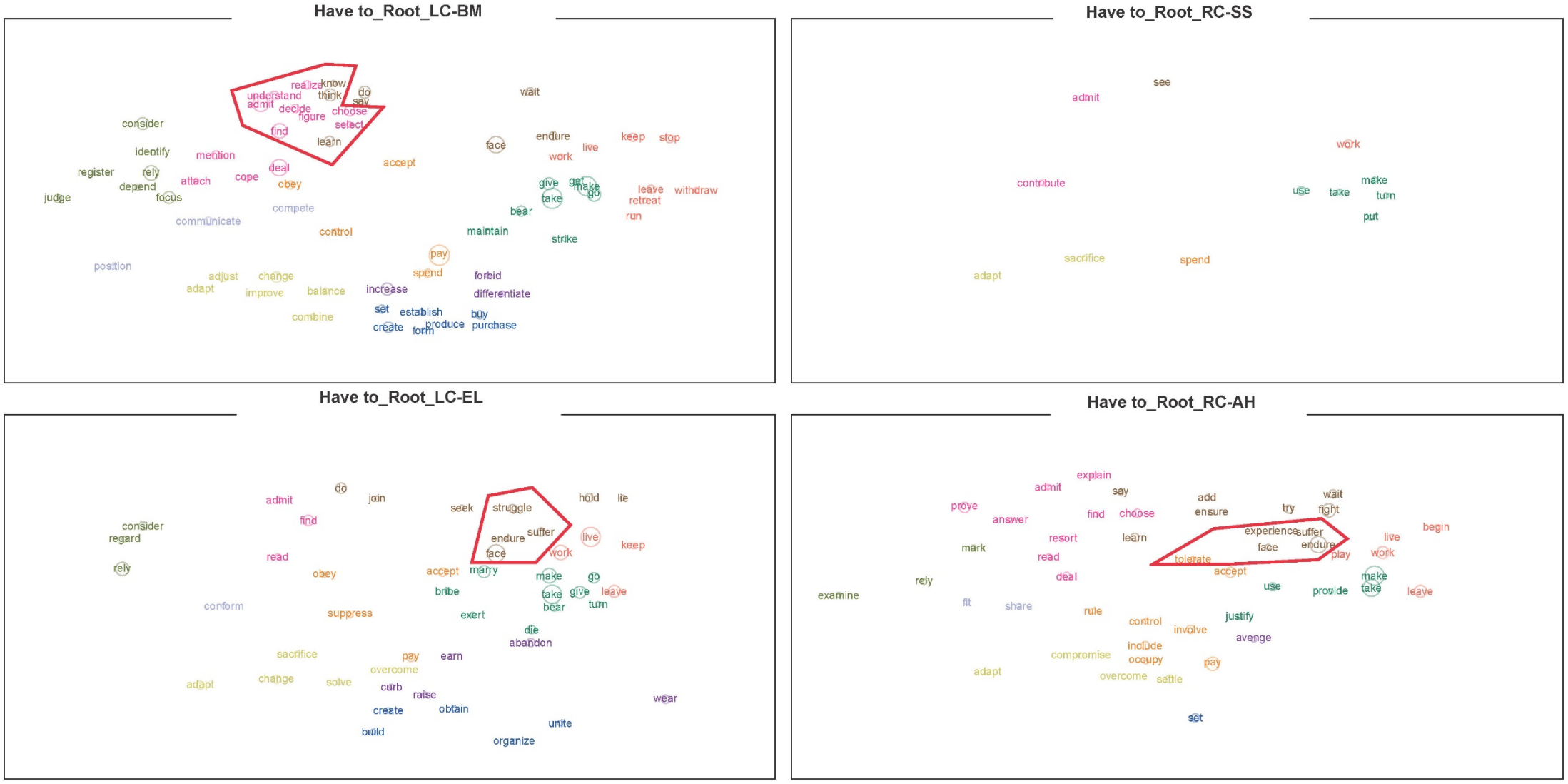


Figure 4.10 Distributional semantic plots of the verb collocates of root *have to* in the four sub-corpora

The figure indicates that root *have to* co-occur with a wider range of verbs in LC-BM than in RC-SS. This is partly due to the low absolute frequency of root *have to* in RC-SS in general. In addition, one of the distinctive clusters in LC-BM denotes cognitive activities such as *realize* and *understand*, whereas this cluster is almost absent in the other three sub-corpora, except *admit* (which will be discussed shortly). It seems that Chinese students in LC-BM prefer to give suggestions at the cognitive level to achieve a goal, as illustrated in 4-34 and 4-35.

4-34 The advertiser *has to* understand the product deeply with the sense of the country producing it. (Root_LC-BM_LC-BM 04607)

4-35 Before officially enter the Chinese market, the companies *have to* know the attitude of the Chinese citizens toward their own countries. (Root_LC-BM_UGBM 11009)

The variety of verb collocates in LC-EL and RC-AH, on the other hand, does not show marked differences. There is one similar verb cluster in these two sub-corpora, denoting actions such as *suffer* and *endure*. This cluster in LC-EL includes verbs such as *suffer* and *struggle*. These verbs are likely to be followed by unpleasant experiences, as illustrated in 4-36 and 4-37 below. Both *suffer* and *struggle* co-occur with root *have to* three times in LC-EL. The cluster in RC-AH includes not only these verbs but also *tolerate* and *experience*, demonstrating a broader range of such verbs, as exemplified in 4-38 below.

4-36 When the successive storms destroy Grampus, Pym and his group *have to* suffer intolerable starvation and thirst. (Root_LC-EL_L 02508)

4-37 Living in the peaceful Marsh Farm away from the outside world of modern industrialization, there is not much difficulty for her to quietly preserve her female self inside, for she can easily ignore the outside world. However, Anna is more exposed to the increasing industrialization. Thus, she *has to* struggle bitterly to keep her own self from being dominated by her husband Will and the outside world. (Root_LC-EL_L 07409)

4-38 As Anne Dormer possessed no power to prevent the 'oppressions I have lain', lacking in physical strength and the law unable to protect her, she simply *had to* tolerate the physical demands of her husband. (Root_RC-AH_0039f)

In 4-36, the description of 'the successive storms destroy Grampus' implies that Pym and his group are in an extreme situation and the obligation to suffer seems to stem from the urgent need to survive rather than from an external authority. The life-threatening environment creates an immediate need to find food and water, compelling Pym and his group to respond to these pressing circumstances. In 4-37, Anna's struggle is unavoidable and *thus* implies that the reasons are presented in previous sentences. The use of *have to* highlights that her efforts are not optional but are necessary responses to the external pressures that threaten her individuality such as interaction with her husband and broader background of increasing industrialisation. Similarly, the last example, 4-38, also starts with presenting the reasons for the

necessity for Anne to tolerate, that are her lack of physical strength and insufficient legal protection. *Have to* expresses the obligation that arises not from a specific party or person but from the overall situation she is in.

The three examples analysed above share a feature that the source of the obligation in these examples is not the authority, but the objective situations or circumstances, and it can be paraphrased as 'it is necessary for'. According to Coates (1983), root *have to* is mostly used to demonstrate the objective sense of the obligation, whereas root *must* is rarely used when the source of obligation is external. Thus, replacing root *have to* with *must* may not be suitable in these instances, as *have to* better conveys the external forces that compel the actions or responses of the characters in the literary work. This demonstrates that both Chinese and British students understand the difference in subjectivity between root *have to* and root *must*, and are able to appropriately choose between these two modals.

This verb cluster related to unpleasant experiences also reveals disciplinary variation in that it is almost absent in the semantic plots of LC-BM or RC-SS. One explanation is that writings in LC-EL and RC-AH tend to require the students to analyse the characters' behaviour or describe general situations. The use of root *have to* with these verbs is to emphasise the unavoidable actions needed to be taken based on the situations or background described, a pattern that is not commonly observed in LC-

BM and RC-SS.

Another disciplinary variation is related to the use of root *have to* with *admit*. The absolute frequency of root *have to* used with *admit* in three sub-corpora is two, and it is eight in LC-BM. Considering the difference in the absolute frequency of root *have to* in each sub-corpus (see Section 4.5.1, Table 4.12), root *have to* appears to show a stronger association with *admit* in LC-BM compared to the other three sub-corpora. It reflects a recognition of limitations or unexpected findings. Moreover, Chinese students in LC-BM use this combination mostly with *we* (6 out of 8 instances) to engage with the readers to agree with their viewpoints, as exemplified in 4-39.

> 4-39 A great amount of products are sold around the world. However, we *have to* admit that the majority of China's export commodities do not have their own brands. (Root_LC-BM_UGBM 05507)

4.6 Profiles of *should* across four sub-corpora

This section will discuss the profiles of the last modal in question, *should*. The meaning distribution, dispersion, co-occurrence with syntactic features and verb collocates will be examined and compared in the four sub-corpora.

4.6.1 Meaning distribution of *should* in the four sub-corpora

Table 4.16 presents the absolute frequency of different meanings of *should* in the four corpora, and the normalised frequency per million words is illustrated in Figure 4.11.

Table 4.16 Absolute frequency of different meanings of *should* in the four sub-corpora

	Epistemic	Root	Unclear	Total
LC-BM	51	1,613	9	1,673
RC-SS	24	249	5	278
LC-EL	42	594	5	641
RC-AH	40	332	7	379

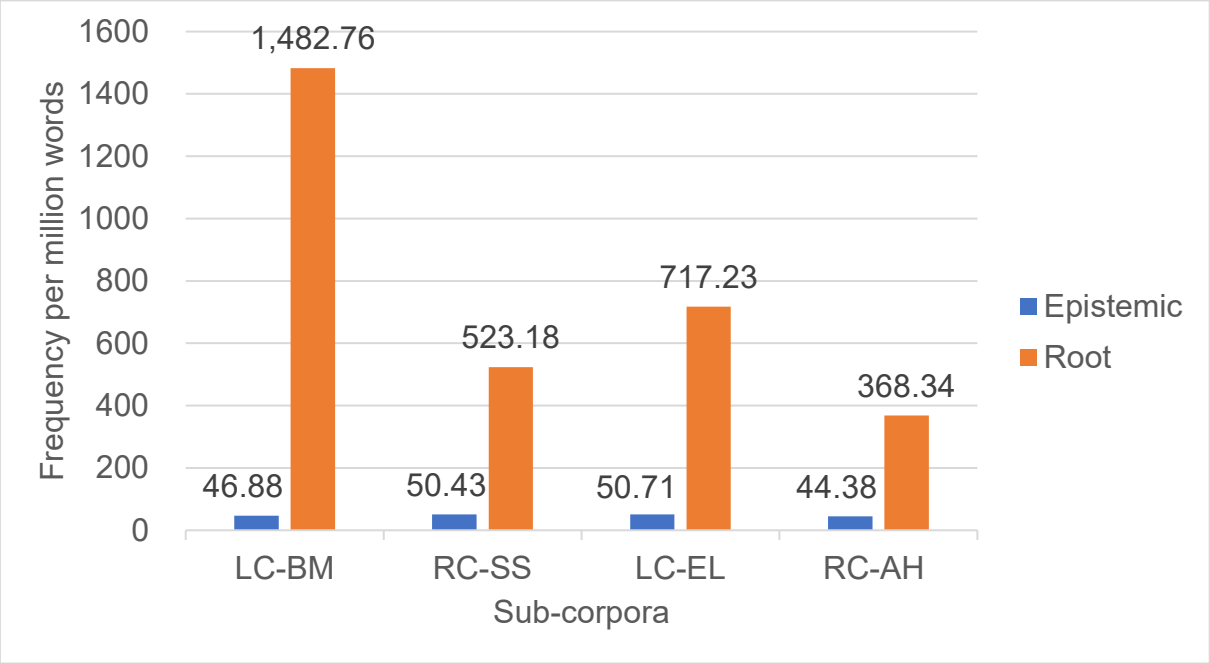


Figure 4.11 Normalised frequency per million words of epistemic and root *should* in the four sub-corpora

As indicated by the table and figure, epistemic *should* is used similarly in terms of its normalised frequency across the four sub-corpora. Root *should*, on the other hand,

shows differences between student groups and between disciplines. Chinese students in both disciplines use markedly more root *should* than their British counterparts, and the difference is larger between writings in LC-BM and RC-SS than between those in LC-EL and RC-AH.

One explanation is that root *should* is the first modal introduced to Chinese students to give suggestions in their junior high school English textbooks. Consequently, the students may be more familiar with root *should* compared to the other two modals, potentially leading to its over-representation in the learner corpus. Another reason might be that compared to root *must*, root *should* generally express a relatively weak sense of obligation. Since students do not act as the authority to give advice, root *should* seem to express a more reasonable degree of force in terms of the suggestion.

As for the disciplinary variation, writings of the two student groups show a similar pattern, with a higher normalised frequency of root *should* in LC-BM and RC-SS compared to LC-EL and RC-AH. Specifically, Chinese students in LC-BM use root *should* twice as often as those in LC-EL, and the frequency is 1.5 times higher in RC-SS than in RC-AH. One explanation, as mentioned in Section 4.4.1, is that writings in applied disciplines (e.g., LC-BM and RC-SS) tend to emphasise the impact on practice, and thus may include more root *should* to give suggestions.

4.6.2 Dispersion of *should* in the four sub-corpora

As to the dispersion of *should*, its range% and Juilland's D are shown in Table 4.17.

Table 4.17 Dispersion (range% and Juilland's D) of *should* in the four sub-corpora

	Epistemic		Root	
	Range%	Juilland's D	Range%	Juilland's D
LC-BM	21.43%	0.82	98.21%	0.94
RC-SS	12.78%	0.74	64.66%	0.89
LC-EL	21.90%	0.80	89.05%	0.92
RC-AH	10.61%	0.80	53.03%	0.90

As shown in the table, epistemic *should* is less evenly distributed across the texts than its root sense. The highest range% and Juilland's D value are observed in LC-BM, where root *should* is absent from only three out of 168 texts. Both senses of *should* are more evenly distributed in both sub-corpora of the learner corpus than in those of the reference corpus, and they show similar dispersion between disciplines.

4.6.3 Co-occurrence of *should* with syntactic features in the four sub-corpora

This section will discuss the profiles of *should* in terms of its co-occurrence of several syntactic features, including negation, voice, and aspect.

Epistemic *should* is not used in negation in the learner corpus, and only two instances are found in RC-AH in the reference corpus, with one example shown in 4-40 below.

4-40 I do not trust this story, because if the girl had successfully kept the miscarriage a secret, Plutarch *should* not have been aware of it, unless the foetus had been found at a later date, enabling the story to be told, or if he was the father of the unwanted child. (Epistmeic_RC-AH_6109c)

Root *should* is more frequently used in negation compared to its epistemic sense, accounting for 3.53% of all instances of root use in LC-BM and 11.11% in LC-EL. The percentages in the reference corpus are 16.87% in RC-SS and 13.55% in RC-AH. It seems that Chinese students use less root *should* in negation than their British counterparts. As for the disciplinary variation, the use of root *should* in negation is less frequent in LC-BM than in LC-EL, while the opposite pattern is observed in the two disciplines in the reference corpus.

As for the voice, Table 4.18 shows the normalised frequency per million words of *should* used in the passive voice in the four sub-corpora.

Table 4.18 Normalised frequency per million words of *should* used in the passive voice in the four sub-corpora

	Epistemic	Root
LC-BM	1.84	386.09
RC-SS	8.40	268.94
LC-EL	1.21	185.95
RC-AH	6.66	143.12

Epistemic *should* in the passive voice is under-represented in the learner corpus compared to the reference corpus, whereas the case for root *should* shows an opposite trend. Chinese students use approximately 1.4 times more root *should* in passives in both disciplines than their British counterparts. It seems that Chinese students tend to omit explicit mention of the source of the obligation to emphasise the action advised to take. Another explanation might be that they try to describe the suggestion objectively. As Biber et al. (1999, p. 475) observed, 'an objective detachment' is expressed by the use of the passive in academic prose. In terms of disciplinary variation, epistemic *should* is slightly over-represented in passives in LC-BM and RC-SS compared to LC-EL and RC-AH respectively. The use of root *should* shows a similar trend, and exhibits a more distinct difference in normalised frequency across these disciplines.

In regard to aspect, Table 4.19 shows the co-occurrence of *should* with the perfect aspect across the sub-corpora.

Table 4.19 Absolute and normalised frequency per million words of *should* used in the perfect aspect in the four sub-corpora

	Epistemic		Root	
	AF	NF	AF	NF
LC-BM	4	3.68	5	4.60
RC-SS	4	8.40	2	4.20
LC-EL	9	10.87	9	10.87
RC-AH	8	8.88	11	12.20

As shown in the table, epistemic *should* in the perfect aspect is under-represented in the writings in LC-BM compared to those in RC-SS, whereas the trend is the opposite when comparing the writings in LC-EL and RC-AH. Regarding the disciplinary variation, this co-occurrence is under-represented by students in LC-BM and RC-SS compared to those in LC-EL and RC-AH. Epistemic *should* used with the perfect aspect indicates a prediction regarding what happened in the past, as shown in 4-41.

> 4-41 If the crisis had been recognized in this stage, it *should* have been very easy to deal with and been possible to prevent the crisis from going to the next stage. (Epistemic_LC-BM_UGBM 03509)

Compared to epistemic *should*, root *should* in the perfect aspect accounts for a lower percentage (less than 4%) among all the instances of root *should* in all four sub-corpora. Its co-occurrence with the perfect aspect does not show marked differences between the two student groups, but it is observed to be under-represented in LC-BM and RC-SS compared to LC-EL and RC-AH.

In addition, Coates (1983) classifies two situations where root *should* is used in the perfect aspect, as discussed in Section 2.2.5. One is the genuine perfective, and the main predication is habitual, as exemplified in 4-42 below. It can be interpreted as ‘It

would be advisable that the reputation of the brand founder have played an important role in their brand build.'. The other usage of the perfect aspect is to describe the suggested actions in the past in a hypothetical sense. It is to emphasise that the subject has not taken the recommended action. 4-43 demonstrates the recommended action for Ping An group, which they failed to implement.

4-42 Fourthly, the reputation of the brand founders *should* always have been play important role in their brand build. (Root_LC-BM_UGBM 10105)

4-43 If a profound investigation about the assets of Fortis had been made, Ping An Group *should* have thought twice before making such a risky investment (Root_LC-BM_UGBM 03109)

Among the five instances observed in LC-BM, two of them are used as the genuine perfective, while the remaining three are used in the hypothetical sense. All the two instances in RC-SS, on the other hand, are used in the latter way, and one example is 4-44 shown below. The hypothetical sense is delivered by both the conditional clause and root *should*. Eight out of nine cases are categorised as the hypothetical use in LC-EL, as illustrated in 4-45, and the number is eight out of eleven instances in RC-AH, with one example provided in 4-46. Example 4-46 describes a hypothetical situation in which Indians still had the chance to devise strategies to drive the Euro-Americans from the continent, but in reality, they made the wrong decision. It seems that both groups of students have a preference for using root *should* to imply that the suggestion

is actually not taken, and this use is more frequently used in LC-EL and RC-AH compared to the other two sub-corpora.

4-44 If the pact was obeyed to its law, these countries *should* have been fined. (Root_RC-SS_0202c)

4-45 […] he has helped himself to big portions of the shrimp, not realizing he *should* have taken only a spoonful, until everybody has taken a little. (Root_LC-EL_L 08008)

4-46 'Indians yearned to destroy the Euro-Americans and drive them from the continent' but by this stage it was too late. They *should* have acted sooner, but made the wrong decision in choosing trade over annihilation. (Root_RC-AH_0029n)

Apart from the perfect aspect, the progressive aspect is also used in both corpora, but it only co-occurs with root *should*. There are five instances of root *should* used in the progressive aspect in the learner corpus, and all of them are in LC-BM. One example is 4-47, which emphasises the current state of what the subject is advised to do. In the reference corpus, the number of co-occurrences is two in RC-SS and one in RC-AH, as illustrated in 4-48 and 4-49 respectively.

4-47 Culture gives people a sense of who they are, of belonging, of how they should behave, and of what they *should* be doing. (Root_LC-BM_UGBM 03105)

4-48 The EU enjoys the privileged position of representing a large group of Western countries, some of whom enjoy extreme riches and wealth, and therefore it *should* be pushing hard to dramatically improve the state of the environment. (Root_RC-SS_0244m)

4-49 It is all too easy for us to wish to relieve the pain of a friend, whilst underneath our supposed 'suffering with' him, we are really relishing their active gratitude and homage, indeed one may be familiar with the feeling of rejection often felt when a friend in need has refused our help, when really we *should* be admiring his independence and strength. (Root_RC-AH_0407d)

4.6.4 Verb collocates of *should* in the four sub-corpora

Having discussed the distribution of syntactic features used with *should*, I will now move on to explore its verb collocates.

4.6.4.1 Verb collocates of epistemic should

Similar to epistemic *must*, epistemic *should* frequently collocates with main verb *be*. It accounts for 70.59% of all epistemic uses in LC-BM, 33.33% in RC-SS, 45.24% in LC-EL, and 47.50% in RC-AH. Table 4.20 below presents the distribution of the co-text following this combination.

Table 4.20 Normalised frequency per million words of the co-text following epistemic 'should + be' in the four sub-corpora

	Adjective	Noun	Preposition	Clause	End of the sentence	Total
LC-BM	7.35	22.98	0.92	1.84	0.00	33.09
RC-SS	4.20	8.40	4.20	0.00	0.00	16.81
LC-EL	6.04	12.07	1.21	3.62	0.00	22.94
RC-AH	12.20	6.66	0.00	0.00	2.22	21.08

The table indicates that the over-representation of 'should + be' in the learner corpus is mostly due to the higher frequency of this combination being used before the noun, especially in LC-BM. Two examples taken from the learner corpus are illustrated in 4-50 and 4-51.

4-50 The result *should* be an increase in community cohesion, a reduction in poverty and […]. (Epistemic_LC-BM_UGBM 06005)

4-51 All his callous treatment *should* be a reason for Fanny's tragedy. (Epistemic_LC-EL_L 08504)

In addition, there is one structure that is exclusively used by each student group. It is 'should + be + clause' in the learner corpus and 'should + be' placed at the end of a sentence in the reference corpus, as exemplified in 4-52 and 4-53 respectively. As for the disciplinary difference, 'should + be + noun' is over-represented in LC-BM compared to LC-EL, and 'should + be + adjective' is used more frequently in RC-AH

than in RC-SS. One of the examples in RC-AH is shown in 4-54.

4-52 But the reasonable, according to the Maslow Hierarchy of Needs theory, relationship between them *should* be that the former is premise to the latter. (Epistemic_LC-EL_L 04908)

4-53 It should be stressed that a stereotype is not a fact; it is not what something is but what a majority perceives it to be or what a majority believes it *should* be. (Epistemic_RC-AH_6066c)

4-54 During this period in England the patriarchal ideal was that women *should* be quiet and subservient to their husbands. (Epistemic_RC-AH_0040b)

Thus far, we have discussed the case of *should* collocating with main verb *be*. The following paragraphs will explore other verb collocates. Figure 4.12 below shows the semantic plots of verb collocates of epistemic *should* in each sub-corpus.

Figure 4.12 Distributional semantic plots of the verb collocates of epistemic *should* in the four sub-corpora

The verb collocates in the learner corpus show generally less variety than those in the reference corpus, partly due to their more frequent use of epistemic *should* with *be*, as mentioned above. It is hard to identify characteristic verb clusters in LC-BM and RC-SS as most of the verbs are loosely distributed in the plots. By contrast, there is one distinctive verb cluster in LC-EL and RC-AH respectively. The cluster in LC-EL denotes emotional states such as *feel* and *envy*, while in RC-AH, British students use verbs indicating attitudes such as *expect* and *want* with epistemic *should*. Examples are illustrated in 4-55 and 4-56 in turn.

4-55 […] as she feels she is better than having a furze-cutter for a husband, and feels that Clym *should* feel the same way. (Epistemic_LC-EL_ L 10907)

4-56 When educated senators and deep intellectuals look on this type of behavior it is quite within reason that they *should* want him out of office, stoic or not. (Epistemic_RC-AH_ 6195b)

These two clusters are related to each other as they both include mental verbs, but they are exclusively used in LC-EL and RC-AH respectively. In addition, they are absent in the writings of LC-BM and RC-SS. One reason could be that students in LC-EL and RC-AH often engage in the analysis of human behaviours, which may lead to more evaluations of human emotions and attitudes compared to those in LC-BM and RC-SS.

4.6.4.2 Verb collocates of root should

Figure 4.13 below shows the distributional semantic plots of verb collocates of root *should* in each sub-corpus.

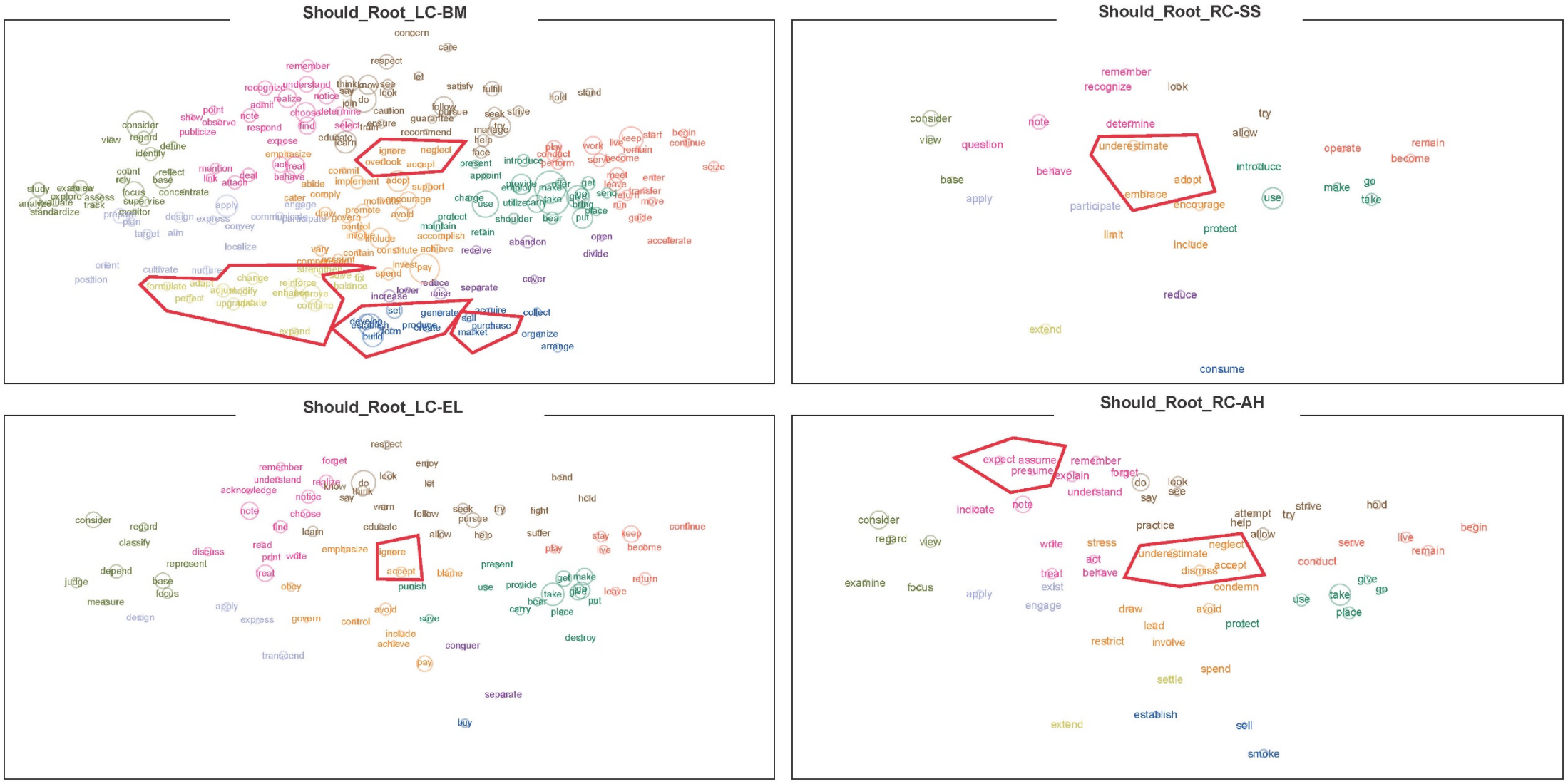


Figure 4.13 Distributional semantic plots of the verb collocates of root *should* in the four sub-corpora

The figure shows that verb collocates of *should*, similar to those of *must*, show a high degree of semantic homogeneity in LC-BM than in RC-SS. Three verb clusters at the bottom half of the plot are distinctively used in LC-BM, including those denoting improvement (e.g., *adjust*, *modify*, and *adapt*), development (e.g., *generate* and *create*), and description closely related to the disciplinary knowledge (e.g., *sell*, *purchase*, and *market*). The examples are shown in 4-57, 4-58 and 4-59 respectively. The absence of these three clusters in RC-SS may be explained by differences in cultural values and disciplines covered, which will be discussed further in Section 6.4.2.

4-57 More importantly, customers' demands are changing and growing all the time, so marketers *should* adjust or update marketing strategies timely. (Root_LC-BM_UGBM 02005)

4-58 In the material layer, China *should* create a healthier environment for taxation. (Root_LC-BM_UGBM 00909)

4-59 For example, if a company sell pork related food, then it *should* never sell its product in restricts where Hui people live. (Root_LC-BM_UGBM 05607)

A cluster of verbs including *accept*, *underestimate*, and *dismiss*, among others, is used with root *should* by both groups of students, but serving different purposes. British students prefer using these verbs with root *should* to assess arguments and viewpoints, as exemplified in 4-60 and 4-61 below. Conversely, although Chinese students

sometimes use this combination in a similar manner, they mostly use it to give suggestions for business practices in LC-BM (three out of eight instances) or for characters in LC-EL (six out of seven instances), thereby engaging more with actions rather than with propositions, as demonstrated in 4-62 and 4-63.

4-60 Lastly, following the decline of ideology as a legitimating force for the CCP, the importance of providing the people with what they want *should* not be underestimated. (Root_RC-SS_0135f)

4-61 Although these revolts can then be seen to have been driven by economic hardship, their potential to represent political behaviour *should* not be dismissed. (Root_RC-AH_0144b)

4-62 International business implies knowledge and an understanding of the behavior, the culture, the customs and the needs of the customers. For example firms *should* not neglect people in Chinese market which is an emerging one. (Root_LC-BM _UGBM 04207)

4-63 Mr. Collins proposes to two women in less than three days. From the point, he is not serious to love. Were Charlotte clever, she *should*n't have accepted his proposal, but why she still chooses such a man to marry? (Root_LC-EL_ L 10505)

Another difference between the two student groups is the use of mental verbs placed at the top left of the plots. Chinese students use a wider range of these verbs with root *should* compared to their British counterparts, and one of the reasons could be the

higher absolute frequency of root *should*. 4-64 is an example in LC-BM where *realize* is used with root *should* to give suggestions on how the company can improve their performance at the cognitive level. The instances in LC-EL are sometimes used with *we* to ask the reader to agree with the writer on what actions are suggested to be taken, as shown in 4-65. Besides, this verb cluster is also used with root *should* in LC-EL to describe the obligation laid by the characters in the literary works, as in 4-66. British students display a less diverse use of mental verbs, suggesting a potential reluctance to offer suggestions at a cognitive level.

4-64 Consequently, manufacturers *should* realize the problem customers need to solve and provide a tool with strength and specialty to stand out and differentiate the tool with others. (Root_LC-BM_UGBM 01209)

4-65 As a result, facing the prejudice of genders in different cultures, the first thing we *should* realize is that gender is independent from culture and there should not be any prejudice. (Root_LC-EL_L 01608)

4-66 In the parents' opinion, Ted's profession did not allowed him to have a yellow-skinned wife, and Rose *should* realize it and leave by herself. (Root_LC-EL_L 07707)

The use of mental verbs also shows disciplinary variation in that, as discussed above, the combination of root *should* and mental verbs seems to serve more functions in LC-EL than in LC-BM. As for the two disciplines in the reference corpus, there is a

characteristic cluster in RC-AH that is absent in RC-SS or in the learner corpus, which includes verbs related to expectations such as *expect*, *assume*, and *presume*, as illustrated in 4-67 and 4-68.

4-67 However we *should* not always assume that the development of single grave burials demonstrates the status of the individual and that multiple burials do not. (Root_RC-AH_6060a)

4-68 Therefore if localisation theories are to be accepted they *should* be expected to include bilinguals, illiterates and populations who speak tone languages but Bassoal. (Root_RC-AH_6174d)

Another disciplinary variation is that, as mentioned earlier, one cluster in LC-BM is closely related to disciplinary knowledge, whereas it is difficult to identify such a cluster in LC-EL. In addition, similar to root *must*, root *should* is found to be frequently used with the pronoun *we* in LC-EL and RC-AH compared to LC-BM and RC-SS, accounting for 12.29 and 18.37% of all the instances respectively, and one of the examples is 4-67 shown above. In LC-BM and RC-SS, the percentages are only about 4%.

4.7 Summary

This chapter has reported the quantitative findings regarding the profiles of *must*, *have to*, and *should*. Initially, the analysis focuses on the overall frequency and meaning distribution of these three modals between the learner and the reference corpus. There

is a notable correlation between the students' first languages and their use of the three modals. Further examination reveals that epistemic use of these modals does not show statistically significant differences between the two student groups. In contrast, root sense of these modals demonstrates a significant association with the students' first languages. Specifically, root *must* is under-represented by Chinese students, whereas root have to and should show an opposite pattern.

Following this overview, each modal is discussed separately in terms of meaning distribution, dispersion, co-occurrence with syntactic features, and semantics of the main verbs used with the modals. In terms of dispersion, root use of the three modals shows a more even distribution across texts compared to their epistemic use. There are generally no significant variations in the usage of the three modals between student groups and between disciplines, though minor variations exist.

As for the co-occurrence with syntactic features, epistemic use of the modals does not exhibit significant differences when used with the passive voice across the four sub-corpora, whereas variations in their root use are more pronounced, possibly due to the characteristics of their verb collocates. Root *have to* shows disciplinary variation to be more frequently used in the past tense in LC-EL and RC-AH than in LC-BM and RC-SS. As for the aspect, a strong correlation is observed between epistemic *must* and the perfect aspect, and this association is more prevalent in the reference corpus.

Additionally, root *should* serves two functions in the perfect aspect, with one function—used in a hypothetical sense—being more commonly used by both student groups. Findings on the remaining two aspects, the meaning distribution and semantics of the main verbs co-occurring with the three modals, are outlined in Tables 4.21, 4.22, and 4.23 below with each table presenting findings for one of the modals.

This chapter has helped to build comprehensive quantitative profiles of the three modals in Chinese EFL students' academic writing and provided a new perspective by exploring the semantics of their verb collocates. Similarities, but mostly differences, were revealed between the two student groups and between the disciplines.

The chapter that follows moves on to present the qualitative findings, providing a finer-grained view of the modal use in academic writing.

Table 4.21 Summary of the meaning distribution and verb collocates of *must* across four sub-corpora (LC = learner corpus, RC = reference corpus, BM = Business and Management, EL = English Literature, SS = Social Science, AH = Arts and Humanities)

Meaning	Factor	Meaning distribution (see Section 4.4.1)	Semantics of the verb collocates (see Section 4.4.4)
Epistemic	Student group variation	• Slightly under-represented in LC compared to RC	• Frequently used with main verb *be*. Among those instances of 'must +be' followed by a noun, over 40% of them preceded by the existential subject *there* in all sub-corpora • Equally used with mental and causative verbs across the four sub-corpora, with a slightly more use of causative verbs in LC-EL
	Disciplinary variation	• Markedly under-represented in LC-BM and RC-SS compared to LC-EL and RC-AH	• 'Must + be + adjective' is used less frequently by students in LC-BM and RC-SS than those in LC-EL and RC-AH
Root	Student group variation	• Under-represented in LC compared to RC	• Used with three verb clusters denoting exploration/evaluation, improvement, and development in LC-BM, but only one of them can be observed in RC-SS • Used with verbs related to statistics in RC-SS exclusively • Chinese students in LC-EL use fewer mental verbs with root *must* compared to their British counterparts in RC-AH • British students in both disciplines use verbs with root *must* to challenge and critically treat propositions, whereas this cluster is not observed in the writings of Chinese students
	Disciplinary variation	• Markedly over-represented in LC-BM and RC-SS compared to LC-EL and RC-AH	• Verb collocates denoting exploration are used in LC-BM and RC-SS but not in LC-EL and RC-AH • Root *must* is frequently used with the pronoun *we* in LC-EL and RC-AH, whereas this is not prevalent in LC-BM and RC-SS

Table 4.22 Summary of the meaning distribution and verb collocates of *have to* across four sub-corpora (LC = learner corpus, RC = reference corpus, BM = Business and Management, EL = English Literature, SS = Social Science, AH = Arts and Humanities)

Meaning	Factor	Meaning distribution (see Section 4.5.1)	Semantics of the verb collocates (see Section 4.5.4)
Epistemic	Student group variation	• rarely used across the four sub-corpora	• Hard to identify patterns due to its low absolute frequency • Mostly used with main verb *be* and stative verbs
	Disciplinary variation		
Root	Student group variation	• Under-represented in RC-SS compared to the other three sub-corpora	• Used with verbs denoting cognitive activities in LC-BM but not in the other three sub-corpora
	Disciplinary variation		• Used with verbs related to unpleasant experiences in LC-EL and RC-AH but this use is almost absent in LC-BM and RC-SS • Shows a stronger association with *admit* in LC-BM compared to the other three sub-corpora

Table 4.23 Summary of the meaning distribution and verb collocates of *should* across four sub-corpora (LC = learner corpus, RC = reference corpus, BM = Business and Management, EL = English Literature, SS = Social Science, AH = Arts and Humanities)

Meaning	Factor	Meaning distribution (see Section 4.6.1)	Semantics of the verb collocates (see Section 4.6.4)
Epistemic	Student group variation	• Distributed similarly across the four sub-corpora	• 'Should + be + noun' is used more frequently in LC than in RC
	Disciplinary variation		• 'Should + be + noun' is used more frequently in LC-BM than in LC-EL • 'Should + be + adjective' is used less frequently by students in RC-SS than those in RC-AH • Used with verbs denoting emotions and attitudes in LC-EL and RC-AH respectively, but these two verb clusters are absent in LC-BM and RC-SS
Root	Student group variation	• Markedly over-represented in LC compared to RC	• Used with three verb clusters related to improvement, development, and disciplinary knowledge in LC-BM, but they are not observed in RC-SS • Both students use root *should* with verbs to assess arguments and viewpoints, whereas Chinese students use these verbs to propose suggested actions • Used with a wider variety of mental verbs in LC compared to RC
	Disciplinary variation	• Markedly over-represented in LC-BM and RC-SS compared to LC-EL and RC-AH	• Students in LC-EL use root *should* with mental verbs to serve more functions than those in LC-BM • There is a distinctive verb collocate cluster in RC-AH denoting expectations which is absent in the other three sub-corpora • Used with verbs related to disciplinary knowledge in LC-BM but not in LC-EL • Frequently used with the pronoun *we* in LC-EL and RC-AH, whereas this is not prevalent in LC-BM and RC-SS

5 QUALITATIVE ANALYSIS OF *MUST*, *HAVE TO*, AND *SHOULD* BETWEEN STUDENT GROUPS AND BETWEEN DISCIPLINES

5.1 Introduction

The previous chapter has presented the quantitative profiles of *must*, *have to* and *should*, and revealed variations between student groups and between disciplines through the comparisons across the four sub-corpora. This chapter aims to provide a finer-grained view of the modal use in Chinese EFL student academic writing by examining 16 sample texts, with four texts selected from each sub-corpus. This approach helps to elaborate on what is found in the quantitative analysis and, in the meantime, reveal anything that might have been overlooked. Comparisons are also made between sample student groups and between disciplines.

Given the reduced volume of texts for review, the qualitative analysis extends beyond prior explorations, investigating the distinctive features of the modal use in academic writing, including the modal distribution in different parts of a text and textual voice expressed by the modals. In addition, as mentioned in Section 3.2.2, a compromise has been made to select comparable disciplinary groups rather than specific disciplines in the reference corpus for quantitative analysis, owing to a scarcity of texts.

To address this limitation, this chapter conducts the qualitative analysis using more comparable texts to see whether the patterns identified in the quantitative analysis are also observable in these texts.

The overall structure of this analysis is similar to that of quantitative analysis. Section 5.2 describes the methods of the selection and annotation of the sample texts. It is then followed by the overall frequency and meaning distribution of *must*, *have to*, and *should* in the learner and the reference sample. In Section 5.4, a more detailed examination of the profiles of the three modals is conducted across the four sub-samples to identify variations between student groups and between disciplines, and Section 5.5 summarises this chapter.

5.2 Method

This sub-section presents the method used for the qualitative analysis, which consists of two parts. One is to justify the selection of learner and reference sample texts. The other is to explain how the sample texts are annotated to fulfil the objectives of this chapter.

5.2.1 Selection of the sample texts

A total of 16 texts are examined in the qualitative analysis, with four texts from each sub-corpus. The decision on the number of texts to analyse is informed by two

considerations. First, the number of texts should offer sufficient data to yield meaningful insights, while the analysis itself is feasible within the given timeframe for the completion of the doctoral thesis. 16 texts in total seem to be a manageable number and could reveal meaningful patterns.

In addition, as mentioned in Section 3.2, there are 168 texts in LC-BM (Learner Corpus – Business and Management) and 137 in LC-EL (Learner Corpus – English Literature) written by Chinese EFL undergraduates. However, after filtering for factors to ensure comparability with the learner corpus, the reference corpus includes only five texts in Business and 47 in English, the two disciplines that most closely approximate those in the learner corpus. In contrast to quantitative analysis, which requires a large number of texts and thus uses texts in similar disciplinary groups rather than specific disciplines in the reference corpus for comparisons, qualitative analysis allows for a more precise and detailed examination by comparing the texts in the closest approximate disciplines. Given the limited texts in the discipline of Business, with only five texts in total, selecting four texts from each sub-corpus seems to be a practical and justified choice.

The selection process is different for the learner and the reference corpus. In the case of the learner corpus, which comprises a substantial number of texts (168 in LC-BM and 137 in LC-EL), a random sampling method is necessary. To ensure

representativeness, texts with an extremely high or low total number of the target modals were excluded from the sampling process because the modal use in these texts might be highly related to the content or individual preference.

The detailed sampling procedures are presented as follows. The concordance lines of all three modals were extracted using the CQL query [lemma = "must"] | [lemma = "should"] | [lemma = "have"] [lemma = "to"] in Sketch Engine. The extracted instances, however, are not all used as modals (see Section 4.2.1 for a detailed discussion). Thus, I went through the meanings/labels of the modals that were previously annotated in the quantitative analysis and excluded the instances that are not modals. Unlike in the quantitative analysis, direct quotations containing the modals were included in the analysis to explore the textual voice expressed by the modals and how students present their viewpoints through these quotations.

After the clean-up process mentioned above, the frequency of the three modals in total reduced from 3,946 to 3,921. The frequency function in Sketch Engine was then used to calculate the total frequency of the three modals in each text. The absolute frequency was not used for sampling due to the considerable variations in token numbers across different texts. The number of tokens in LC-BM varies from 12,499 to 3,133, and in LC-EL, it ranges from 10,184 to 2,567. Thus, total normalised frequencies

of the three modals in each text in LC-BM and LC-EL were used for sampling. I selected a base of 10,000 tokens for normalisation instead of one million in this chapter to ensure comparability relative to the token size of the texts. Normalising per million words might inflate small absolute frequencies, potentially leading to a misperception of these rare uses. The results are displayed in the form of a boxplot in Figure 5.1, generated by Lancaster Stats Tools online (Brezina, 2018).

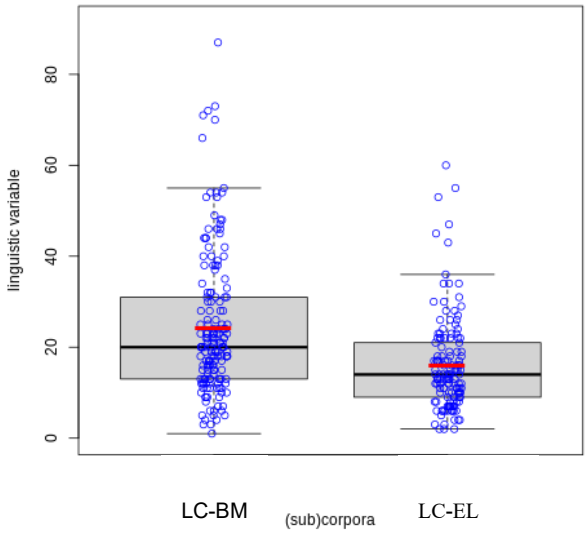


Figure 5.1 Boxplot of the total normalised frequency per 10,000 words of the three modals in each text in LC-BM and LC-EL in the learner corpus

Each dot represents one text, with the y-axis values indicating the total normalised frequency of the modals in each text. The grey box in the middle refers to the interquartile range, which includes the middle 50% of the values. Specifically, this range spans from the first quartile (25th percentile) to the third quartile (75th percentile).

Inside the box, the red short line represents the mean, and the black bold line refers to the median. The whiskers below and above the box represent the minimum and maximum values, while the individual values outside the scope of the whiskers are extreme values.

As shown in the figure, the total normalised frequency per 10,000 words of the three modals in each text varies markedly, from zero to over 80. To choose the representative texts for qualitative analysis, only texts in the interquartile range (those in the grey box) were used, thereby minimising the influence of extreme values. The file names of these texts from each sub-corpus were recorded in the first column of two separate spreadsheets, one for each sub-corpus. In the second column, random numbers generated by the Excel formula '=rand()' were listed. The two columns were sorted together in descending order based on the random numbers. The top four texts from each sub-corpus, as listed in the spreadsheets, were chosen for the qualitative analysis. Table 5.1 below presents the basic information of the eight learner sample texts. Despite their random selection, these texts were arranged within the table based on the ascending order of their file names to ensure systematic organisation. As mentioned in Section 3.2.1, the titles of the learner texts were not documented by Zou (2018) when compiling the corpus. These texts were thoroughly read and used as a reference for selecting comparable sample texts in the reference corpus.

Table 5.1 Selected learner sample texts in LC-BM and LC-EL

File name	Sub-corpus	Tokens
UGBM 04307	LC-BM	6,075
UGBM 05706	LC-BM	6,014
UGBM 10205	LC-BM	6,735
UGBM 10407	LC-BM	6,329
L 00504	LC-EL	6,369
L 02508	LC-EL	7,926
L 04106	LC-EL	5,783
L 10408	LC-EL	7,393

As for the reference corpus, the number of texts to select from is markedly smaller than that in the learner corpus, and each text is documented with titles and course information. Apart from the texts in the discipline of Business and English mentioned above, three texts categorised as *other* discipline were included in the selection as well since they are the assignments of the course of English. Thus, five texts in Business and 50 texts in English were used for subsequent selection. I first skimmed the titles and module names of the texts to exclude those that differ markedly from the learner sample texts and then reviewed the full texts to make a more reliable judgement. In the end, four texts each from Business and English (or *other* discipline) were chosen as the reference sample. Details of the texts are listed in Tables 5.2 and 5.3 below respectively.

Table 5.2 Selected reference sample texts in Business

Text ID	Title	Level	Module	Discipline	Grade	Words	Course
0202k	1. Why is quality important to EHL? (20% of the marks) 2. What are the underlying causes of the quality problems at EHL? (40% of the marks) 3. What steps would you advise Paul Stone to take to improve quality performance at EHL? (40% of the marks)	3	Business Studies II Operations Management	Business	M	1,551	Business
0202l	'Conflict is always a consideration in Managing Employment Relations (MER)'. Critically assess this statement in relation to different perspectives on the employment relationship and in relation to sectoral, labour force and unionisation issues	3	Business Studies II Managing Employment Relations	Business	M	1,635	Business
0202m	Critically discuss the extent to which group processes influence individual behaviour and group performance. In your answer, draw on theories and concepts from across the module, and illustrate your views with examples.	3	Business Studies II Organisational Behaviour	Business	D	1,551	Business
0202n	'Business strategies demand discipline in the execution of long-term strategic plans and flexibility to address emergent changes'. Discuss. Explain which one of the two features is more critical in your view.	3	Business Studies II Business Policy	Business	D	1,529	Business

Table 5.3 Selected reference sample texts in English

Text ID	Title	Level	Module	Discipline	Grade	Words	Course
0229b	Discuss love in Emma and The Sorrows of Young Werther.	2	The European Novel	Other	M	5,142	English
3007a	Consider the ways in which Orwell articulates the relationship between power and language in at least two texts	2	Modern Lit. core 1900-1955	English	D	2,827	BA English Studies
3008f	‘Women talk: men are silent: that is why I dread women’ (Dickinson). How is the question of gender and speech or eloquence rendered in any of the texts?	3	Special Subjects II - American Lit.	English	D	4,794	BA English Studies
3110a	Examine the relationship between language and spectacle in the texts of at least two authors	2	Renaissance Literature	English	D	3,870	BA English (major) with History of Art

The text identification (ID) corresponds to the file name, which consists of four digits representing the writers, and a lowercase letter denoting different assignments by the same writer. The level indicates which year the student is in, and the grade includes two types: merit (M) and distinction (D).

The four texts in English are written by four different students, as indicated by the different numbers in each text's ID. However, the texts in Business are written by the same student because student 0202 is the only person who fits all the criteria mentioned before. This may raise concerns about the representativeness and generalisability of the findings, as texts from one individual might not be able to fully reveal the patterns of how British students use the target modals in this discipline. I have tried to check whether there are other texts suitable for analysis and written by different students, such as texts in the discipline of Economics. However, I decided not to use these texts to maintain comparability, as they were not sufficiently similar to the learner sample texts. Thus, I continued to use texts written by student 0202 in Business, while being aware that the findings should be cautiously applied beyond the scope of this study.

In sum, four sub-samples are extracted from the two corpora. The learner and the reference sample are referred to as LS and RS respectively. Within these, the two sub-

samples in LS are labelled LS-BM (Business and Management) and LS-EL (English Literature), while those in RS are designated as RS-Business and RS-English. Table 5.4 presents the number of texts and the total token size for each sub-sample.

Table 5.4 Token numbers of the learner and the reference sub-samples

Sub-sample	Number of texts	Total tokens
LS-BM	4	25,153
LS-EL	4	27,471
RS-Business	4	6,266
RS-English	4	16,633

5.2.2 Annotation of the sample texts

After selecting the sample texts for qualitative analysis, the next step is to annotate the texts to explore the modal use. As mentioned in Section 3.4, the purpose of the qualitative analysis is twofold: firstly, to support the arguments made in the quantitative analysis and reveal the underlying reasons; secondly, to explore the profiles of *must*, *have to*, and *should* in a fine-grained view, complementing the quantitative analysis with a focus on their distinctive features in academic writing. With these objectives in mind, I followed Merriam and Tisdell's (2015) step-by-step process of qualitative analysis and applied it to my data. I began the analysis by reading the first sample text and annotating any relevant information in the margins that could address the research questions and fulfil the objectives of the analysis.

These annotations were reviewed and organised into categories, forming the basis for coding the next text. As more texts were examined, the categories expanded to accommodate new insights, reflecting the inductive nature of category construction. Initially, these categories were provisional and frequently adjusted. Ultimately, a relatively comprehensive coding scheme was developed. Instances of the target modals from the sample texts, presented in full sentences, were listed in a spreadsheet, and the coding categories were organised into columns. The coding categories are listed as follows:

- Meaning of the target modals
- Main verbs collocating with the target modals
- Co-occurrence with syntactic features, such as negation, voice, tense, and aspect
- Textual voice (averral or attribution) and the classification of attribution (direct/indirect quotations)
- Parts of the text in which the target modals are used
- Co-text proceeds and follows the modals
- Subject of the sentence

Most of the categories are straightforward to understand, while the annotation of text parts and textual voice may need more explanation. These two coding categories are

intended to highlight the distinctive features of the modal use in academic writing.

The distribution of the modals in different parts of a text could reveal if a modal is frequently used in one part of the text over another, helping to explain the differences in frequency distribution between student groups and between disciplines.

There are two approaches to explore the modal distribution across different parts of a text. One approach is to use the concordance plot function of AntConc 3.5.9 (Anthony, 2020). A search term list, ‘must, have to, has to, had to, should’, was inputted in AntConc, and the concordance plot function generated a figure showing the relative position of the modals in each text, as will be presented in Section 5.4.2.

While this approach could give us a general idea of how the modals are distributed in different parts of the text, it omits two pieces of information. One is the position of each modal in the text because all three modals are presented the same as vertical lines in the figure. The other is what meaning is expressed by the modals. In addition, the non-modal uses will also appear in the plot since the search term list includes all the forms of ‘must, have to, has to, had to, should’, which may not always function as modals, as noted in Section 4.2.1. Specifically, four texts, 0202k (RS-Business), 0202m (RS-Business), 3007a (RS-English), and 3110a (RS-English), each contain one modal that should be excluded from the analysis (the brackets following each file name indicate

the corresponding sub-sample name). However, manually removing these non-modal uses before generating the plot is challenging to achieve in AntConc.

Therefore, another approach is used, manually dividing the texts into different parts and presenting the modal use in a table. This can on the one hand, provide the two pieces of information that are overlooked by the figure generated by AntConc, and on the other hand, divide the texts into meaningful parts, helping to explain the variations in the modal use across the sample texts in a fine-grained view.

As mentioned in Section 3.2.2, both the learner and the reference texts share similar text parts, including an introduction, an analysis, and a conclusion. This division is applicable to the sample texts, and these parts were manually labelled in the texts. A further division of the text parts was conducted through the in-depth reading of the full sample texts. This is to investigate whether there are more detailed parts that constitute a significant portion of the texts, which may relate to the modal use. Some texts have signposting sentences indicating the overall structure at the beginning, which helps with the division. Although a more detailed description such as ‘historical background’ could be used to substitute the introduction part, it does not apply to all the texts. To ensure consistency and make comparisons across texts, it is not included as a text part.

Apart from the three generic parts mentioned above, the review of literature also constitutes a major part in the learner's sub-sample of LS-BM and in two texts in LS-EL (L 00504 and L 04106), whereas this part is almost absent in RS-English. In the case of RS-Business, the examination of literature is mostly used to facilitate the discussion of the topic, and it is integrated with the analysis rather than constituting an independent part. Thus, although 'literature review' is still recognised as one part of the text, it is absent in most reference sample texts. In sum, four parts were identified in the learner and the reference sample: introduction, literature review, analysis, and conclusion. The distribution of the different meanings of the modals across these parts was then displayed in a table, which will be presented in Section 5.4.2.

Another coding category that may need additional clarification is textual voice. As discussed in Section 2.3.1, researchers such as Sinclair (1988) and Tadros (1993) divided textual voice into two types, *averral* (the writer's voice) and *attribution* (the voice of a third party, such as antecedent authors), and this classification will be used in the present study. In addition, the attribution is further categorised into two sub-types, direct and indirect quotations, to explore how students integrate their viewpoints with those of others. Direct quotations are words taken directly from another person's work, whereas indirect quotations involve paraphrasing other people's words with an appropriate citation indicating the source.

A clarification of terms is necessary. As mentioned in Section 2.3.1, w*riter* refers to the Chinese and British students who produce the texts in the corpora. By contrast, *author* is used to describe the ones apart from the writers, such as those who wrote the previous literature (scholarly works) or the literary work (creative works such as novels and poems) discussed in the sample texts.

The list of coding categories includes aspects investigated in the quantitative analysis and expands beyond them to provide new perspectives. While the list appears comprehensive, some categories will not be examined in detail due to a lack of recurring or meaningful patterns. The exploration will focus on meaning distribution, main verbs used with the modals, parts where the modals are used, and textual voice expressed by the modals. Comparisons will be made between sample student groups and between disciplines.

5.3 Overview of the three modals used in the Chinese EFL students' sample texts

Before examining specific texts, this section provides an overview of how *must*, *have to*, and *should* are used in the learner and the reference sample in terms of frequency and meaning distribution. It also compares these observations with the quantitative findings.

This overview focuses on reporting the modal frequencies to provide comprehensive detail. However, it is important to note that no claims for generalisation should be made due to the small sample size and the sampling method. As stated in Section 5.2.1, the learner and the reference sample together comprise only 16 texts, with four in RS-Business written by the same student due to availability. Despite these limitations, the primary aim of the qualitative analysis is to provide a fine-grained view of the modal use in academic writing by examining individual texts. A detailed analysis of examples from the corpora and their usage by the students will be presented in Section 5.4.

5.3.1 Overall frequency of the three modals in the Chinese EFL students' sample texts

Table 5.5 shows the absolute and normalised frequency per 10,000 words of *must*, *have to*, and *should* in the learner and reference sample. The table also shows the percentage (%) of each modal relative to the total occurrences of all target modals.

Table 5.5 Frequency distribution of the three modals in the learner and the reference sample

Modals	Learner sample			Reference sample		
	AF	NF	%	AF	NF	%
Must	21	3.99	20.19	18	7.86	40.00
Have to	17	3.23	16.35	7	3.06	15.56
Should	66	12.54	63.46	20	8.73	44.44
Total	104	19.76	100.00	45	19.65	100.00

Figure 5.2 is a figurative presentation of the normalised frequency per 10,000 words so that the comparison between the two sample student groups is clearly demonstrated.

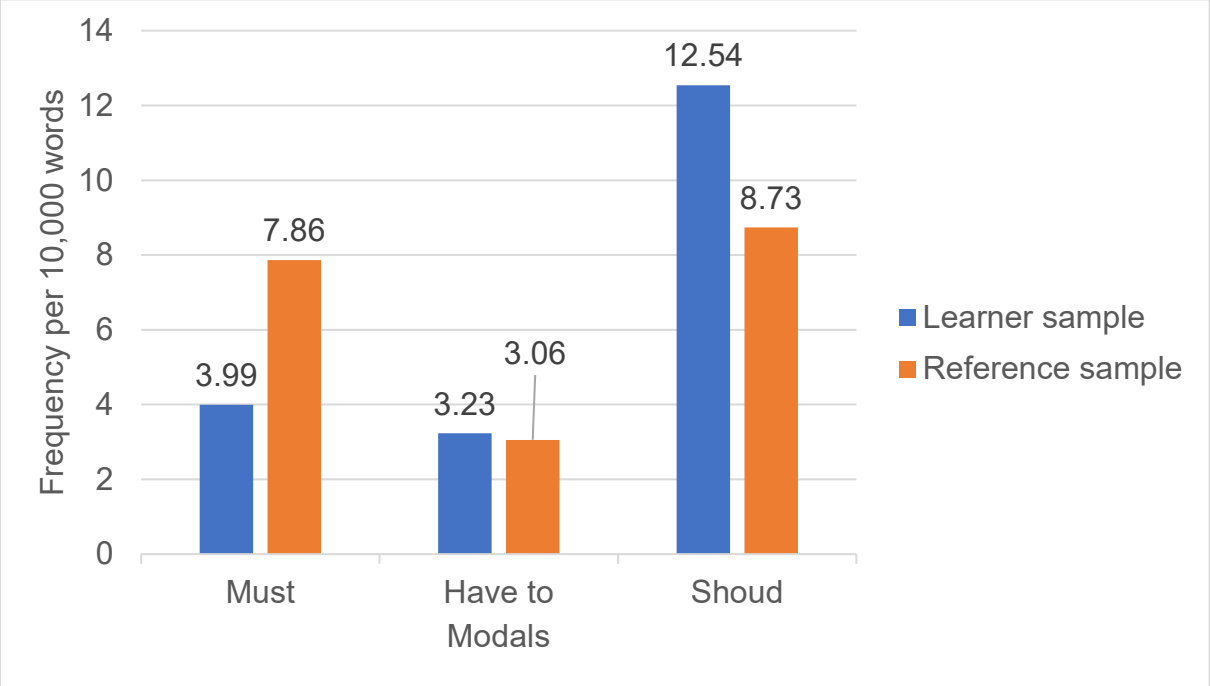


Figure 5.2 Normalised frequency per 10,000 words of the three modals in the learner and the reference sample

The results are mostly consistent with the quantitative findings (see Figure 4.2). *Must* is under-represented and *have to* and *should* are over-represented in the learner sample compared to the reference sample. However, the normalised frequency difference for *should* between the two sample student groups is less significant than that in the quantitative analysis. This difference between the quantitative and qualitative findings, as well as those to be addressed shortly, might result from the

sampling method used. Despite efforts to select sample texts that are as representative as possible, there remains the possibility that they may not entirely capture the trends observable in a broader dataset, as mentioned at the beginning of Section 5.3. Frequency distribution presented here is to provide an overview of the three modals, while the focus of the qualitative analysis is on the examination of the modal use within each text.

5.3.2 Overall meaning distribution of the three modals in the Chinese EFL students' sample texts

Table 5.6 below shows the absolute and the normalised frequency per 10,000 words of each meaning of the three modals in the learner and the reference sample.

Table 5.6 Meaning distribution of the three modals in the learner and the reference sample

Modals	Meanings	Learner sample		Reference sample	
		AF	NF	AF	NF
Must	Epistemic	2	0.38	1	0.44
	Root	19	3.61	17	7.42
Have to	Epistemic	0	0.00	0	0.00
	Root	17	3.23	7	3.06
Should	Epistemic	5	0.95	2	0.87
	Root	61	11.59	18	7.86
Total	Epistemic	7	1.33	3	1.31
	Root	97	18.43	42	18.34

As shown in the table, root sense of the three modals is predominantly used in the texts compared to their epistemic use. In the learner sample, there are only two instances of epistemic *must*, and in the reference sample, only one. Epistemic *should* shows a slightly higher absolute frequency, with five and two instances in the two samples respectively, while there is no instance of epistemic *have to*.

As for the normalised frequency per 10,000 words, epistemic uses of the three modals do not show marked differences between the learner and the reference sample. This corresponds with the quantitative findings (see Section 4.3.2), which also indicate no statistical differences in the epistemic usage of modals between the two student groups.

When examining the proportion of epistemic use of the three modals in the learner and the reference sample shown in Figure 5.3, a similar pattern can also be observed between the two student sample groups.

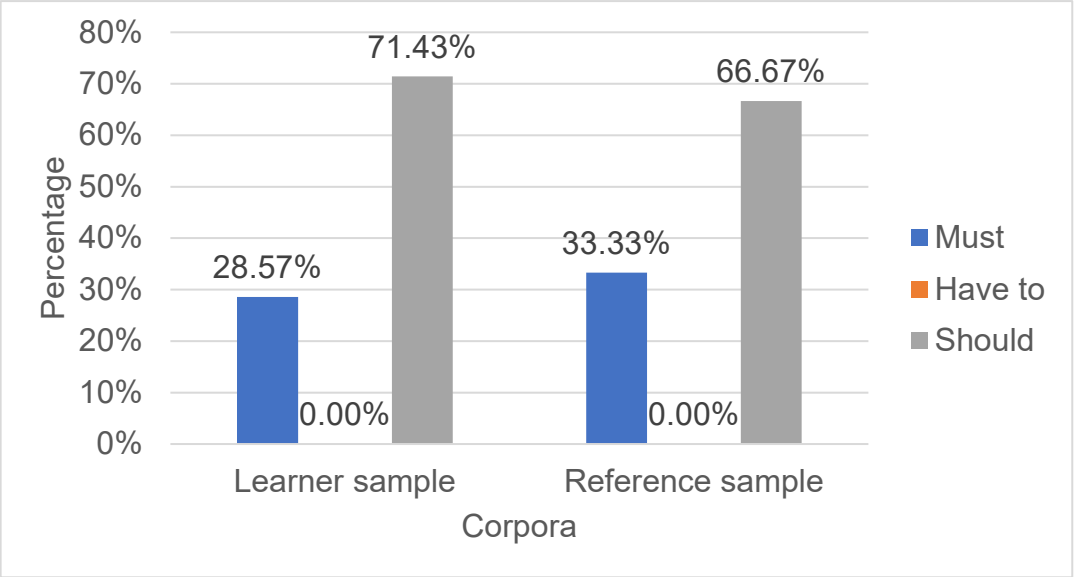


Figure 5.3 Proportion of the epistemic use of the three modals in the learner and the reference sample

As shown in the figure, *must* only represents around 30% of instances among the three modals to express epistemic meaning in the learner and the reference sample, whereas *should* accounts for approximately 69%. This contradicts the quantitative findings presented in Figure 4.3, where epistemic *must* holds a slightly higher proportion than epistemic *should*, with a difference of about 15%. As shown in Table 5.6, the absolute frequencies of the epistemic use of the three modals are extremely low, and thus each instance carries greater weight. This could disproportionately influence the overall analysis and lead to differences from the quantitative findings.

Regarding the root use, the differences in using the three modals between the two sample student groups align with the quantitative findings presented in Table 4.6. Chinese sample students use less than half the normalised frequency of root *must*

compared to their British counterparts, whereas their use of root *have to* and *should* is higher.

When examining the proportion of the root use of the three modals in the learner and the reference sample, as shown in Figure 5.4, a consistent pattern emerges. In both samples, root *should* constitutes the highest percentage, followed by root *must* and *have to*. This finding is mostly consistent with the results presented in the quantitative analysis (see Figure 4.4).

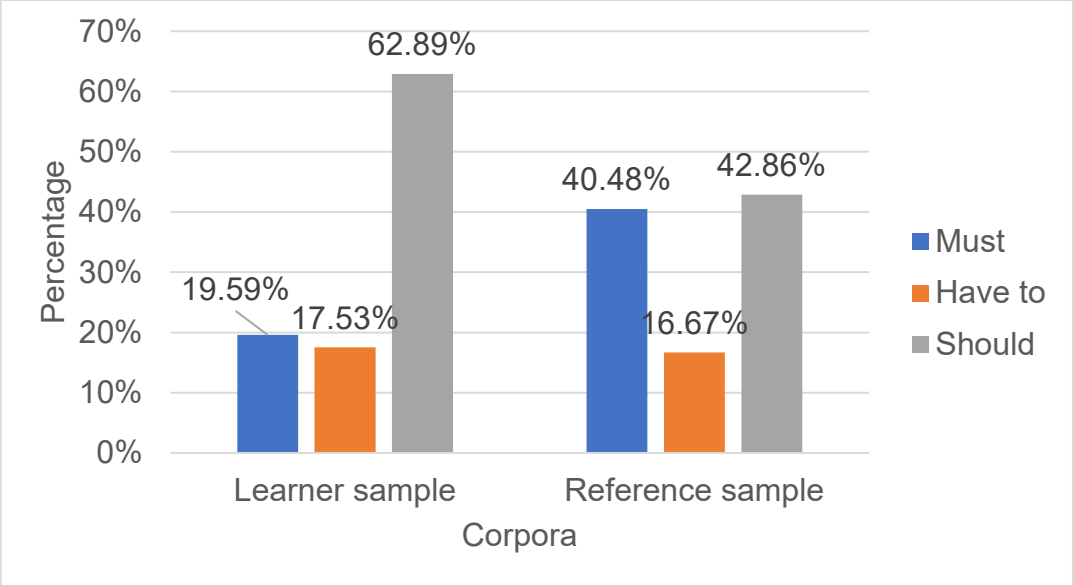


Figure 5.4 Proportion of the root use of the three modals in the learner and the reference sample

5.4 Profiles of the three modals across four sub-samples

Having analysed the overall pattern of *must*, *have to*, and *should* used in the learner and the reference sample, I will now proceed to examine the modal profiles across the four sub-samples in the following four sub-sections. The findings will be structured similarly, beginning with a discussion of the variations between sample student groups and followed by an examination of disciplinary differences.

The analysis will start by presenting the frequency distribution of the modals across the four sub-samples. As previously noted, the sampling method and size constrain the generalisability of the qualitative analysis. However, the primary focus of the qualitative analysis is to examine the use of modals within individual texts. This involves comparing their distribution across different parts of the texts and discussing their epistemic and root uses separately in terms of textual voice and other aspects (e.g., degree of subjectivity and category of directives) by analysing examples, which will be presented in the Sections 5.4.2, 5.4.3, and 5.4.4.

5.4.1 Frequency distribution of the three modals across four sub-samples

Table 5.7 below presents the absolute frequency of the three modals in each sub-sample, and Figure 5.5 graphically illustrates their normalised frequency per 10,000 words.

Table 5.7 Absolute frequency of the three modals in the four sub-samples

	LS-BM	RS-Business	LS-EL	RS-English
Must	11	6	10	12
Have to	6	3	11	4
Should	46	13	20	7

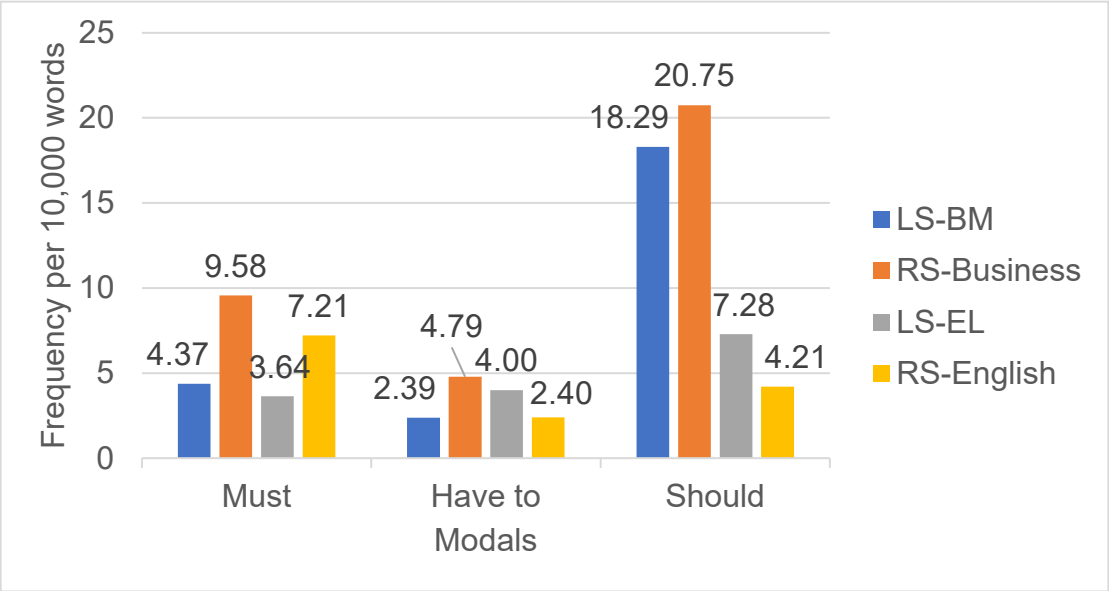


Figure 5.5 Normalised frequency per 10,000 words of the three modals in each sub-sample

As shown from the table and the figure, *have to* does not show marked differences between groups of sample students and between disciplines. *Must*, on the other hand, is used less frequently by Chinese sample students in both disciplines compared to their British counterparts. There is a higher normalised frequency of *must* in LS-BM and RS-Business than in LS-EL and RS-English. Regarding *should*, it is under-represented in LS-BM compared to RS-Business, while the opposite trend is seen between LS-EL and RS-English. A marked variation is observed between disciplines,

with *should* being over-represented in LS-BM and RS-Business compared to LS-EL and RS-English.

Sample texts are also examined individually, and Table 5.8 below presents the absolute frequency of the three modals in each text. To clearly differentiate the writing of the two sample student groups, the learner sample is highlighted with an orange background, while the reference sample is shown on a blue background.

Table 5.8 Absolute frequency of the three modals in each text

Sub-sample	Text	Must	Have to	Should	Total
LS-BM	UGBM 04307	3	3	6	12
	UGBM 05706	4	1	8	13
	UGBM 10205	1	0	18	19
	UGBM 10407	3	2	14	19
RS-Business	0202k	1	1	4	6
	0202l	1	0	2	3
	0202m	1	2	1	4
	0202n	3	0	6	9
LS-EL	L 00504	5	1	5	11
	L 02508	1	4	4	9
	L 04106	2	3	7	12
	L 10408	2	3	4	9
RS-English	0229b	1	2	4	7
	3007a	3	1	1	5
	3008f	4	1	0	5
	3110a	4	0	2	6

What is generally similar across the texts is the order of the absolute frequency of the

modals, with *should* being the most frequent, followed by *must* and *have to*. An exception is British sample students in RS-English, who slightly prefer *must* over *should* and *have to*. Another observation is that writers of UGBM 10205 and UGBM 10407 in LS-BM use *should* markedly more than other writers, which could result from individual differences in specific topics or structures of the texts.

5.4.2 Frequency and meaning distribution of the three modals across different parts of the texts

Beyond examining frequency distribution, it is also important to explore how the three modals are distributed in different parts of academic writing. This analysis can help to explain variations in the modal frequency and meaning distribution between student groups and between disciplines revealed through the quantitative and qualitative analyses.

As mentioned in Section 5.2.2, two approaches were used to examine the frequency distribution of the modals across different parts of the sample texts. First, an overview is provided using AntConc, followed by manually dividing the text parts and presenting the results in a table.

Figure 5.6 below is generated by AntConc using the concordance plot function, showing the relative position of the three modals in each text.

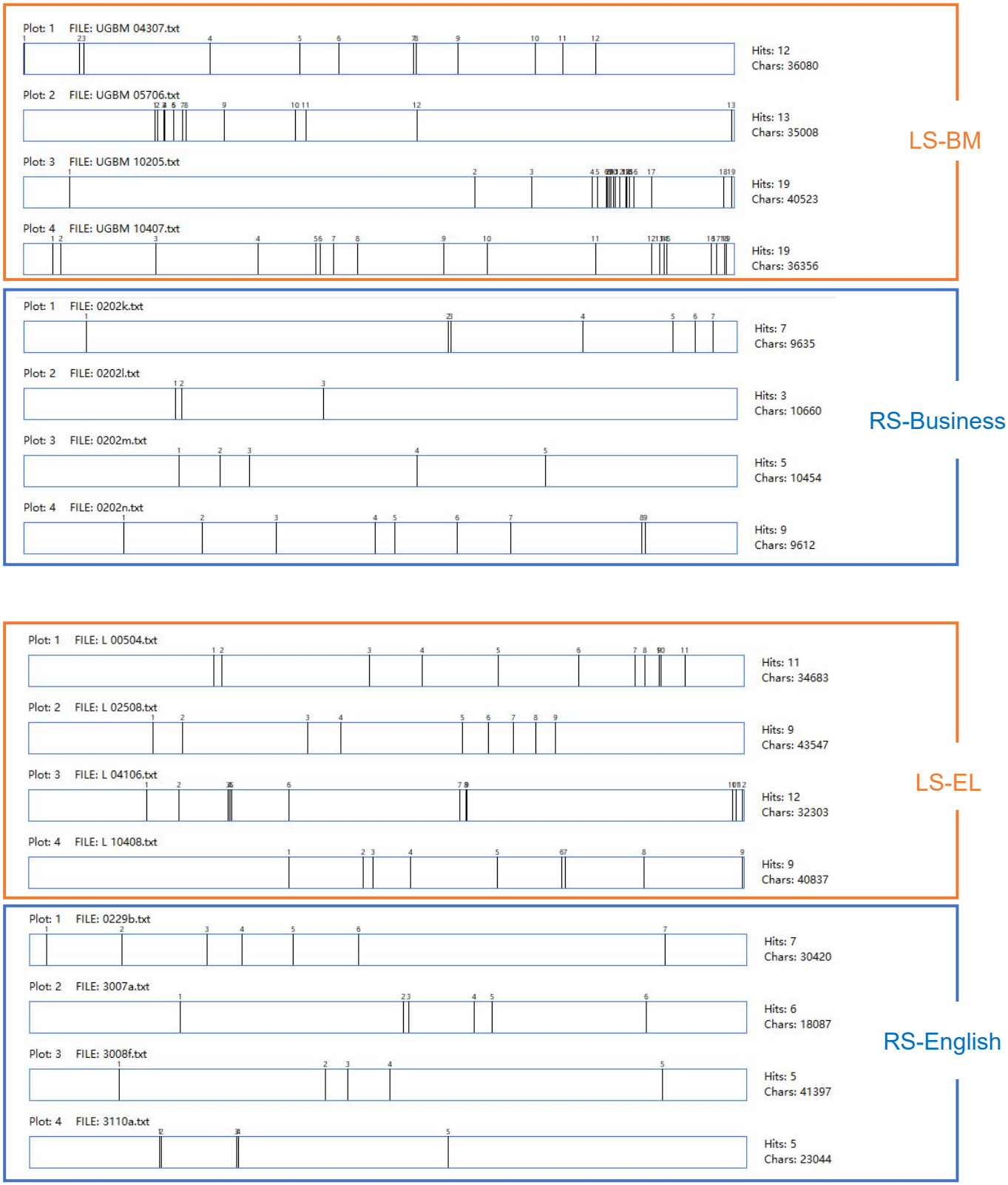


Figure 5.6 Distribution of the three modals in each text

There are four plots in the figure: the two within the orange rectangle are from the learner sample, and the two within the blue rectangle are from the reference sample. Captions to the right of each plot are added to indicate the names of the sub-samples. Vertical lines in the bars represent one of the three modals, and the length of each text is normalised to the width of the bar for comparison. The caption at the top of each bar indicates the file name, while that on the right presents the total absolute frequency of the three modals in each text.

In the learner sample, the three modals tend to cluster together in both sub-samples. For instance, in UGBM 10205 (LS-BM) and L 04106 (LS-EL), there is a noticeable increase in the use of the three modals towards the end of the texts compared to other parts. The modals also cluster at the start of UGBM 05706 (LS-BM). In contrast, the reference sample shows no clustering of modals in particular parts.

To investigate disciplinary variations, we compare the plots of LS-BM with LS-EL and RS-Business with RS-English. There is no marked difference between the two disciplines in the reference sample. However, Chinese sample students in LS-BM use the target modals more frequently in particular parts of a text compared to those in LS-EL.

This approach, while offering an overview of how modals appear in different parts of a text, has certain limitations, as discussed in Section 5.2.2. For example, the total absolute frequency is not necessarily accurate since the non-modal uses are not excluded when generating the figure. In addition, it is challenging to explain the reasons for the variations across sub-samples because the figure does not distinguish between the three modals or their meanings, and the division of text parts is not specific enough. To address these, a more detailed examination of the texts is undertaken, manually dividing the texts into introduction, literature review, analysis, and conclusion.

Tables 5.9 and 5.10 below display the absolute frequency and meanings of the three modals in each part of the text. Table 5.9 displays the findings for LS-BM and RS-Business, while Table 5.10 presents the results for LS-EL and RS-English. The first column lists the text parts, and the second column shows the modals analysed. The subsequent eight columns represent the individual texts in the sample. Cells highlighted in orange correspond to the learner sample texts, and those in blue represent the reference sample texts. Within the cells, numbers and abbreviations indicate the absolute frequency and the meaning of the modals: *e* refers to the epistemic meaning, and *r* represents the root sense. Hyphens denote the absence of a modal in that part. Hyphens are used instead of zeros to differentiate them from other numbers. *N/A* indicates that the literature review part is not included in the text.

Table 5.9 Distribution of different meanings of the three modals in each part of texts in LS-BM and RS-Business

		LS-BM				RS-Business			
		UGBM 04307	UGBM 05706	UGBM 10205	UGBM 10407	0202k	0202l	0202m	0202n
Introduction	must	-	-	-	-	-	-	-	-
	have to	-	-	-	-	-	-	-	-
	should	1r	-	-	2r	-	-	-	-
Literature review	must	N/A	-	1r	1r	N/A	N/A	N/A	N/A
	have to		-	-	1r				
	should		-	-	1e, 3r				
Analysis	must	3r	4r	-	1r	-	1r	1r	3r
	have to	3r	1r	-	-	-	-	2r	-
	should	5r	7r	15r	2r	-	2r	1r	6r
Conclusion	must	-	-	-	1r	1r	-	-	-
	have to	-	-	-	1r	1r	-	-	-
	should	-	1r	3r	6r	1e, 3r	-	-	-

Table 5.10 Distribution of different meanings of the three modals in each part of the texts in LS-EL and RS-English

		LS-EL				RS-English			
		L 00504	L 02508	L 04106	L 10408	0229b	3007a	3008f	3110a
Introduction	must	-	-	-	-	-	-	-	-
	have to	-	-	-	-	-	-	-	-
	should	-	1r	-	-	1r	-	-	-
Literature review	must	1r	N/A	1r	N/A	N/A	N/A	N/A	N/A
	have to	1r		-					
	should	-		-					
Analysis	must	2e, 2r	1r	1r	2r	1r	3r	1e, 3r	4r
	have to		4r	3r	2r	2r	1r	1r	-
	should	2e, 3r	1e, 2r	1e, 3r	4r	1e, 2r	1r	-	2r
Conclusion	must	-	-	-	-	-	-	-	-
	have to	-	-	-	1r	-	-	-	-
	should	-	-	3r	-	-	-	-	-

Let us first focus on the epistemic use of the three modals. As indicated in the tables, it is difficult to compare epistemic use of the modals in LS-BM and RS-Business because each sub-sample contains only one instance. Both are examples of epistemic *should*, used in the literature review part of UGBM 10407 (LS-BM) and the conclusion part of 0202k (RS-Business) respectively. By contrast, epistemic *must* and *should* are used similarly in LS-EL and RS-English, appearing exclusively in the analysis part.

This observation is also related to the disciplinary variation. Epistemic sense of the modals is used in the analysis part in LS-EL and RS-English but not in LS-BM and RS-Business, likely due to the differences in analytical approaches. Students in LS-EL and RS-English tend to focus more on examining characters' behaviours or plot settings, which could lead to a more frequent use of epistemic modals to make judgements about the truth of statements based on evidence in literary works or the author's background. The epistemic use of the modals will be fully discussed in the next sub-section.

Regarding the root use, Chinese sample students tend to use root sense of the modals in the introduction part, whereas British sample students rarely do so. Chinese sample

students appear to use the modals in this part primarily to explain the background and importance of their study, as illustrated in 5-1 to 5-3.

5-1 In the financial market, moral hazard has been changed into an important role, sometimes we *should* admit that can cause financial crisis. (Root_LS-BM_UGBM 10407)

5-2 Every successful business *should* be based on a strong culture. (Root_LS-BM_UGBM 04307)

5-3 The significance of this thesis lies in the following three aspects. First, compared with Poe's other works, his sea exploration tales does not gain as much attention as they *should* have, especially Pym. (Root_LS-EL_L 02508)

A complementary perspective to analyse examples of root use is Hyland's (2002) categories of directives, which have been briefly mentioned in Section 2.3.2. The classification of the three main categories is based on the types of actions the addressee is requested to perform: textual acts (which direct the addressee to another part of the text or to a different text), physical acts (which instruct the addressee to carry out research-related or real-world actions), and cognitive acts (which introduce new arguments, guide reasoning, or highlight specific points) (Hyland, 2002). Examples of the latter two categories will be presented and analysed in subsequent

sub-sections, but the first category is not present in the samples. A detailed account of directives and their sub-categories will not be addressed in this thesis, as they do not form the primary focus of the qualitative analysis. Nonetheless, this topic could serve as a valuable direction for future studies, details of which will be elaborated in Section 7.5.

The three examples above could all be classified as cognitive acts, as they involve the influence of the addressee's thought processes or beliefs. Specifically, in 5-1, the writer makes a suggestion rather than imposing a strong obligation. This is evidenced by the use of *sometimes*, indicating that admitting the issue is only recommended occasionally. The pronoun *we* not only involves the writer which shows a degree of subjectivity, but also engages the reader, suggesting a collective acknowledgement of the issue that moral hazard could lead to a financial crisis. Additionally, the verb *admit* following *should* highlights a cognitive action to recognise or accept an idea, which falls under the category of cognitive acts to influence the reader's understanding of a claim.

Root *should* in 5-2 is used to express a relatively strong suggestion, evidenced by its use with *every*, which implies that the suggestion applies to all businesses. Based on Halliday and Matthiessen's (2004) classification, this usage is subjective and implicit,

as it reflects the student's own suggestion but does not explicitly state this. It represents a cognitive act in that the verb *base* describes a foundational guideline for business practices, which is used at the beginning of this text to introduce the argument. The use of passive voice avoids mentioning who is responsible for implementing the suggestion, a feature that Hyland (2002) observes as prevalent in research reports by second language students.

5-3 is an example from LS-EL in which the student emphasises what gap the study fills compared to the previous studies. In this case, root *should* is used with the perfect aspect to convey the hypothetical sense, as explained in Section 2.2.5. This co-occurrence indicates that it is known that previous researchers did not place sufficient emphasis on Poe's sea exploration tales. Collins (2009) notes that this use also conveys a sense of criticism. The example implies a writer-imposed suggestion but does not explicitly state it. It could be classified as a cognitive act as it encourages the reader to re-evaluate their viewpoints about Poe's work.

In contrast to Chinese sample students, British sample students rarely use root sense of the target modals in the introduction part. The only instance is in 0229b (RS-English), where root *should* is used in a direct quotation of a critic's comment on the significance

of Jane Austen's work. This example as well as the further discussion on the textual voice expressed by the modals will be presented in Sections 5.4.3 and 5.4.4.

One explanation for this variation in the modal distribution in the introduction part lies in the slight genre differences between the learner and the reference sample, as discussed in Section 3.2.2. The learner sample consists of dissertations, while the reference sample includes essays. Due to the higher word limit of dissertations, Chinese students tend to start their work by introducing new arguments or highlighting specific viewpoints, establishing a foundation for following analyses. They then devote the rest of the text to elaborating on these initial ideas. However, this approach can appear overly assertive if not properly moderated with preliminary reasoning and explanation. By contrast, in the reference sample, the background or importance of the study is generally presented in fewer than three sentences because of the shorter text length. British sample students tend to focus on the aim and structure of the essay in this part and introduce their arguments at a later stage.

Another difference between the two sample student groups is the use of root modals in the analysis part. Both sample students use more root modals in this part compared to the other parts, but the distribution of the three modals differs. Chinese sample

students in both disciplines predominantly use root *should* in the analysis part, in contrast to their British counterparts. This is also observed in the quantitative analysis (see Table 4.6) that Chinese students prefer root *should* over root *must* and *have to* when giving suggestions. A relatively extreme case is UGBM 10205 (LS-BM), which contains 15 instances of root *should* in the analysis part. In this part, there are two consecutive paragraphs describing what is advised for charity marketing in a company, and eight root *should* are used. 5-4 is one of these paragraphs, containing root *should* in every sentence.

> 5-4 Both consumers and enterprises *should* keep a clear mind on charity marketing, both of its merits and shortcomings. To consumers, they *should* realize the importance and necessity of an organization's charitable activities and stop being immersed into skepticism. At the same time they *should* keep a clear mind to not to be misled my some deceptive charity marketing. To companies, they *should* understand the essence of charity marketing and make use of it in a right way. (Root_LS-BM_UGBM 10205)

These uses could be considered as suggestions rather than strong obligations for several reasons. First, root *should* is used, which typically conveys a weaker sense of obligation compared to root *must*. Additionally, the verbs used with root *should, keep*, *understand* and *realize,* express stative meaning, and the student does not have

authority over the subjects in the sentence to enforce the suggested actions. Lastly, the source of the suggestion, the writer, is not explicitly stated but rather implied. As for the directive categories, all uses seem to describe physical acts because they aim to influence real-world behaviours among consumers and companies. The dense usage of root *should* may be due to the student's lack of alternative expressions for suggestions, resulting in repetitive use of the same modal. This reason will be discussed further in Section 5.4.4, where additional examples are provided.

The over-representation of root *should* in LS-BM compared to RS-Business can also be observed in the conclusion part, as shown in Table 5.9. One reason for the over-representation of root *should* in the conclusion part in LS-BM could be the rubric of the essay or guidelines provided by lecturers. Although the titles of the learner sample texts are not recorded (see Section 3.2.1), an analysis of the texts reveals that the conclusion parts in LS-BM typically offer suggestions for future practice, as exemplified in 5-5.

5-5 To balance benefits between corporation and society is a practical problem. Firstly, stress openness and transparency in financial market. This principle can keep information symmetry and avoid information asymmetry. The investors will know about financial institutes and their products clearly. Secondly, it *should* be

sensitive and pay more attention to social responsibility when designing financial derivatives. (Root_LS-BM_UGBM 10407)

Root *should* in this example indicates a physical act that the entities designing financial derivatives are advised to perform. The suggestion is relatively weak in force, as the student lacks the authority to request the relevant party to undertake this action, and although it is implied that the student is the source of the suggestion, this is not explicitly stated. In addition, the verbs used with root *should, be* and *pay*, both express stative meanings related to cognitive activities. These suggestions are difficult to prove as being followed due to their abstract nature.

This reason related to rubrics could also explain the sole exception in the reference sample, 0202k (RS-Business), where the student uses the root modals frequently in the conclusion part. The third rubric question of this essay is 'What steps would you advise Paul Stone to take to improve quality performance at EHL? (40% of the marks)'. The prompt of *advise* and *improve*, along with the percentage of marks assigned, may encourage the students to give suggestions so that they could receive a higher mark, potentially leading to more use of root modals. One example is 5-6 shown below, which can be categorised as a physical act since it involves a suggestion for real-world practice. The suggestion is subjective and explicit, as 'I think' indicates that it is the

student's viewpoint. The use of passive voice in this example serves to avoid specifying who is responsible for implementing the suggestion, focusing instead on the action to be taken.

> 5-6 Finally I think the pay scheme *should* also be changed to a weekly or hourly wage, so everyone is paid the same amount regardless of output, since this would discourage working for quantity rather than quality. (Root_RS-Business_0202k)

As for the disciplinary variation, root *should* is used more frequently in the analysis and conclusion parts in LS-BM than in LS-EL. This difference could be attributed to disciplinary conventions. Writings in LS-BM generally place more emphasis on practical impact than those in LS-EL, reflecting the predominant features of applied disciplines as opposed to pure disciplines. Students in LS-BM tend to give suggestions on how to enhance performance or resolve issues for companies and brands, as demonstrated by the previously analysed examples 5-4 and 5-5.

The examination of the frequency of the modals in different parts provides a perspective for exploring the modal use in academic writing and offers potential explanations for variations in frequency and meaning distribution between sample

student groups and between disciplines. To gain a deeper understanding of the modal profiles in the samples, the following two sub-sections will discuss the epistemic and root uses of modals separately and analyse more examples.

5.4.3 Epistemic use of the three modals in the four sub-samples

This section will examine the epistemic use of *must*, *have to*, and *should* across the four sub-samples, focusing on their frequency distribution, co-occurrence with harmonic phrases, and textual voice expressed. Epistemic *have to* is not used in the learner and the reference sample, and thus is not mentioned in this sub-section.

Let us start with the discussion on the epistemic *must*. It only appears in LS-EL and RS-English, with two occurrences in the analysis part of L00504 (LS-EL) and one in the same part of 3008f (RS-English), as shown in Table 5.10. The two instances of epistemic *must* in LS-EL are shown in 5-7 and 5-8.

5-7 In the myth of the archetype often comes the dual system of good and evil, though they are contradictory. There *must* be the evil as opposed to the good. (Epistemic_LS-EL_L 00504)

5-8 It can be said that part of Harry's ability was given by Voldemort, or it is Voldemort who created Harry, so he *should* be destroyed by the newly born hero,

Harry. Just as in Oedipus and later in the creation myth of God, the god before *must* be replaced by the new god. (Epistemic_LS-EL_L 00504)

In 5-7, the combination of ‘there must be’ and a noun, as discussed in Section 4.4.4.1, is strongly associated with the epistemic use of *must*. This combination implies a strong sense of certainty about the relation between the evil and the good based on the evidence of the myth provided in the previous sentence. In the case of 5-8, epistemic *must* express a relatively strong sense of certainty that the previous god is expected to be replaced by a new god in the myth. This supports the earlier proposition that Harry is likely to destroy Voldemort through the use of epistemic *should*, which is also evidenced by the narratives in the literary work.

Epistemic *must* in RS-English, on the other hand, is used in a direct quotation, as shown in 5-9. The quotation, written by the author of the literary work, is used by the student to support the statement that ‘social evolution lies in centring society on female ethics’.

5-9 Images of decay illustrate the degeneration of patriarchy and strengthen the text’s conclusive nuance that social evolution lies in centring society on female

ethics; ‘The angel and apostle of the coming revelation *must* be a woman.’ (Epistemic_RS-English_3008f)

The use of epistemic *must* in LS-EL and RS-English but not in LS-BM and RS-Business could be explained by the differences in analytical approaches. Students in LS-EL and RS-English tend to make confident judgements regarding the truth of a statement because they can find relatively strong and reliable evidence in literary works, whereas students in LS-BM and RS-Business are less inclined to make such judgements.

In regard to epistemic *should*, its usage is consistent across the four sub-samples, being used with harmonic words or phrases, a combination described as ‘modal harmony’ by Huddleston and Pullum (2002, p. 179), as discussed in Section 2.2.3. Examples from different sub-samples are presented in 5-10, 5-11, and 5-12.

5-10 As expected, one particular pool of loans *should* be worth what the loans supposed to be worth. (Epistemic_LS-BM_UGBM 10407)

5-11 What *should* happen is that when effort into prevention increases, this will reduce internal failure costs, which reduces external failure costs which then eventually reduces appraisal costs. (Epistemic_RS-Business_0202k)

5-12 The Magic Stone can achieve the desire of eternal life of the people with its particular functions. Dumbledore *should* also know this capacity, but perhaps he will not use it. (Epistemic_LS-EL_L 00504)

In 5-10, ‘as expected’ is consistent with the epistemic meaning of *should*, demonstrating the student’s assessment of the loan’s value. In the case of 5-11, epistemic *should* is used in a noun clause functioning as the subject. ‘What should happen’ could be paraphrased as ‘what is likely to happen’. The student evaluates the likelihood of the truth of the following proposition, where the predictive use of *will* is compatible with epistemic *should*. Dumbledore in 5-12 is a character in the novel, and the student makes assumptions regarding Dumbledore’s knowledge of the power of the Magic Stone. The use of *will* in the following coordinate clause facilitates the epistemic interpretation of *should*. These three examples all demonstrate subjective uses of epistemic *should*, as they imply the writer's involvement in the judgement. However, this involvement is not explicitly stated.

Both epistemic *must* and *should* show variations across the sub-samples in the textual voice they express. They appear in direct quotations only in RS-English but not in the other three sub-samples. There are two instances of epistemic use of the modals in RS-English, both of which appear in direct quotations. One instance involves epistemic

must, which has been previously discussed (see 5-9). The other is an instance of epistemic *should*, as illustrated in 5-13.

> 5-9 Images of decay illustrate the degeneration of patriarchy and strengthen the text's conclusive nuance that social evolution lies in centring society on female ethics; 'The angel and apostle of the coming revelation *must* be a woman.' (Epistemic_RS-English_3008f)
>
> 5-13 At the end of Austen's novel, there is the sense that the union of Mr Knightley and Emma will be a marriage of true minds. The heroine herself remarks to her fiancé that 'I can hardly imagine that anything which pleases or amuses you, *should* not please and amuse me too' (Austen 492, ch. 54). (Epistemic_RS-English_0229b)

In 5-13, the student uses the quotation of Jane Austen, the author of the literary work, to prove their proposition in the previous sentence, that Mr Knightley and Emma's marriage will be a union characterised by true minds. Writers of 5-9 and 5-13 both use the quotations of the author containing the epistemic modals to support their own statements.

The absence of quotations with the epistemic modals in LS-EL may be due to different analytical approaches used in LS-EL and RS-English, which will be discussed further

in the next sub-section with examples of the root use of the modals. The disciplinary variation of not using epistemic sense of the target modals in direct quotations in LS-BM and RS-Business could be explained by the fact that students in RS-English tend to use quotations from the literary work to support their analyses, an approach that is not as commonly used by students in LS-BM and RS-Business.

5.4.4 Root use of the three modals in the four sub-samples

Let us move on to discuss the case for root sense of the three modals in the four sub-samples. As mentioned in Section 5.4.2, root *should* is used most frequently in the learner sample among the three modals compared to the reference sample. There is an extreme instance where it appears 15 times in the analysis part of UGBM 10205 (LS-BM). The excessive use of root *should* in the learner sample may result from a lack of alternative devices to give advice. British sample students, on the other hand, use a broader range of expressions for their suggestions, as shown in 5-14 to 5-16. The writer of 0202k (RS-Business) uses ‘I would recommend’, ‘it is important to’, and ‘another solution’ to imply the suggestions and prevent repeated use of the same root modals.

5-14 I would recommend a move to a Total Quality Management approach to quality to improve quality performance at EHL. (Root_RS-Business_0202k)

5-15 To ensure that external customers are happy, it is important to push the idea that every part of the firm and therefore every internal supplier contributes to this by ensuring the satisfaction of internal customers. (Root_RS-Business_0202k)

5-16 Another solution, which could be combined with TQM, is Kaizen, which essentially means continuous improvement. (Root_RS-Business_0202k)

These expressions could be positioned at different points on the scale of subjectivity. For instance, 'I would' in 5-14 explicitly indicates that the suggestion is given by the student, making it relatively more subjective compared to 5-15, which contains an adjectival predicate with *important*, a description used by Hyland (2002). He also characterises this expression as impersonal to mitigate the risk of obligating the reader to perform an action.

Moreover, a closer analysis reveals that Chinese sample students in LS-BM sometimes use the same main verb with root *should*. One example is UGBM 10205 (LS-BM), where the student uses *include* with root *should* three times in the analysis part, as illustrated in 5-17 to 5-19. These instances appear in two adjacent paragraphs. The student repeatedly stresses the importance of planning in charity marketing, but they almost express the same meaning.

5-17 A successful charity marketing campaign *should* include the whole marketing planning process: define the mission of the organization, determine organizational objectives, […]. (Root_LS-BM_UGBM 10205)

5-18 To some companies, doing charitable activities just belong to the fifth step—implement strategy through operating plans. This is obvious incorrect. Charity marketing *should* include all the marketing mixes and should be implemented step by step. (Root_LS-BM_UGBM 10205)

5-19 Successful charity marketing doesn't merely mean huge donations or just charitable activities; it *should* include a whole process of marketing planning. (Root_LS-BM_UGBM 10205)

The verb used with the root *should* in these instances, *include*, conveys the relationship between entities, as described by Biber et al. (1999). These examples seem to convey suggestions rather than strong obligations indicated by the co-textual features. The use of *include* to express a stative meaning, the inanimate subjects 'charity marketing', and the absence of explicitly stated writer involvement all contribute to weakening the strength of force behind implementing the actions. In addition, these suggestions could be classified as cognitive acts as they prompt reconsideration and strategic thinking about charity marketing. For instance, in 5-18, the student first states what is incorrect and then proceeds to introduce a new argument concerning recommended steps for

charity marketing. The same strategy applies to example 5-19 as well, in which the student first outlines what does not constitute effective charity marketing and then introduces a broader, more strategic approach involving comprehensive marketing planning. The writer effectively challenges existing perceptions, prompting the reader to reassess and potentially revise their views based on the new insights offered.

The use of the same main verb with root *should* can also be observed in UGBM 10407 (LS-BM), as shown in 5-20 and 5-21 below. *Avoid* is used twice with root *should* in the same paragraph in the conclusion part to give similar suggestions for preventing moral hazard.

> 5-20 The greedy plans, such as creating various financial derivatives, which may exacerbate moral hazard will lead to even more excessive risk-taking and *should* be avoided. (Root_LS-BM_UGBM 10407)
>
> 5-21 If I take a risk, then we want to ensure that I be made to bear it. But if I take a risk at your expense, then that is moral hazard -- a bad thing *should* be avoided. (Root_LS-BM_UGBM 10407)

Both examples use root *should* with passive voice, which allows the student to underscore the importance of avoiding a point while not specifying who is responsible

for undertaking it. This co-textual feature tends to weaken the strength of obligation, indicating that these examples express suggestions rather than strong obligations. Another indicator is that the student does not have the authority over the party involved to perform the suggested action. Additionally, it seems that 5-20 is relatively less subjective than 5-21 in that 5-20 presents a general principle about risky financial behaviours without explicitly referencing the viewpoint of the writer, but 5-21 uses pronouns such as *I* and *we*, which directly involve the writer's perspective in the assessment of risk. Both examples could fall into the cognitive acts category in that they present new arguments regarding moral hazard.

The repeated use of the same verb with root *should*, as observed in the previous two sets of examples, potentially indicates that Chinese sample students in LS-BM try to emphasise a particular argument but seem to face challenges in elaborating on the suggested action and clearly explaining its reasoning.

Another difference between the two sample student groups is related to textual voice. In the case of root modals, textual voice is sometimes discussed in relation to the source of obligation, referring to the subjective and objective distinction. As mentioned in Section 2.3.1, the source of obligation concerns who imposes the obligation. It is

generally considered subjective when the source is the speaker, and more objective when the source is external forces, such as rules and laws. However, Collins (2009) questions the consistency of this distinction in application to all root uses in that some instances do not specify the source. Although the textual voice and the source of obligation overlap to some extent, they seem to have different focuses. The source of obligation is related to the foundation of the writer's proposition (e.g., personal beliefs or objective evidence), whereas textual voice highlights how the writers engage with previous literature or other third parties. The latter deals more specifically with features of academic writing. Thus, in the following paragraphs, I will compare the root modal use, focusing mainly on the perspective of textual voice.

Table 5.11 below presents the absolute frequency of the root use of the three modals in each (sub-)type of textual voice in the four sub-samples.

Table 5.11 Absolute frequency of the root use of the three modals categorised by textual voice in the four sub-samples

	Averral	Attribution		Total
		Direct quotation	Indirect quotation	
LS-BM	57	4	1	62
RS-Business	19	2	0	21
LS-EL	29	2	4	35
RS-English	9	9	3	21

The root sense of the modals in direct quotations is used similarly by students in LS-BM and RS-Business. The instances convey what previous researchers recommend and express the writer's viewpoints on them, either implicitly or explicitly, as exemplified in 5-22 and 5-23 below.

5-22 Almost all the governments are regarded as hero in the financial disaster because they issue bailouts. To quote just one writer out of many others with a same opinion, "the pendulum will swing -- and *should* swing -- towards an enhanced role for government in saving the market system from its excesses and inadequacies" (Summers 2008). Free markets have been tried and failed; so it needs more regulation and more active macroeconomic management. (Root_LS-BM_UGBM 10407)

5-23 Furthermore, 'To engage in strategic planning, an organisation *must* be able to predict the course of its environment, to control it, or simply to assume

its stability...otherwise, it makes no sense to set the inflexible course of action that constitutes a strategic plan'. (1998 2001:67). Mintzberg calls this the 'fallacy of predetermination'. Predicting discontinuities like technological breakthroughs and economic change is difficult. Indeed, in the context of the above quote, it makes no sense to demand discipline in a strict sense if your environment is not stable, since a firm would be unable to react to change. (Root_RS-Business_0202n)

In 5-22, the writer presents an argument about the significant role governments play in financial crises by quoting a researcher who shares this view. This direct quotation used subtly indicates that the writer also endorses this view, though it is not explicitly stated. The co-occurrence of *will* implies an inevitable trend towards an enhanced role of the government, which seems to strengthen the force of this suggestion. Following the quotation, the writer expands on this argument, emphasising the inadequacies of free markets and advocating for more government involvement and regulation.

The writer of example 5-23 uses a direct quotation to illustrate the concept of the ‘fallacy of predetermination’. Root *must* is used with ‘be able to’ to specify the capabilities an organisation is obliged to possess. The strength of obligation is relatively weak, as indicated by the inanimate subject ‘an organisation’, the main verb *be*, and the lack of authority held by the antecedent author over the organisation, co-

textual features which Coates (1983) discusses in relation to the strength of obligation. In the last sentence of the example, the writer revisits the quotation to further clarify the suggestion, using an *if* clause and the pronoun *your* to enhance engagement with the reader.

Both examples appear to express cognitive acts, though they are not directly conveyed by the writers but rather through direct quotations to implicitly present the writers' viewpoints. This method of introducing new arguments via established research might prove more persuasive than presenting them directly, as it leverages the credibility of antecedent authors to support the claims.

As mentioned in Section 5.4.3, epistemic sense of the modals is used in direct quotations in RS-English but not in LS-EL. This pattern is also evident in the root use of the modals, with a more pronounced difference in frequency. British sample students in RS-English use over 40% of all instances of the root modals in direct quotations, compared to only 6% in LS-EL. This difference can be attributed to analytical approaches. Chinese sample students in LS-EL mostly take a top-down approach, grounding their analysis in theories or background information and focusing on examining how characters develop and the impact of literary works. The only two

instances used in direct quotations in LS-EL are in the same text, L 04106 (LS-EL), where both examples include quotations from antecedent authors, as illustrated in 5-24 and 5-25.

> 5-24 Some scholars in China have already done some research on the topic of characteristics of western Children's literature and the phenomenon of widespread of Harry Potter series. Zhang Ying and Kong Dan, who teach in College of Foreign languages, Northeast Normal University, Changchun, China, commented in Harry Potter and Characteristics of Children's Literature in New Period, "[…] Different ages produce different literary works; therefore writers of children's literature *must* update their ideas and advance with the age. […]." (Root_LS-EL_L 04106)

> 5-25 As Shelley E. Taylor, Letitia Anne Peplau and David O. Sears wrote in the book Social Psychology, "[…] The distinguishing feature of compliance is that we are responding to a request from another individual or group. In some social situations, we perceive one person or group as having the legitimate authority to influence our behavior. In these cases, social norms permit legitimate authority to make requests and dictate that subordinates *should* obey them." (Root_LS-EL_L 04106)

The student takes the full paragraph from the previous researchers' work to describe their suggestion on how to be the writers of children's literature in 5-24. Root *must* is

used with several co-textual features identified by Coates (1983) as indicators of core use of root *must*, thus conveying a relatively strong sense of obligation. These features include an animate subject ('writers of children's literature'), activity verbs (*update* and *advance*), and the antecedent author's interest in obliging the subject to act. Additionally, this obligation could be categorised as a cognitive act, as it introduces a new argument encouraging authors of children's literature to integrate contemporary ideas into their work, with reasoning provided beforehand. The writer's attitude toward the obligation proposed by the antecedent author seems to be neutral, as indicated by the reporting verb *comment*. The quotation serves merely to illustrate research topics related to the Harry Potter series phenomenon, rather than directly impacting the analytical framework of the research. In contrast, a subsequent quotation (not presented in 5-24) forms the foundation for further analysis.

Similarly, in 5-25, the writer's attitude toward the obligation imposed by the antecedent author is also neutral, as indicated by the reporting verb *write*. In the quotation, the antecedent author appears to convey an objective suggestion because the source of the suggestion is the social norm, which is a collective and external standard. The suggestion is relatively strong, indicated by co-textual features such as the animate subject *subordinates* and the activity verb *obey*. However, this may not be considered

as an obligation due to the use of root *should* rather than *must*, and the social norm, while powerful, lacks the enforceability of law. This quotation introduces the concept of compliance as a factor in the success of Harry Potter films, yet the writer does not provide additional commentary or analysis, leaving the quotation somewhat isolated. This pattern is also observed in 5-24, and both examples discussed above are placed at the end of paragraphs, leaving no space for further discussion on the views of the antecedent authors.

British sample students in RS-English also use root modals in direct quotations of antecedent authors, but they seem to integrate these quotations more closely into their arguments than their Chinese counterparts in LS-EL, as exemplified in 5-26.

> 5-26 Indeed, existing criticism always qualifies any statements made to liken them. One critic has even moved from remarking on a resemblance, albeit relatively slight, in saying that, ‘Jane Austen’s novels could, indeed, be called educational novels though they bear little resemblance to the ‘Bildungsromane’ of Goethe’ (Klieneberger 33), to reminding the reader that, ‘It is significant that Jane Austen whose work marks the transition from the eighteenth-century novel of manners to the social realism of the nineteenth-century, *should* have started her career [....] by satirizing Werther and the novel of sensibility’ (Klieneberger 15). (Root_RS-English_0229b)

In 5-26, instead of using the direct quotation in isolation, the British sample student uses two quotations from the same critic and integrates them in one sentence, with suitable use of ellipsis to omit unnecessary words. Root *should* is used with the perfect aspect to describe a suggestion in the past and implies that the suggested action is not taken by Jane Austen, reflecting a hypothetical use of root *should* noted by Coates (1983), as discussed in Section 2.2.5. The writer uses these two direct quotations to support the previous statement that critics generally discuss the similarities rather than differences between the works of Jane Austen and Johann Wolfgang. It appears that British sample students may employ a more effective strategy by using direct quotations with root modals to support their statements compared to Chinese sample students. However, since this is the only example found in RS-English, generalising this observation should be approached with caution. Further evidence is needed to substantiate this finding more robustly.

Apart from using quotations from critics, British sample students in RS-English also take direct quotations from the literary work, mostly words of the characters, and then add comments to develop their arguments using a close reading approach. This practice is not observed in LS-EL, and the potential reason will be discussed in Section

6.3.2. One such example in RS-English is 5-27 shown below, in which the writer quotes a character's words as evidence of the author's viewpoint on the importance of having a career as a woman. Root *must* expresses a relatively strong and subjective sense of obligation in that the obligation is imposed by the character herself, and the strong intention is further emphasised by 'want to' in the following sentence. Directly quoting the character's words in the literary work could be an effective strategy to reveal the author's viewpoint for analysis.

> 5-27 Although written prior to the sexual revolution, the text demonstrates an awareness of the debate regarding the proper place of women. They were beginning to seek social change, increased freedom and economic security by forging their own careers. Gilman's depiction of the power and pride felt by the narrator in having 'work' could be indicative of her support for this as; 'I *must* get to work. I want to astonish him,' demonstrates her need to prove herself, worth and value. (Root_RS-English_3008f)

As for indirect quotations, students in LS-EL and RS-English use them similarly to express the suggestions or obligations given by the characters in the literary work, as shown below in 5-28 and 5-29.

5-28 The last clue is when Pym, Augustus and Dirk Peters plane to retrieve control of Grampus from the mutinous crew, Pym plays the role as real rescuer rather than Augustus and Peters. There are three clues to this point. […] Secondly, when Peters proposes that he *should* go up on deck and throw the watch to the sea; after that, they can make a rush together and secure the companion-way before any opposition could be offered; it is Pym who objects the proposal and points out its weaknesses. (Root_LS-EL_L 02508)

5-29 Faustus is strongly defacing the head of the Catholic church here, firstly by snatching from him, then by telling him that crossing is a 'trick' and *should*n't be done, and thirdly by hitting him. This short scene is a somewhat comic spectacle of mocking, considering the climate at the time toward the Catholics and how they were discarded, and mocked, this is a good example in blank verse how the common people of the country may have viewed the Catholic church and how the rise in Protestantism had no place for it. (Root_RS-English_3110a)

In 5-28, root *should* expresses a suggestion made by Peter, a character in the literary work, to 'retrieve control of Grampus', which Pym later objects to. This indirect quotation is used to support the previous statement that 'Pym plays the role as real rescuer'. Root *should* in 5-29 conveys what is suggested not to do, and the description of this scene serves to illustrate people's perception of the Catholic church in that specific historical setting in which the literary work was written. In both examples, verbs such as *propose* and *tell* are used to describe how the characters give suggestions,

but these do not reflect the writer's viewpoints. It appears that the use of indirect quotations primarily serves to describe the plots within the literary works as evidence to support arguments.

In regard to LS-BM and RS-Business, the latter sample group does not use target root modals in indirect quotations, as shown in Table 5.11. The sole instance in LS-BM, 5-30, is problematic in terms of how the quotation is paraphrased.

> 5-30 As the famous cultural theorists Robert A. Cooke defined, Passive/ Defensive Cultures refer to a culture where members believe they *must* interact with people in ways that will not threaten their own security. (Root_LS-BM_ UGBM 04307)

The instance of 5-30 in LS-BM is in the analysis part and at the start of a paragraph. The student paraphrases a definition written by a theorist, which forms the basis for the subsequent discussion on the cultural causes of counterproductive behaviour. The reporting verb used is *define*, which is a neutral one without showing the writer's attitude toward this definition. The following co-text elaborates on the definition, implicitly indicating that the writer agrees with this definition. To confirm if root *must* was added by the student or taken directly from the literature, I searched for the article

and the original sentence is ‘Passive/Defensive cultures, characterized by Approval, Conventional, Dependent, and Avoidance norms, encourage or implicitly require members to interact with people in ways that will not threaten their own personal security.’ (Cooke & Szumal, 2000, p. 148). This indicates that the root *must* is added by the writer. When comparing 5-30 with the original sentence, the latter part of the sentence remains almost identical. The differences appear in the front part, where the original definition is altered. The original sentence emphasises that the action is encouraged implicitly, whereas 5-30 highlights that the action is obligated to do. The paragraphed sentence shifts the original sentence from a mild encouragement to a strong obligation, which is indicated by the use of root *must* with an agentive subject *they* and an activity verb *interact*. Despite being a single example, it indicates that Chinese students may add root modals in indirect quotations, potentially altering the meaning of the original sentence.

Regarding the disciplinary variations of root use of the modals, as mentioned before, students in RS-English tend to use the root modals in the direct and indirect quotations of the characters in the literary work to conduct close reading, whereas students in RS-Business seem to use quotations to describe the suggestions they (dis)agree on. This difference is related to how the analysis is conducted in different disciplines.

In addition to the variation in textual voice, verb collocates of root use of the three modals in LS-BM seem to be more specific to the disciplinary knowledge (e.g., *sell*, *charge*, and *publicise*) compared to those in LS-EL, as exemplified in 5-31. This difference is also observed in the quantitative analysis presented in Section 4.6.4.2.

> 5-31 The company should take into considerations like what price *should* be charged, sensitiveness to price changes of the target segment, if the consumers use price as a cue to value or a cue to quality in the specific industry, and the discounting and premium charging. (Root_LS-BM_UGBM 05706)

In 5-31, root *should* expresses a suggestion for the company, and its strength is relatively weak, as indicated by co-textual features such as the inanimate subject *price*, the use of passive voice to avoid specific mentioning of who is responsible to act, and the lack of authority by the student over the company. This suggestion could be considered as a physical act as it involves practical applications in the real world. The verb *charge* used with root *should* is closely related to disciplinary knowledge of Business, describing the importance of appropriate pricing strategy.

Another noticeable disciplinary difference is that Chinese sample students in LS-EL

more frequently use the pronoun *we* with root use of the modals to encourage the reader to accept the suggestion or obligation, whereas this combination appears only once in LS-BM. This pattern is also reflected in the quantitative findings presented in Sections 4.4.4.2 and 4.6.4.2. Examples in LS-EL and LS-BM are illustrated in 5-32 and 5-33 respectively.

5-32 With regard to its success, England has undergone an unprecedented economic growth and cultural development turning into the leading capitalist country of the rest. Nevertheless, this is only one side of the ruling, we *should*, on the other side, also be aware of the existing serious social conflicts. From that perspective, the Elizabeth is definitely not a flourishing age, as it seemed. (Root_LS-EL_L 10408)

5-33 In the financial market, moral hazard has been changed into an important role, sometimes we *should* admit that can cause financial crisis. (Root_LS-BM_UGBM 10407)

Both examples appear to fit into the cognitive act category as they direct the reader's attention to the importance of particular points, such as another aspect of the Elizabethan age and the impact of moral hazard. This is highlighted by the use of ‘be aware of’ and *admit* with root *should* as both of them describe cognitive activities. Root *should* seems to convey a suggestion rather than an obligation, implied by the use of

stative verbs and the lack of authority by the students over the reader to enforce the action. The frequent use of *we* with root *should* by students in LS-EL might aim to create a sense of community with the reader, making the suggestion less authoritative and more like a shared conclusion. This interactive tone to encourage the readers to agree with the students' viewpoints might be more accepted in disciplines like EL than in BM, which will be further discussed in Section 6.4.2.

5.5 Summary

This chapter has presented the qualitative findings to explore *must*, *have to*, and *should* in a finer-grained view compared to the previous quantitative analysis, revealing distinctive features of the modals in academic writing such as the modal distribution in different parts of a text and the textual voice expressed by the modals.

There are variations between the two sample student groups and between the disciplines, and most findings are consistent with the quantitative results. In general, normalised frequencies of epistemic use of the three modals do not show marked differences between the student groups. However, their root use differs, especially root *must*, which is under-represented in the learner sample.

In terms of distribution in different parts of the texts, Chinese sample students tend to use root sense of the modals densely in specific parts, whereas British sample students use them more evenly. For example, Chinese sample students over-represent root *should* in the analysis part compared to their British counterparts, possibly because they lack alternative devices to give suggestions. Root *should* is also over-represented in the conclusion part in LS-BM compared to RS-Business, and most instances could be categorised as physical acts according to Hyland's (2002) framework. This could result from the influence of rubrics or guidelines provided by lecturers. As for the disciplinary variation, epistemic sense of the modals is used in the analysis part in LS-EL and RS-English but not in LS-BM and RS-Business, likely due to different analytical approaches. In addition, root *should* is used more frequently in the analysis and conclusion parts in LS-BM than in LS-EL.

The two senses of the modals are further discussed separately. Epistemic *should* is used similarly with harmonic words or phrases across the four sub-samples. However, the sub-samples differ in that epistemic *must* and *should* are used in direct quotations in RS-English but not in the other three sub-samples.

In regard to the root use of the modals, it is found that not only do Chinese sample

students predominantly use root *should* among the three modals compared to their British counterparts, but they also sometimes use the same main verbs repeatedly with the modals in LS-BM, indicating their struggle to explain their suggestions clearly and flexibly. Additionally, root use of the modals is more frequently used in direct quotations in RS-English than in LS-EL, and students in RS-English show better integration of these quotations with their arguments effectively. What is worth noting is one instance where a student in LS-BM adds root *must* when paraphrasing a definition but alters its original meaning. In regard to disciplinary variations in the root use, students in RS-English and RS-Business use the modals in quotations for different purposes. In addition, compared to writings in LS-BM, those in LS-EL demonstrate a more interactive tone, as epistemic use of the modals frequently co-occurs with the pronoun *we*.

The previous and the current chapters have presented the quantitative and qualitative findings respectively. The next chapter will summarise these findings and discuss potential explanations for variations in the modal use in Chinese EFL students' academic writing.

6 DISCUSSION

6.1 Introduction

The previous two chapters have presented the quantitative and qualitative findings of the profiles of *must*, *have to*, and *should* in the learner and the reference corpus. This chapter will summarise the findings relevant to the research questions and provide potential explanations regarding the variations in the frequency, meaning distribution, and semantics of verb collocates of the modals 1) between student groups and 2) between disciplines. To recap, the research questions stated in Section 1.3 are:

RQ 1: How frequently are *must*, *have to*, and *should* used in the Chinese EFL learner corpus and the reference corpus?

RQ 2: How are the meanings of the three modals distributed?

RQ 3: What semantic patterns can be identified regarding the main verbs that collocate with the three modals?

RQ 4: Do the profiles of the three modals differ between the two student groups, Chinese and British students?

RQ 5: Do the profiles of the three modals differ between the disciplines?

The rest of the chapter will be organised as follows. Sections 6.2 to 6.4 answer the first

three research questions respectively and provide discussions. The last two research questions, concerning the variation between the student groups and between the disciplines, are explored simultaneously with the discussion of the first three questions. The final section summarises this chapter.

6.2 Frequency distribution of the three modals across sub-corpora

The overall frequency of *must*, *have to*, and *should* in the learner and the reference corpus gives us a starting point for looking at the profiles of the three modals, from which we can gradually unfold the picture and examine the potential reasons.

This section will begin with the comparison between the learner (LC) and the reference corpus (RC) to examine the overall pattern between Chinese and British students. It will be followed by comparisons of the four sub-corpora (two from each corpus) to examine the variations between student groups in each comparable discipline and between different disciplines in each student group. To recap, the four sub-corpora are Business and Management (LC-BM) and English literature (LC-EL) in the learner corpus and Social Science (RC-SS) and Arts and Humanities (RC-AH) in the reference corpus.

6.2.1 Comparison between student groups

A significant association can be found between the first languages students use and the three modals. The total normalised frequency of the three modals is higher in the learner corpus (3745.10 per million words) than that in the reference corpus (2735.58 per million words). *Have to* is slightly over-represented in the learner corpus compared to the reference corpus, with a normalised frequency of 717.52 and 565.66 respectively. By contrast, the use of *should* and *must* shows marked differences. Chinese students use *should* over two times more frequently than British students, while their use of *must* is only half the frequency observed in the reference corpus.

The over-representation of *should* is also observed in previous studies such as Ma and Lu (2007) and Tang (2013). As for *must*, it is under-represented in the present study and in Tang (2013), but it is shown to be overly used by Chinese students in Cheng and Qiu (2007), Liang (2008), and Long (2013). Another stream of studies (e.g., Bai, 2015 and Yang, 2008) demonstrates that there is no difference in the use of *must* between Chinese students and native English speakers. This divergence in findings of frequency distribution may result from the nature of the writing examined, as discussed in Section 2.4.2. Most of the previous studies examine short argumentative writings on various topics, which differ from the writings examined in the present study in several aspects, such as length of the texts, genre, and discipline. Yang (2018), on the other hand, investigates the undergraduates' writings in International Business and Trade,

which is similar to one of the disciplines in the present study, Business and Management. Yang compares the results with relevant journal articles and reports no significant difference in the use of *should* and *must* between the two corpora. The difference in findings between Yang's (2018) analysis and the present study might be due to the selection of reference corpus. As explained in Section 3.2.2, British students' writing seems to be a more appropriate reference compared to journal articles since the two groups of writers are of similar age and have similar readers.

Having compared the present findings with the previous literature, I will now discuss the potential reasons for the over-representation of *should* and under-representation of *must* in the learner corpus. One explanation is related to a phenomenon called 'lexical teddy bear', which is proposed by Hasselgren (1994, p. 237) to describe the tendency for learners to use the words that they feel safe with. Hasselgren concludes that this feeling can be derived from two aspects, one of which is the resemblance between the learner's first and second language, and the other is the early exposure to the words.

As mentioned in Section 2.4.1, Chinese translations of *should*, *yinggai [应该]* and its variation *gai [该]*, apply to both the epistemic and root meanings, which share the polysemous feature of *should* in English. In addition, it is recognised as a modal

auxiliary verb and is therefore used in the same syntactic position as *should*. By contrast, one of the main translations of root *must* in Chinese, *bixu [必须]*, is controversial regarding its part of speech. It is recognised as a modal auxiliary verb by Lin (2012), but Cai (2010) proposes that it is a modal adverb. Besides, *must* has a wider variety of Chinese translations, some of which can express both root and epistemic meanings such as *dei [得]*, while the others can only express one of the meanings such as *yiding [一定]* and *bixu [必须]* expressing epistemic and root sense respectively. The use of *should* in Chinese is more similar to its English equivalent than *must*, which may lead to the difference between the student groups.

Another reason could be the earlier exposure to *should* than *must* in junior high school English textbooks where Chinese students are first introduced explicitly to the use of modals as a grammar point. It is important to clarify that there are eight editions of junior high school English textbooks, used by schools in different provinces, as stated by the Chinese Ministry of Education (2020). All these textbooks, despite differences in presentation and design, follow the same guidelines developed by the Ministry of Education in China, the English Language Curriculum Standards for Compulsory Education, and thus they are similar in content and topics covered. The present study will mainly discuss one of the editions, titled *English,* published by Yilin Press in 2013 as a representative since the compiler of the learner corpus did not document which

textbook edition the students used. This textbook series consists of two books for each Grade (Grade 7 to 9) and was analysed by Sun (2018) for the use of modals. If not specified, the textbooks in discussion are from this edition. An additional edition, *Go for it,* which is used by schools in Guangdong Province and examined by Li (2020), will also be mentioned occasionally.

Should is the first modal among the three that is explicitly introduced to the Chinese students in the junior school English textbook. It is covered in the first semester of Grade 8, while *must* and *have to* are introduced in the second semester of the same Grade. Li (2020) examines the essays written by junior high school Chinese students and finds that *should* appears earlier in their writings than *must*. It is also used more frequently and with a lower error rate compared to *must*. Although these three modals may be used by teachers in the classroom and exposed to the students prior to their introduction in textbooks, it is hard to measure the extent of this exposure. The explanation of the three modals in the textbook, on the other hand, is in detail with examples to demonstrate their meanings and functions (will be presented in Section 6.3.1). They will be tested in the exams since it is an important part of the curriculum, and thus this will raise the student's awareness of the importance of learning these modals.

6.2.2 Comparison between disciplines

The previous sub-section has presented the variations between the two student groups and discussed potential reasons, such as the influence of first languages and the order in which target modals are introduced. To examine disciplinary variations, the comparison is made between the four sub-corpora, two from each corpus. Table 6.1 is a summary of the normalised frequency per millions words of the three modals in LC-BM, RC-SS, LC-EL, and RC-AH.

Table 6.1 Normalised frequency per million words of *must*, *have to*, and *should* in the four sub-corpora

	Must	Have to	Should
LC-BM	444.00	364.95	1,537.92
RC-SS	666.05	220.62	584.11
LC-EL	271.68	352.58	773.98
RC-AH	499.26	345.05	420.49

The overall frequency difference between the two student groups mentioned before can be observed in each pair of comparable sub-corpora, between LC-BM and RC-SS, and between LC-EL and RC-AH. Chinese students in both disciplines over-represent *should* and *have to* and under-represent *must* compared to their British counterparts, and the potential reasons have been discussed above. This consistent pattern suggests that the impact of the first languages on the use of the modals in Chinese EFL students' academic writing may be stronger than that of the disciplines.

Let us now move on to look at different disciplines in the same corpus. Generally, the difference between LC-BM and LC-EL in the learner corpus is greater than that between RC-SS and RC-AH in the reference corpus. Disciplinary variations of *must* and should are similar in both corpora, with higher normalised frequencies in LC-BM and RC-SS than in LC-EL and RC-AH. *Have to* is used in similar normalised frequency between the two disciplines in the learner corpus, whereas it is less frequently used in RC-SS than in RC-AH.

The disciplinary variations may be attributed to the different disciplinary conventions and analytical approaches, which will be explored further in Section 6.3.2. A detailed discussion is deferred because the three modals in question are polysemous, and a deeper understanding of their disciplinary variation requires analysing which meanings are most influential. As highlighted in previous studies (see Section 2.4.2), the overall frequency alone provides only limited insights into the use of these modals. To gain a clearer understanding of the modal profiles in Chinese EFL students' academic writing, the second research question is posed, focusing on the distribution of their meanings.

6.3 Meaning distribution of the three modals across sub-corpora

This section follows the same structure as the previous one, starting with a comparison between the learner and the reference corpus as a whole. This is followed by a comparison of the four sub-corpora to further investigate the difference between student groups and between disciplines. The two meanings of the modals, epistemic and root, are discussed separately.

6.3.1 Comparison between student groups

The normalised frequency distribution of the two meanings of *must*, *have to* and *should* shows some similarities between the learner and the reference corpus. Epistemic sense of the modals is used markedly less frequently than their root use, especially for *have to*. *Have to* displays the largest normalised frequency difference between the two meanings, followed by *should*, while *must* demonstrates the smallest difference. This result is similar to what Biber et al. (1999) find in academic prose, that *must* is used most evenly between the two meanings compared to the other two modals. The proportions of the three modals expressing epistemic meaning show similar patterns between the two student groups as well, with *must* ranking first (over 50%), followed by *should* (around 40%) and *have to* (under 5%). The Pearson Chi-squared test conducted in Section 4.3.2 suggests that there is no statistically significant difference in the distribution of epistemic use of the three modals between the two student groups.

Despite not being statistically significant, Chinese students slightly under-represent epistemic *must* compared to their British counterparts, while epistemic *have to* and *should* are used similarly by the two student groups. This observation contrasts with Chen's (2012) findings, where Chinese undergraduates use epistemic *must* and *should* more frequently in argumentative essays on general topics in CLEC than native English speakers in LOCNESS. Hu and Li (2015) also find that epistemic *must* is over-represented by Chinese students in ICNALE. On reason for the differences in findings could result from the nature of the writings examined, as discussed in Section 6.2.1. Students in previous studies tend to focus primarily on constructing persuasive propositions in argumentative essays on topics such as friendship or smoking. Conversely, those writing discipline-specific dissertations in the present study are required to engage in a broader range of rhetorical functions, including providing information and conducting detailed analyses. This variation may lead to different patterns in the epistemic use of the modals.

Hu and Li (2015) further argue that the use of epistemic modality is related to student's language proficiency. Chinese students with higher proficiency levels tend to be more tentative and use epistemic expressions more similarly to native speakers. Gao (2023) supports this view, noting that students rated at lower proficiency levels (CSE 1-5) tend

to over-represent *must*, whereas those at higher levels (CSE 6-7) under-represent it compared to native English speakers. This also helps to explain the inconsistency between the present findings and the previous literature. The writings of Chinese undergraduates in Chen's (2012) and Hu and Li's (2015) studies were collected in 2003 and 2010 respectively, whereas the learner corpus in the present study was compiled around 2018. Advancements in English education in China over the past decade may have elevated the proficiency levels of students, thus narrowing the gap in the use of epistemic modals between Chinese students and native speakers.

The slight under-representation of epistemic use of *must* may be due to the lack of textbook input as the epistemic sense is not explicitly introduced in junior high school textbooks in China. *Must* is exclusively used in its root sense in the textbooks, without explicit explanations or implicit examples of their epistemic usage throughout the text (Sun, 2018). With no detailed presentation and examples, it is hard for the students to notice its epistemic sense and use it in their writing.

Even when a different edition of the Grade 9 textbook introduces the epistemic use of *must*, as examined by Li (2020), the junior high school students still do not it in their writings. Li suggests that this could be because students lack the ability to express reasoning and inference at that age. This observation aligns with Pienemann's

Teachability Hypothesis (1989), which suggests that language acquisition is closely tied to cognitive development and that students can only effectively employ linguistic structures such as epistemic modals when they have reached the necessary cognitive level. In our case, although Chinese undergraduates may have developed the cognitive ability to express their assessment of the truth of a proposition, this remains challenging because epistemic use of the modals is acquired later than their root use by language learners, as pointed out by Papafragou (1998). As a result, they may avoid these expressions or opt for more familiar alternatives, potentially causing a slight under-representation of epistemic *must*.

Another reason could be the cultural differences. Confucianism, which is deeply rooted in Chinese culture, values harmony and moderation and avoids extremes (Li, 2016). This preference for balance and moderation may lead to a less assertive and more tentative tone when assessing the truth of a proposition compared to their British counterparts. However, this is merely an assumption which needs further investigation of other epistemic expressions to support this.

In regard to the root use of the modals, it shows a more complicated picture than their epistemic use. There is a significant association between the first language of the students and root use of the three modals, and the association is medium (see Section

4.3.2). The total normalised frequency of the root use of the three modals is higher in the learner corpus than in the reference corpus, and it is mostly due to the predominant use of root *should*. Root *should* is used over 2.5 times more frequently in the learner corpus than in the reference corpus. Similarly, root *have to* is over-represented by Chinese students, though with a smaller difference in normalised frequency between the student groups compared to root *should*, being about 1.2 times more frequent than in the reference corpus. By contrast, Chinese students use only about half as many instances of root *must* as their British counterparts. These trends can also be observed when looking at the proportion of the root use of the three modals. In the learner corpus, *should* accounts for over 60% of all instances expressing root meaning, and it is followed by *have to* and *must*, each constituting approximately 18%. By contrast, root *must* and *should* share a similar proportion in the reference corpus, each accounting for about 38%, while root *have to* only constitutes 24%.

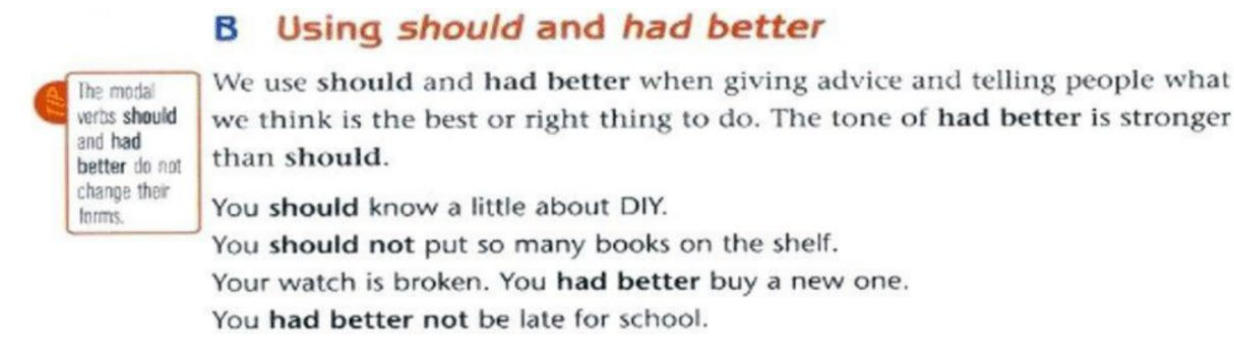
As mentioned above, root *have to* is used similarly by the two student groups concerning the proportion across the three modals, exhibiting a marginally higher normalised frequency in the learner corpus. The following paragraphs will thus focus on exploring the reasons for the under-representation of root *must* and over-representation of root *should* in the learner corpus. As discussed in Section 2.2.5, root *must* conveys a stronger sense of obligation compared to root *should*. Root *must* is

mostly used when the writer has authority or power over the addressee and expects the addressee to fulfil the obligation. *Should*, by contrast, allows for 'non-actualization' (Collins, 2009, p. 45), that is it does not imply a high expectation for the addressee to follow the suggestion. This difference is not equally observed in the Chinese translation of root *must* and *should*. Li (2020) proposes that root use of *yinggai [应该] 'should'* has little pragmatic difference from its English equivalent, whereas the Chinese translation of root *must* shows major differences since it does not imply the writer's or the speaker's authoritative position. Therefore, Chinese students might find root *should* more familiar due to its resemblance to the Chinese translation.

The cultural and social values may also contribute to the different preferences in the root modals between student groups, as proposed by Hinkel (1995) and Kecskes and Kirner-Ludwig (2017). As discussed before, Chinese culture is influenced by Confucianism, which encourages the principle of moderation and avoids extremes. This may lead to more use of root *should* rather than root *must*. Similarly, Bu (2011) relates the difference in using the suggestion strategies to cultural perceptions of suggestions, arguing that Chinese culture values harmony and collectivism and considers suggestions as a way to maintain amicable interpersonal relations. The sense of duty and accountability toward the community and the nation seems to have been culturally ingrained in Chinese students' identity from a young age, leading to a

higher total normalised frequency of the three root modals than their British counterparts. The cultural focus on harmony and moderation may also result in a preference for using root *should* over root *must*, given that the latter conveys a more intense degree of compulsion. By contrast, such a positive image of suggestions is less dominant in English culture, where sometimes suggestions may be taken as an offence because English society is more of an individualistic society and an unsolicited suggestion might not be appreciated if not given cautiously. For example, 'You *must* try the food at this restaurant, it's really delicious!' is socially accepted in China to express a strong recommendation of the restaurant and show hospitality. It does not expect or require the listener to try this food. In English, this sentence may sound overly authoritative or commanding in certain contexts such as when the speaker and the listener are not familiar with each other.

Another possible explanation is related to teaching material. Figures 6.1 and 6.2 below show how the three modals are presented in junior high school textbooks in China. You may notice that the relevant part is labelled as B, as in 'B Using *should* and *had better*'. This is because each unit (eight units per textbook) introduces approximately two grammar structures which are labelled as Part A and B, and they are mostly related to each other.



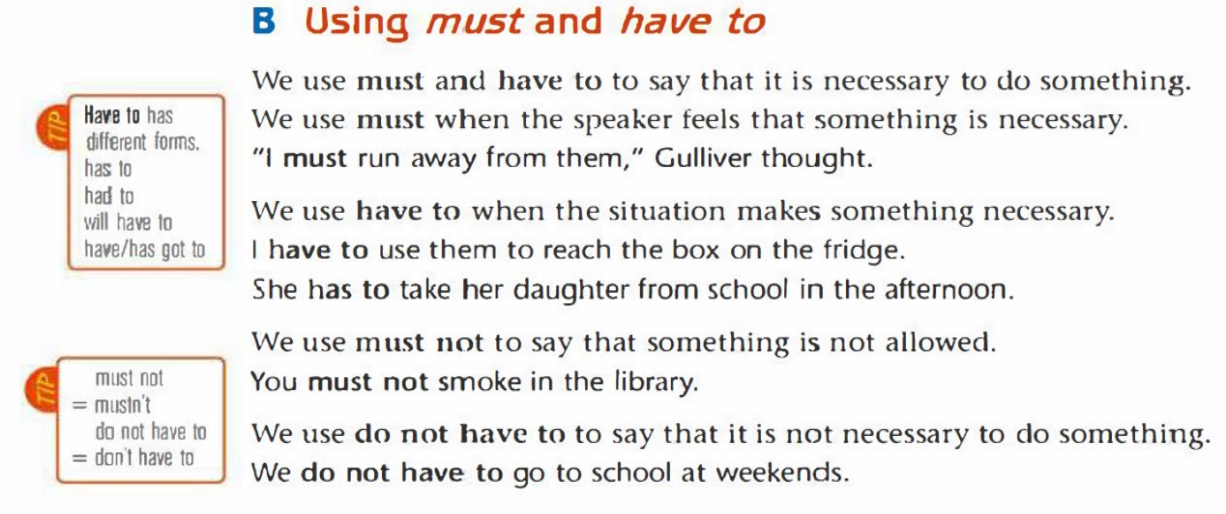
B Using *should* and *had better*

TIP The modal verbs **should** and **had better** do not change their forms.

We use **should** and **had better** when giving advice and telling people what we think is the best or right thing to do. The tone of **had better** is stronger than **should**.

You **should** know a little about DIY.
You **should not** put so many books on the shelf.
Your watch is broken. You **had better** buy a new one.
You **had better not** be late for school.

Figure 6.1 Screenshot of *should* introduced in the junior high school English textbooks (Grade 8 Semester 1) in China



B Using *must* and *have to*

TIP **Have to** has different forms.
has to
had to
will have to
have/has got to

We use **must** and **have to** to say that it is necessary to do something.
We use **must** when the speaker feels that something is necessary.
"I **must** run away from them," Gulliver thought.

We use **have to** when the situation makes something necessary.
I **have to** use them to reach the box on the fridge.
She **has to** take her daughter from school in the afternoon.

TIP must not
= mustn't
do not have to
= don't have to

We use **must not** to say that something is not allowed.
You **must not** smoke in the library.

We use **do not have to** to say that it is not necessary to do something.
We **do not have to** go to school at weekends.

Figure 6.2 Screenshot of *must* and *have to* introduced in the junior high school English textbooks (Grade 8 Semester 1) in China

Before introducing root *should*, Part A presents how to give instructions and 'tell people what (not) to do' using imperatives, which is related to the function of root *should*. However, the Part A before root *must* and *have to* is irrelevant to their use, titled 'Using question words + to-infinitives'. It would make more sense to the students if the three modals are introduced in the same unit, implying their similarity in meanings and

functions. In addition, introducing them in a bundle may also help the students to differentiate them in strength through contrasting examples in various contexts.

The difference in strength is implicitly shown in the textbooks, as demonstrated in Figures 6.1 and 6.2. *Should* is introduced as a device to give advice and to demonstrate what is the best or right thing to do, which is more related to personal opinions. By contrast, *must* is introduced with *have to* to describe what is necessary to do, and its emphasis on necessity shows a higher degree of force. The introduction of root *should* and *must* may give the students the impression that root *must* can be replaced by *should* since what is necessary to do is highly likely to be the right thing to do. With no explicit comparison across the modals, it is difficult for the students to tell the difference, and thus whichever modal is introduced first may become the 'lexical teddy bear' (Hasselgren, 1994, p. 237) that the students feel more comfortable using. Root *should* is explicitly introduced one semester earlier than root *must* and *have to*, which may explain its over-representation in the learner corpus.

Another reason might be that Chinese students lack the knowledge of alternative devices to give suggestions, and thus they tend to rely heavily on root *should*. British students use a wider variety of expressions to describe suggestions such as 'I would recommend' or 'it is important to', as observed in the qualitative analysis in Section

5.4.4. Similarly, Li (2016) finds that Chinese EFL learners tend to use simplified modal sequences to express suggestions, whereas native speakers use a wider range of devices such as ‘be + required/allowed/supposed’. As previously discussed, the perception of suggestions among Chinese students differs from that of native English speakers, leading to differences in their suggestion strategies. Chinese students tend to rely on direct suggestion strategies including using obligation modals, whereas native English speakers prefer more indirect strategies that do not specify the suggestive force, such as describing the situation as in ‘The classroom is really noisy.’ (Bu, 2011, p. 31). Consequently, British students may use a wider variety of expressions compared to Chinese students.

In sum, the difference in the root use of the three modals between the two student groups can be explained by factors such as first language, cultural and social value, teaching material and knowledge of alternative devices to express suggestions. Among these factors, I would argue that the interplay of first language and cultural value emerges as the most significant, which is supported by the findings presented above, as well as corroborated by previous literature.

6.3.2 Comparison between disciplines

This sub-section will compare the four sub-corpora to examine disciplinary variations. I will first look at whether the student group variations mentioned above can be

observed in both sub-corpora in the learner and the reference corpus, and then analyse the disciplinary similarities and differences.

Let us start with the epistemic use of the three modals. The distinctions between the student groups mentioned earlier are generally consistent across comparable disciplines. Epistemic *have to* and *should* do not show marked differences across the four sub-corpora. By contrast, epistemic *must* is slightly under-represented in both sub-corpora in the learner corpus compared to those in the reference corpus. British students seem more confident in expressing their judgement of the truth of a proposition in both disciplines, using epistemic *must* more frequently than Chinese students. This could be attributed to how this sense is introduced to the Chinese students in the textbook and cultural differences, which have been discussed above.

As for the disciplinary variation, epistemic *must* is markedly under-represented in LC-BM and RC-SS compared to LC-EL and RC-AH. This result reflects that of Takimoto (2015) who finds that writers in social science use fewer boosters than those in humanities. This finding is consistent with the present results because epistemic necessity modals can also function as boosters, as illustrated in 6-1 and 6-2 below. Epistemic *must* is used to demonstrate the writer's confidence and certainty in the truth of the following proposition.

6-1 American consumer premature consumption culture *must* lead to a large amount of risk-taking consumer behavior. (Epistemic_LC-BM_UGBM 10005)

6-2 Here Wharton doesn't give much direct introduction on Ellen yet; but from the reaction of people in other boxes we have a feeling that this woman *must* have a past that is mysterious as well as regretful. (Epistemic_LC-EL_L 00041)

Additionally, in terms of the proportions of epistemic use among the three modals, students in LC-BM and RC-SS use epistemic *must* and *should* relatively evenly, whereas those in LC-EL and RC-AH use a markedly higher percentage of epistemic *must* than *should*.

These disciplinary variations can be attributed to differences in analytical approaches and rhetorical conventions used within each discipline, as mentioned in Section 2.5. In terms of the approaches, students in LC-BM and RC-SS primarily rely on empirical data collected from fieldwork or business reports for analysis, and their investigations are mostly fact-oriented and objective. Thus, these students have a lower chance to make subjective judgements, leading to the under-representation of epistemic *must*. By contrast, as discussed in Section 5.4.4, one of the approaches used by students in LS-EL and RS-English is close reading. This approach involves the use of quotations from literary texts to construct arguments, a practice not observed in the sub-samples

of LS-BM and RS-Business. The interpretations of, for example, the behaviour of the characters or the setting of the plots are relatively subjective, reflecting the students' personal insights which requires the use of epistemic modals to make subjective judgements.

Another reason is related to rhetorical conventions in specific disciplines. Peacock (2014) points out that writings in different disciplines may value different types of arguments and ways of persuasion. For example, pure disciplines are more concerned with understanding or interpreting the world whereas applied disciplines are more related to acting and making an impact (Squires, 2005). Accordingly, students in LC-EL and RC-AH, which are categorised as pure disciplines (see Section 3.2.2 for discussion), are more likely to be accepted by readers within their discipline for expressing strong commitments to their argument. The conventions in LC-BM and RC-SS (categorised as applied disciplines), on the other hand, seem to require the students to use a more hedged and impersonal tone, leading to the under-representation of epistemic *must* compared to writings in LC-EL and RC-AH.

Having explored the epistemic use of the three modals, let us examine their root use. The differences in normalised frequency of the root use of the three modals between the student groups mentioned earlier remain consistent across comparable disciplines.

Root *have to* and *should* follow a similar trend, with more instances found in LC-BM and LC-EL than in RC-SS and RC-AH respectively. In addition, the normalised frequency differences of root *should* across the sub-corpora are greater than those of root *have to*. By contrast, root *must* is markedly under-represented by the Chinese students compared to the British students, and this pattern also applies to both disciplines. The potential reasons between the student groups have been discussed in Section 6.3.1.

Apart from these reasons, the predominant use of root *should* in LC-BM requires further discussion because it is used nearly three times more frequently than writings in RC-SS, and this over-representation is much more salient than root *have to* across sub-corpora. This might be related to the rubric of dissertations or instructions given by lecturers, which asks the students to give suggestions or recommendations in the dissertation. It is a pity that the guidelines of the learner texts were removed by Zou (2018) when compiling the corpus. However, evidence can still be found in the qualitative analysis (see Section 5.4.2) where the sample texts were divided into different parts. Root *should* is more frequently used in the conclusion part of LC-BM than in the other three sub-corpora. Since the conclusion part normally describes what actions need to be taken, as confirmed by a thorough review of selected texts, the higher frequency of root *should* in this part may result from the guidelines provided by

lecturers, or a convention to give suggestions at the end of a dissertation in LC-BM.

As for the disciplinary variation, root *must* and *should* show similar patterns, with a markedly higher normalised frequency in LC-BM and RC-SS compared to LC-EL and RC-AH. This could be explained by the fact that writers in applied disciplines such as BM and SS are primarily concerned with 'acting rather than knowing' (Squires, 2005, p. 130). Thus, they place more emphasis on reflecting upon current practices and giving suggestions to impact the world compared to writers in pure disciplines like EL and AH.

The normalised frequency of root *have to* is similar in the three sub-corpora, being over 1.5 times as frequent as in RC-SS. The under-representation of root have to in RC-SS compared to RC-AH could be attributed to the objectivity conveyed by root *have to*. Compared to root *must*, the source of obligation of root *have to* is mostly external without the involvement of speakers (Collins, 2009). Given that students in RC-SS over-represent root *must* compared to those in the other three sub-corpora (see Section 4.4.1), it seems that they are more inclined to express obligation subjectively rather than objectively.

6.4 Verb collocates of the three modals across sub-corpora

We have examined two aspects of the profiles of *must*, *have to*, and *should*, their frequency and meaning distribution. The third research question shifts the focus from frequency to the semantics of main verbs used with the modals, offering a new perspective on modal use through distributional semantic analysis and providing potential explanations for variations across sub-corpora. The epistemic and root uses of the three modals present different patterns concerning their verb collocates, and thus they will be addressed separately in the following two sub-sections. The comparisons between the student groups and between the disciplines will be discussed in each sub-section.

6.4.1 Verb collocates of epistemic *must*, *have to* and *should* between student groups and between disciplines

In general, epistemic *must*, *have to*, and *should* are used mostly with stative verbs, which is in accordance with Coates's (1983) and Biber et al.'s (1999) findings. Among the stative verbs, main verb *be* is the most frequently used one. Epistemic *must* predominantly collocates with main verb *be* in every sub-corpus. The percentage of this combination relative to the total absolute frequency is approximately 50% in three sub-corpora except in RC-AH, which shows a relatively lower percentage of 39%, leaving more instances to collocate with other verbs. The co-text following 'must + be' shows a similar pattern across the four sub-corpora, with nouns ranking first, followed

by adjectives and prepositions. Coates (1983) and Warchał (2007) identify a strong association between existential subject *there* and epistemic *must*, which can also be observed in the present study. However, they overlook the co-text following 'there must'. In this study, all the instances of 'there must be' are followed by a noun.

There is no marked difference in the co-text following epistemic 'must + be' between the student groups, whereas disciplinary variation can be found in that students in LC-EL and RC-AH use slightly more combinations of 'must + be + adjective' than those in LC-BM and RC-SS. This could result from a higher frequency of adjectives in LC-EL and RC-AH, leading to a higher chance to modalise them. However, this reason has been ruled out in Section 4.4.4.1 as the number of adjectives shows the opposite trend. Another reason could be related to disciplinary conventions. Writings in LC-EL and RC-AH tend to be more interpretative and descriptive in the analysis and thus are more likely to include the assessment of the truth of the quality conveyed by an adjective compared to those in LC-BM and RC-SS.

Since nearly half of the instances of epistemic *must* co-occur with main verb *be*, the variety of lexical verbs collocating with it is limited, and there is no marked variation observed between student groups or between disciplines. Two distinctive verb clusters can be identified in all four sub-corpora, one which includes mental verbs such as *know*

and *see*, and the other consisting of causative verbs such as *cause* and *influence*. The identification of the first cluster aligns with prior research, which has found a strong association between epistemic modals and stative verbs, a category that includes mental verbs as a sub-group. However, the latter cluster, causative verbs, involves actions that bring about changes and impacts, corresponding to the characteristics of dynamic verbs. Students in LC-EL use relatively more of these verbs, indicating that an essential element of their analysis is to make judgements about the influential factors that lead to the situation based on evidence in literary works or previous literature.

As noted earlier, *have to* is rarely used in the epistemic sense, and its low absolute frequency across the four sub-corpora makes it difficult to identify patterns between student groups and between disciplines. However, it still shares similarities with epistemic *must* and *should*, as it is mostly used with either *be* or other stative verbs.

As for epistemic *should*, it frequently collocates with main verb *be*. This combination accounts for 71% of all the instances in LC-BM, whereas the percentage is relatively lower in the other three sub-corpora, with 33% in RC-SS and 46% in LC-EL and RC-AH. The highest percentage in LC-BM results from that students in this discipline use more combinations of ‘should + be + noun’ than those in LC-EL and RC-SS. As for the

other verbs collocating with epistemic *should*, no distinctive verb cluster can be identified in LC-BM and RC-SS. By contrast, writings in LC-EL and RC-AH each have one featured cluster, denoting emotional states (e.g., *feel* and *envy*) and attitudes (e.g., *expect* and *want*) respectively. These two verb clusters both describe the mental states of a person and are related to human behaviour. Writings in LC-EL and RC-AH often involve the analysis of human emotion and experience, and thus are more likely to use these verbs with epistemic *should*.

So far, we have discussed the semantics of the main verbs collocating with epistemic use of the three modals. The most frequently used one is main verb *be*, and writings in sub-corpora demonstrate preferences for different sub-groups of stative verbs. Due to the low absolute frequency of lexical verbs used with the three epistemic modals, the distributional semantic analysis does not seem ideal for identifying the semantic patterns of their verb collocates. The verb collocates show a wide variety of meanings and are distributed loosely in the semantic plots. However, this approach does uncover some distinctive patterns of verb collocates of the root use of the three modals, given their high absolute frequency. These patterns will be summarised in the next sub-section.

6.4.2 Verb collocates of root *must*, *have to* and *should* between student groups and between disciplines

This section will examine the semantics of verb collocates of the root use of the three modals and present them separately. Comparison between the student groups (LC-BM vs. RC-SS and LC-EL vs. RC-AH) will be discussed first, followed by the comparison between disciplines (LC-BM vs. LC-EL and RC-SS vs. RC-AH).

The verb collocates of root *must* in LC-BM show the highest degree of semantic homogeneity than those in the other three sub-corpora. Chinese students in LC-BM tend to give practical suggestions for business in similar aspects. Three characteristic clusters can be identified, including verbs denoting exploration and evaluation (e.g., *assess* and *evaluate*), improvement (e.g., *adjust* and *strengthen*), and development (e.g., *create* and *produce*). The first cluster can also be identified in the writings in RC-SS but not the latter two. Additionally, based on the discussions in the qualitative analysis regarding the categories of directives proposed by Hyland (2002), these suggestions could be classified as physical acts, since they recommend specific practices applicable in the real world.

One explanation is the difference in cultural and social values between the student groups, which has been mentioned in Section 6.3.1. According to Li (2016) and Bu (2011), Chinese students tend to have a strong sense of duty and accountability

towards their communities and nation since they value harmony and collectivism. Thus, although Chinese students do not necessarily have the power to ask the company to follow these obligations, they use root *must* with these two verb clusters to emphasise the importance and urgency of taking actions and demonstrate their concern for business development. Another reason could be that verbs denoting improvement and development are more related to the topics in LC-BM, which is to lay obligations on the company or brand to improve performance. Writings in RC-SS, on the other hand, cover a broader range of topics and disciplines, leading to a wider variety of suggestions. Among these, one distinctive cluster is related to statistics. British students in RC-SS use verbs such as *increase*, *rise* and *equal* with root *must* to give specific suggestions on the change in statistics.

As for the comparison between LC-EL and RC-AH, one marked difference is the use of mental verbs. Chinese students in LC-EL use fewer mental verbs than British students in RC-AH, which is partly due to the lower absolute frequency in general. This may also indicate that Chinese students, unlike their British counterparts, do not prefer to discuss what is advised to do at the cognitive level.

Another difference between student groups is that British students in both disciplines use verbs such as *question* and *address* with root *must* to emphasise the necessity of

challenging and critically treating propositions, whereas this verb cluster is absent in Chinese students' academic writing. As mentioned before, root *must* expresses a relatively strong sense of obligation. British students use it with these verbs to make authoritative and assertive assessments of propositions. This usage may not be accepted by Chinese students, as Chinese culture prefers moderation and maintaining harmony.

Regarding the disciplinary variation in the use of root *must*, verb collocates denoting exploration such as *study* and *examine* can be found in LC-BM and RC-SS but not in LC-EL and RC-AH. It seems that students in LC-BM and RC-SS pay more attention to the research as an activity and describe what is advised to be examined in their analysis, whereas students in LC-EL and RC-AH do not mention the examination of a problem as overtly as they do. In addition, Warchał (2007) notes the frequent use of root *must* with the pronoun *we* in linguistic-related journal articles. This combination also represents a higher proportion of all instances of root *must* in LC-EL and RC-AH compared to LC-BM and RC-SS. A similar pattern is noted with the use of root *should*, which will be mentioned later. The use of the pronoun *we* with the target root modals shows an interactive tone, encouraging the readers to agree with their suggestions. This approach seems to be more accepted in LC-EL and RC-AH, helping to build the students' authoritative voice and claim their membership in this specific research

community.

Concerning root *have to*, Chinese students in LC-BM frequently use mental verbs (e.g., *realize* and *understand*) with root *have to*, but this pattern is almost absent in the other three sub-corpora. It seems that Chinese students in LC-BM emphasise the necessity to change at the cognitive level to achieve a goal. The verb collocates in LC-EL and RC-AH do not show marked differences, with one shared cluster standing out, including verbs related to hardship and unpleasant experiences (e.g., *suffer* and *endure*). Students in RC-AH use a slightly broader variety of verbs (e.g., *tolerate* and *experience*) in this cluster than those in LC-EL possibly due to its wider coverage of disciplines. This cluster is exclusively used with root *have to* but not root *must* and *should*. The necessity to *suffer* or *endure* is derived from objective situations rather than personal and subjective viewpoints. In these cases, root *have to* cannot be replaced by *must*. This usage implies that both Chinese and British students understand this difference in subjectivity between root *have to* and *must*. This cluster is also related to the disciplinary variation since it is almost absent in LC-BM and RC-SS. Students in LC-EL and RC-AH tend to interpret characters or plots in literary works, as well as general situations and circumstances, to give suggestions. As a result, they are more likely to discuss how to deal with unpleasant and unavoidable experiences.

In addition, root *have to* shows a stronger association with *admit* in LC-BM than in the other three sub-corpora. It is also reported that this combination is mostly used with *we* in LC-BM, engaging the readers to agree with the writers' viewpoints. The use of root *have to* in conjunction with communication verbs, such as *admit,* is rarely reported. Most of the previous studies (e.g., Coates, 1983; Palmer, 1990) associate this use with root *must*, a pattern that is not salient in the present study. Furmaniak (2020) is one of the few studies to acknowledge this pairing of root *have to*, suggesting that it tends to make the statement more tentative. This combination implies that the proposition following *admit* describes an objective limitation or unexpected findings. Compared to LC-EL and RC-AH, writings in LC-BM tend to be more objective in the analyses, as they tend to rely on concrete evidence to make arguments, leading to more use of this combination.

As for the last modal to discuss, root *should*, it is similar to root *must* in that its verb collocates in LC-BM show a high degree of semantic homogeneity. Three verb clusters are identified to be exclusively used in LC-BM but not in RC-SS, and they are verbs denoting improvement (e.g., *adjust* and *modify*) and development (e.g., *generate* and *create*), and that are closely related to the disciplinary knowledge (e.g., *sell* and *market*). The first two verb clusters are also used with root *must*, as previously discussed. The similarity in the semantics of verb collocates of root *should* and *must* supports Cappelle

et al.'s (2019) finding that *must* is the modal most similar to *should* compared to *have to* and *need to*. The absence of these three verb clusters in RC-SS may be attributed to similar reasons as those discussed for root *must*, including differences in cultural values and the disciplines covered.

Verb collocates such as *accept* and *underestimate* are used by both student groups, but serve different purposes. British students mostly used them to assess arguments and viewpoints. While Chinese students occasionally use these verbs in a similar way to their British counterparts, they mostly use the combination to give suggestions for business practices in LC-BM and for characters in LC-EL. This variation in the verb collocates between the student groups is similar to the pattern observed with root *must*, which British students specifically use with a cluster of verbs associated with challenging and critically treating propositions. This use of root *should* and *must* could be categorised as cognitive acts within Hyland's (2002) functional framework of directives since it involves either introducing new arguments or directing the reader's focus to the importance of specific statements. British students seem to use root modals with these two verb clusters to assess propositions and engage with arguments, whereas Chinese students tend to focus more on suggesting practical actions. In addition, Chinese students use a wider variety of verbs denoting mental states compared to their British counterparts possibly due to the higher absolute frequency

of root *should* in the learner corpus.

As for disciplinary differences, Chinese students in LC-EL seem to use mental verbs with root *should* to serve more functions than those in LC-BM, such as describing what is advised to perceive, or presenting the suggestions given by the characters in literary works. British students in RC-AH use a cluster of mental verbs denoting expectations (e.g., *expect* and *assume*), but such a cluster is absent in RC-SS or in the learner corpus. Additionally, as mentioned earlier, there is one verb cluster that is closely related to disciplinary knowledge in LC-BM, but this cluster is absent in LC-EL. This is probably because the suggestions given in LC-BM shared a similar goal of improving the business or brand. However, students in LC-EL offer suggestions to a broader range of addressees, such as authors of literary works, practitioners, and researchers in the field. Moreover, some suggestions in LC-EL are not given by the writers but by the characters in literary works through quotation (see Section 5.4.4). Another disciplinary variation discussed with root *must* also applies to root *should*. Root *should* in LC-EL and RC-AH frequently co-occurs with the pronoun *we* whereas this usage is not prevalent in LC-BM and RC-SS, and the potential reasons have been discussed before.

6.5 Summary

In this chapter, I have summarised the quantitative and qualitative findings and discussed the potential explanations of the variations between student groups and between disciplines. The profiles of *must*, *have to*, and *should* are discussed in detail in the order of research questions, including frequency distribution, meaning distribution, and the semantics of main verbs used with them.

In terms of frequency distribution, there is a significant association between the students' first languages and their use of the three modals. *Have to* shows a slightly higher normalised frequency in the learner corpus compared to the reference corpus. *Must* is under-represented in the learner corpus, which contradicts previous literature. By contrast, *should* shows an opposite pattern and is over-represented by Chinese students, consistent with prior findings. Potential reasons for the variations between student groups include the influence of students' first language and how the modals are presented in the textbooks. As for disciplinary variations, *must* and *should* show similar patterns in the learner and the reference corpus, with higher normalised frequencies in LC-BM and RC-SS than those in LC-EL and RC-AH. By contrast, *have to* is used similarly in LC-EL and RC-AH but is under-represented in RC-SS compared to RC-AH.

As for meaning distribution, *must*, *have to*, and *should* are predominantly used to express root meaning rather than epistemic sense. There is no statistically significant difference in the epistemic use of the three modals between the two student groups. Specifically, epistemic *have to* and *should* are used with similar normalised frequency across the four sub-corpora. Epistemic *must*, on the other hand, is slightly under-represented in the learner corpus, contradicting previous research. The inconsistency in the findings could be attributed to differences in the nature of writings examined and the improvement of Chinese students' language proficiency over the past decade. The under-representation of epistemic *must* may result from textbook presentations and cultural differences. As for the disciplinary difference, epistemic *should* and *have to* are used similarly across the four sub-corpora. In contrast, epistemic *must* is markedly under-represented in LC-BM and RC-SS compared to LC-EL and RC-AH, which could be attributed to differences in analytical approaches and rhetorical conventions.

Regarding root use, there is a significant association between students' first languages and root sense of the three modals. Both root *should* and *have to* are over-represented by Chinese students compared to their British counterparts, and this applies to both disciplines. Root should shows a greater difference between the student groups than root *have to*. In contrast, root *must* shows the opposite trend, being markedly less frequently used in the learner corpus. The under-representation of root *must* and over-

representation of root *should* in the learner corpus can be attributed to first language influence, cultural values, textbook presentation, and knowledge of alternative expressions. One reason specifically tied to the marked over-representation of root *should* in LC-BM compared to RC-SS involves the rubrics and guidelines given by lecturers, as evidenced by the exploration of the modal distribution across different parts of a text in the qualitative analysis. Regarding disciplinary variations, root *must* and *should* show similar patterns, being used markedly more frequently in LC-BM and RC-SS compared to LC-EL and RC-AH. Root *have to* shows no disciplinary variation across three sub-corpora but is under-represented in RC-SS, which could be explained by the objectivity conveyed by root *have to* compared to root *must*.

As for the last aspect to explore, the semantics of the verb collocates of the modals, it is found that epistemic use of the three modals frequently co-occurs with stative verbs, especially with main verb *be*. In terms of lexical verbs collocating with the modals, epistemic *must* is found to be frequently used with mental and causative verbs in all four sub-corpora. The verb collocates of epistemic *should* show disciplinary variations in that students in LC-EL and RC-AH use emotional and attitudinal verbs respectively, while these two verb clusters are absent in LC-BM and RC-SS.

Concerning the root use, both root *must* and *should* show a high degree of semantic

homogeneity in their verb collocates in LC-BM compared to RC-SS. Each of them co-occurs with three distinctive verb clusters, two of which are the same (verbs denoting improvement and development), offering practical suggestions for the business. The difference between LC-BM and RC-SS can be attributed to differences in cultural values and the disciplines covered. In addition, root *must* and *should* co-occur with clusters of verbs to assess propositions and engage with arguments in the reference corpus, which is a pattern not observed in the learner corpus. In addition, Chinese students in LC-BM use root *have to* with mental verbs in LC-BM, but this trend is not observed in the other three sub-corpora.

Disciplinary variations can also be observed in the root use of the three modals. Root *must* is used with verbs denoting exploration in LC-BM and RC-SS but not in LC-EL and RC-AH. Both root *must* and *should* are more frequently used with the pronoun *we* in LC-EL and RC-AH compared to LC-BM and RC-SS, showing a more interactive tone. Additionally, a verb cluster denoting disciplinary knowledge is used with root *should* in LC-BM but not in LC-EL, possibly due to the topics covered. Root *have to* is used with verbs denoting unpleasant experiences in LC-EL and RC-AH but not in LC-BM and RC-SS, which can be explained by differences in analytical approaches.

This chapter has discussed the potential explanations for the differences between

student groups and between disciplines, hoping to deepen the understanding of the profiles of *must*, *have to*, and *should* in Chinese EFL students' academic writing. In the chapter that follows, I will conclude this thesis by outlining the main arguments, discussing the contributions and limitations, and suggesting directions for future research.

7 CONCLUSION

7.1 Introduction

Modality plays a crucial role in academic writing, as it concerns the interaction between the writer and the reader, demonstrating the writer's evaluation of the 'accuracy or credibility of a claim' (Hyland, 2005, p. 178) or the necessity to fulfil obligations and follow suggestions. However, EFL students often find modality challenging (Hyland & Milton, 1997; Yang, 2018). While previous research highlighted the difficulties Chinese EFL students may face, it primarily focused on argumentative essays on various topics rather than discipline-specific academic writing and rarely addressed epistemic necessity and obligation modals.

To bridge this gap, this study set out to explore the profiles of three epistemic necessity and obligation modals, *must*, *have to*, and *should,* in Chinese EFL undergraduates' academic writing and examine the impact of first languages and disciplines on modal use. This aim was achieved by applying CIA as the analytical framework, comparing a learner corpus with a reference corpus. The learner corpus was comprised of dissertations written by Chinese EFL undergraduates in two disciplines, Business and Management (BM) and English Literature (EL), sourced from CAEL-CAWE. The reference corpus was assembled from BAWE with careful filtering to guarantee

comparability, including essays written by British students in Years 2-3 in two disciplinary groups, Social Science (SS) and Arts and Humanities (AH). The two corpora were mainly compared in terms of frequency, meaning distribution, and semantics of the main verbs collocating with *must*, *have to*, and *should*. These aspects served as the basis for operationalising the profiles of the three modals. Additionally, a qualitative analysis was conducted to examine comparable texts in a finer-grained way and identify the distinctive features of the modals in academic writing.

Having discussed the overview of the thesis in this sub-section, this chapter proceeds to summarise the main arguments in Section 7.2. It then examines the contributions this work makes to the fields of modality in academic writing and EFL student writing. Section 7.4 identifies the limitations of the study, while Section 7.5 proposes directions for future research. The chapter ends with final remarks.

7.2 Summary of main arguments

The previous chapter has summarised the findings of the study based on the research questions and presented the potential reasons for variations across sub-corpora. In general, Chinese EFL students differed from British students in the root use of the modals. In addition, although disciplinary variations were observed in the modal use, they were consistent in both the learner and reference corpora. The following

paragraphs will outline the main arguments that emerged from the discussion in two overarching themes: the contrasts between Chinese and British students, and the variations between the disciplines. Similarities in the modal use across sub-corpora were noted in the previous chapter. For example, epistemic use of the three modals are similarly used with stative verbs. The following two arguments will concentrate on the differences and their potential explanations.

1. In the epistemic use of *must*, *have to*, and *should*, no statistically significant differences were found between Chinese and British students, contrary to previous findings. However, significant differences were observed in their root use, with root *must* being under-represented and root *should* being over-represented in the learner corpus. Analyses of the semantics of their verb collocates further revealed that Chinese students in LC-BM prefer offering practical suggestions for business, while British students in both disciplines tend to use the root modals for assessing propositions. These variations may be attributed to the influence of first languages, cultural values, and textbook presentations.

The differences in epistemic use between the two student groups were not statistically significant, with a slight under-representation of epistemic *must* in the learner corpus, contradicting previous findings in Chen (2012) and Hu and Li (2015). The differences between the current and previous findings may result from the nature of the writings

examined (argumentative essays on general topics vs. discipline-specific dissertations) and the improvements in the language proficiency levels of the Chinese students over the past decade.

The under-representation of epistemic *must* in the learner corpus could be attributed to the insufficiency of both implicit and explicit exposure to the epistemic sense in junior high school English textbooks in China. In addition, cultural differences could also be one of the factors in that Chinese culture values harmony and moderation and thus a tentative tone is preferred when making judgements about the truth of a proposition.

Statistically significant differences in normalised frequency were observed between the two student groups in their root use of the modals, particularly for root *must* and *should*. Chinese students used approximately half as many instances of root *must* as their British counterparts. By contrast, root *should* was used over 2.5 times more frequently in the learner corpus than in the reference corpus. These differences can be explained by several factors, including first languages, textbook presentations, cultural values, and knowledge of alternative expressions conveying similar meanings. Root *should* is more similar to its Chinese translation compared to root *must* and is introduced earlier in junior high school English textbooks in China. Root *should* is explained following a part that gives instructions in the textbook, while root *must* is introduced in isolation,

without any connection to preceding content. Additionally, root *should* conveys a relatively weaker sense of obligation and aligns more closely with Chinese culture that values moderation. The predominant use of root *should* among the three modals in the learner corpus could also be due to the lack of alternative devices, which is confirmed by the qualitative analysis and previous studies (e.g., Bu, 2011; Li, 2016).

Despite root *must* and *should* exhibiting marked differences in normalised frequency between student groups, their verb collocates in LC-BM demonstrated a notable semantic similarity. This similarity was characterised by two shared verb clusters related to improvement (e.g., *adjust* and *strengthen*) and development (e.g., *create* and *produce*), which were not used in RC-SS. Compared to those in RC-SS, verb collocates of root *must* and *should* in LC-BM showed a high degree of homogeneity in giving suggestions, possibly due to another facet of Chinese culture, emphasising the responsibility and accountability to the community and nation. Correspondingly, Chinese students in LC-BM tended to demonstrate their concern for business and give practical suggestions that could enhance business practices. In contrast, British students in RC-SS did not show such a strong sense of duty. Instead, they used root *must* and *should* with two distinctive clusters of verb collocates respectively in both RC-SS and RC-AH: one associated with challenging and critically treating propositions (e.g., *question* and *address*) and another with assessing arguments and viewpoints

(e.g., *accept* and *underestimate*). This reflects a preference among British students for assessing propositions through the use of the root modals in academic writing, as opposed to the focus on practical actions observed among Chinese students since these two clusters were either absent or served different purposes in the learner corpus.

2. Epistemic *must* was markedly under-represented in LC-BM and RC-SS, whereas root *must* and *should* were used more frequently in these two sub-corpora compared to LC-EL and RC-AH. In addition, only students in LC-EL and RC-AH used root *have to* with verbs related to unpleasant experiences. Root *must* and *should* showed a higher percentage of usage with the pronoun *we* than those in LC-BM and RC-SS. These disciplinary variations can be explained by the differences in analytical approaches and rhetorical conventions.

There was a marked under-representation of epistemic *must* among students in LC-BM and RC-SS compared to those in LC-EL and RC-AH, which may be attributed to differences in analytical approaches and rhetorical conventions. In applied disciplines like BM and those in SS, where data is predominantly empirical and the investigation is likely to be fact-oriented, students tend to engage in less subjective judgement and use an impersonal tone, which aligns with the conventions of their disciplinary community. Conversely, in pure disciplines like EL and those in AH, students are expected to interpret data in a more subjective manner through approaches such as

close reading, which involves analysing quotations from literary works.

Regarding root use, root *must* and *should* showed consistent disciplinary variation between the two student groups, with a higher frequency in LC-BM and RC-SS compared to LC-EL and RC-AH. One explanation is the tendency of students in LC-BM and RC-SS to focus on examining problems and making an impact through the reflection of current practices and giving suggestions, in contrast to their counterparts in LC-EL and RC-AH.

Students in LC-EL and RC-AH used root *have to* with verbs related to unpleasant experiences (e.g., *suffer* and *tolerate*), a pattern not observed in LC-BM and RC-SS. They are more likely to give suggestions for addressing challenging and unavoidable experiences as their analyses include discussions on plots in literary works or general situations. In addition, students in these two sub-corpora frequently used the pronoun *we* with root *must* and *should*, demonstrating a more interactive tone to encourage readers to agree with their suggestions compared to writings in LC-BM and RC-SS. This combination seems to be more accepted by their research community for building an authoritative voice and claiming membership.

7.3 Contributions of the study

In light of these findings, this study has made significant contributions in aspects of theory, methodology, and pedagogy. I will address them in turn.

Theoretically, prior to this study, research on modality in academic writing had a preference for examining epistemic modality, and they tend to explore a list of epistemic expressions to have an overall picture (e.g., Rozumko, 2017; Sameri & Tavangar, 2013). However, this may have the drawback of not being able to examine each modal in depth. This study focused on three epistemic necessity and obligation modals, *must*, *have to*, and *should*, which allows for a fine-grained assessment of their profiles. In addition, epistemic necessity and obligation modals were under-investigated (see Section 2.3.2), and thus the present findings complemented those of earlier studies, contributing to a more comprehensive coverage of modals in academic writing.

The thesis has also provided deeper insight into the profiles of modals in academic writing. Research on modality in academic writing has been largely influenced by studies such as Coates (1983) and Palmer (1990), which offer a comprehensive discussion on modal framework. However, as mentioned in Section 2.3.1, these studies focus primarily on spoken data, and correspondingly their findings are likely to

be more representative of modals in spoken material. The present study pointed out that modality in academic writing may exhibit characteristics that are distinctive from other contexts and require further discussions, such as distribution across different parts of a text and textual voice expressed by the modals. It was observed that Chinese students in LC-BM tended to use root *should* densely in the analysis and conclusion part compared to their British counterparts in RC-SS. In addition, the identification of textual voice in using modals has been briefly mentioned by previous researchers (e.g., Huddleston, 1971) without being examined in depth. This study differentiated between the writer's voice and those of others presented through direct or indirect quotations, finding that students in RS-English used a higher percentage of root modals in direct quotations compared to LS-EL, possibly due to differences in analytical approaches.

The analysis further explored co-textual features, including the co-occurrence with syntactic features and the semantics of the main verbs that collocate with the modals, extending beyond the mere examination of frequency and meaning distribution. Although syntactic features and their association with modals have been discussed by Hermerén (1978) and Perkins (1983) qualitatively, and by Coates (1983) and Biber et al. (1999) quantitatively, the investigations were not specifically for modals in academic writing. As for the discussion on verb collocates, they have only been briefly addressed in the previous literature, such as the association between the epistemic sense of the

modals and stative verbs (Biber et al., 1999). Aiming to explore the semantics of the verb collocates more systematically, a distributional semantic analysis was conducted. This analysis revealed variations between the student groups and between the disciplines, providing insights into the actions modulated by the modals.

At a more specific level, this work contributes to existing knowledge of modal use in Chinese EFL student academic writing by examining discipline-specific academic writing. Previous research has primarily focused on short argumentative essays on general topics (e.g., Bai, 2015; Cheng & Qiu, 2007) due to the availability of Chinese learner corpora. These studies noted an over-representation of *should* in Chinese EFL writing compared to the reference corpus, a trend confirmed by the present study. This suggests that the over-representation of *should* is a consistent pattern across both argumentative essays on general topics and discipline-specific dissertations. However, this study has also demonstrated that not all findings can be universally applied across these writings. For instance, previous studies have frequently reported that *must* is over-represented in argumentative essays by Chinese EFL students compared to native English speakers. (Cheng & Qiu, 2007; Liang, 2008; Long, 2013). In contrast, this study found that *must* was used less frequently in the Chinese EFL student dissertations compared to the reference corpus. This difference in findings underscores the necessity to avoid the over-generalisation of findings from one writing

to another. Besides, while previous research in Chinese EFL writing primarily focused on frequency, this thesis also investigated the distribution of meanings and verb collocates. Therefore, variations in the usage of the three modals across sub-corpora can be more comprehensively discussed in terms of these two aspects, offering potential explanations for the observed differences.

Another theoretical implication is related to disciplinary variations. As mentioned in Section 2.5, the study of modality in academic writing has predominantly focused on research articles, and the applicability of their disciplinary variation findings to the writing of student academic writing remains uncertain. This study examined two disciplines, BM and EL, in Chinese EFL student academic writing and identified disciplinary variations such as the over-representation of root *must* and *should* in LC-BM compared to LC-EL. Additionally, the investigation into these two disciplines, classified respectively as soft-applied and soft-pure, addressed a gap in the existing literature, which has mainly focused on more distinct categorisations, such as the differences between soft and hard disciplines.

Having discussed the theoretical implications, let us now turn to the methodological ones. This study has been the first attempt to use distributional semantic analysis to examine modality in EFL student academic writing and its disciplinary variations. While

there are certain limitations in the use of this method, which will be addressed in Section 7.4, it nonetheless proves useful in expanding our understanding of modals and the actions they modulate. The present study explored the semantic similarities of these verb collocates more systematically compared to Coates (1983) and Biber et al. (1999), identifying the distinction between student groups and between disciplines and enriching the potential explanations. I believe this method is worthy of inclusion in the learner corpus research toolbox.

Although it is beyond the scope of this work to draw pedagogical implications, it is hoped that the findings and discussions presented may shed light on the improvement of language teaching and teaching materials. First, the use of student writing as the reference corpus, instead of research articles, appears advantageous for Chinese EFL students since these writings present a more attainable goal. Incorporating the reference corpus into teaching could be beneficial to provide a more authentic and precise depiction of modal profiles. For example, it could facilitate the exploration of the modal strength differences between *must* and *should* in the context and their respective distributions of meaning. In addition, it is essential to showcase a wider range of devices to express suggestion and obligation in the reference corpus, given Chinese students' pronounced preference for root *should*.

In terms of teaching material, rather than presenting them in different grades in EFL textbooks, it would be helpful to discuss them together with examples to show similarities and differences in their uses. As discussed in Section 6.3.1, Chinese junior high school English textbooks only introduce the root use of the three modals, and it is necessary to expose the students to the epistemic use implicitly or explicitly when they are in higher grades so that they could express reasoning and inference appropriately.

Finally, the study also has implications for EAP practice, highlighting the need for tailored teaching strategies based on both the student group and the disciplinary variations. Unlike their British counterparts, who tended to use root modals with verbs to critically assess propositions, Chinese students predominantly used these modals to suggest practical actions. EAP teachers could introduce exercises that involve presenting various viewpoints, demonstrating how root modals like *must* and *should* can be strategically used to interpret and critically evaluate these propositions. This approach would not only emphasise the significance of assessing propositions but also enhance students' abilities to engage in nuanced critical analysis in academic writing.

In addition, the study identified disciplinary variations, which require more discipline-specific presentations and instructions on the use of the modals in EAP teaching practices. For instance, teachers should raise students' awareness of disciplinary

conventions in using epistemic necessity modals to express subjective judgement regarding the truth of a proposition, as it was found that students in LC-EL and RC-AH used markedly more epistemic *must* than those in LC-BM and RC-SS.

7.4 Limitations of the study

Although the study has successfully presented the answers to the research questions through a mixed-methods design, some limitations remain. In terms of the quantitative strand of this study, one limitation concerns the comparability of the learner and the reference corpus. Although the present study has controlled several variables in the reference corpus such as the first language and the educational background of the students (see Section 3.2.2), there is still some inconsistency in terms of genre and discipline.

The present study adopted a broad definition of the genre, considering dissertations in the learner corpus and essays in the reference corpus to be largely similar in social purposes and generic text parts. However, there are still differences in their length and the detailed divisions of text parts, which were taken into consideration when analysing and interpreting the results. As for disciplines, compromises were made to include texts in similar disciplinary groups rather than specific disciplines in the reference corpus due to the lack of comparable texts. The explanations of the quantitative findings were

thus treated with caution, and a qualitative study on more comparable texts was conducted as a potential solution to address the concern.

Another limitation is related to the approach used, distributional semantics analysis, in terms of the data used to build the model and the effectiveness for the study. As discussed in Section 4.2.2, although using COCA to build the vector-space model for analysing main verbs used with the modals in student academic writing may appear to be problematic, it provides sufficient data to ensure the model's reliability. Moreover, the semantics of verbs are unlikely to markedly differ between COCA and the corpora used in the present study, rendering it a suitable model for this analysis. Regarding the effectiveness of this approach to examine the profiles of the modals, the study revealed some noteworthy observations in the use of root modals. However, its application to the verb collocates of epistemic uses was not as effective as the root use due to the low absolute frequency of epistemic modals and their frequent co-occurrence with the main verb *be*, which was excluded from the model.

Another weakness is in regard to the annotation of modal meanings. As discussed in Section 2.2.2, Coates (1983) points out that there are indeterminate instances which can be described as ambiguity (either/or) and merger (both/and) depending on their relationship between the epistemic and root meanings. Practically, such instances

were annotated as *unclear* in the present study. However, I must admit that there were other instances, although they were annotated as one of the two main meanings, that could be labelled differently due to lack of context or 'contextual neutralisation' (Coates, 1983, p. 17). To avoid this deviation, I invited a second rater whose first language is English to annotate 10% of the instances (see Section 4.2.1) and calculated the inter-rater reliability. Most of the annotations were consistent between me and the second annotator, and instances with disagreements were discussed until we reached a consensus. This could indicate the reliability of the annotations, but it would be ideal to ask the second rater to annotate all the data or invite more raters.

As for the qualitative strand of the study, the limitation is the number of reference texts that can be compared to the learner texts. The qualitative analysis aims to provide more evidence to the quantitative findings and deepen the discussion, which requires the sample texts to be as comparable as possible. However, the number of reference texts in similar disciplines of the learner texts is limited. There are only five texts in RC-Business that met the criteria as comparable texts and four of them were selected for analysis (see Section 5.2.1). While having four texts in each sub-sample appears sufficient for conducting qualitative analysis, this limitation restricts the potential to uncover additional evidence that could support the qualitative findings.

7.5 Suggestions for future study

This research has raised some directions for further investigation. A natural progression of this work is to analyse additional modals beyond *must*, *have to*, and *should* in order to develop a comprehensive profile of modals in Chinese EFL students' discipline-specific academic writing since the previous studies focus mostly on argumentative essays on general topics.

For the same reason, it would also be interesting to assess the influence of proficiency level on modal use in discipline-specific academic writing. Previous research (e.g., Gao, 2023; Hu & Li, 2016) has found that as proficiency level advances, learners use modals more adeptly to qualify propositions in argumentative essays. Nonetheless, research focusing on the application of this observation in Chinese EFL discipline-specific academic writing remains scarce. Given that the source of the learner corpus, CAEL-CAWE, includes dissertations from both undergraduate and postgraduate levels, an analysis comparing these two groups could shed light on variations in modal use between proficiency levels. This comparison is predicated on the assumption that progression through educational levels may serve as an indicator of increased proficiency. Alternatively, researchers can evaluate the dissertations according to the Common European Framework of Reference for Languages (CEFR), thus grounding their analysis in a widely recognised language proficiency framework.

As mentioned in Section 2.6, CIA involves two types of comparison, and this study applied one of them, comparing the interlanguage with the reference language. The other type, comparing different interlanguages of English, is also worth exploring as it helps to distinguish features that are specifically influenced by the first language from those that are commonly observed across interlanguages (Granger, 2015). One example is Kecskes and Kirner-Ludwig (2017), who identify shared features of *must* and *should* used across three groups of Asian English learners, Chinese, Japanese, and Korean, and attribute the features mostly to cultural values rather than first language.

Another possible area of future research would be to identify the subject that is responsible for modulating the proposition using modals. As mentioned in Section 2.3.1, academic writing is concerned with the integration of the voices of the writers and others such as antecedent authors, and this identification of voice is a distinctive feature in the use of modality in academic writing compared to other contexts. If the voice is not the writers', then exploring how and why the writers give their viewpoints implicitly rather than directly is worthwhile. In addition, there might be cases where the modal is added, changed, or omitted from the antecedent author's statements during paraphrasing, as illustrated in Section 5.4.4. This shows the importance of annotating

who is responsible for the judgement and checking the original literature when necessary. Such detailed annotation is more practicable in a qualitative analysis, as it is time-consuming to conduct extensive annotations in a quantitative study.

Another avenue worth exploring is the category of directives proposed by Hyland (2002), which was briefly used in the qualitative analysis chapter to analyse examples. Although this study did not primarily focus on investigating various directive devices and their functions, it did explore some examples expressed by target root modals and briefly mentioned other expressions. One of the findings is that the use of cognitive acts in the introduction part of the texts was mainly observed in the writing of Chinese EFL students, not in that of British students. Vincent (2020) also suggests that this analytical framework could offer insights into how students manage potential risks by imposing obligations or giving suggestions through different expressions. A more detailed annotation of directives and their categories in future studies would be beneficial.

Methodologically, as mentioned in Section 7.3, the present study is the first to use distributional semantic analysis in examining modality in EFL student academic writing and its disciplinary variations. Although there are some limitations, it shows the potential to explore modal use from a new perspective, revealing the semantic

similarity of the main verbs collocating with the modals. Future studies should further explore how this approach can be applied in learner corpus research concerning modality.

With the digitisation of student writing, it has become easier to compile learner corpora. However, as encountered in this study, a significant hurdle is the scarcity of texts after controlling factors to make the learner and the reference corpus comparable. This challenge possibly arises from the lack of learner corpora in academic writing. One promising initiative addressing this issue is the project named ‘Linguistic demands of EMI in Higher Education: A corpus-based analysis of reading and writing in EMI university settings in China, Italy and Thailand’ (Brezina et al., 2022), but there is still a need to compile more such corpora.

The lack of available corpora does not necessarily be the main reason. Issues with accessibility and standardisation in formatting also pose significant challenges. One solution involves establishing a unified platform where detailed information about various corpus-building projects can be submitted, such as the website ‘Learner Corpora Around the World’ (Centre for English Corpus Linguistics, 2024) at the University of Louvain. This could increase the findability and reusability of existing corpora, and the researchers can also contact the compiler to get access if necessary.

Such a platform can also include a call to action, urging researchers to standardise formats for data collection and sharing. Although such platform or manual has been developed, such as Frey et al.'s (2020) project on building a learner corpus infrastructure, and Granger et al.'s (2022) publication of the Louvain error tagging manual, there is still a long way to go. Looking forward, the research community should aim towards a future where learner corpora not only encompass a wider range of texts but are also freely available and standardised across countries. This would markedly ease the research process and foster a more inclusive and extensive research environment. Specifically, encouraging the development of student academic writing corpora and their availability will allow for more nuanced and comprehensive analyses.

7.6 Concluding remarks

This thesis contributes significantly to research on the use of modals in academic writing and learner corpus research focusing on Chinese EFL undergraduate discipline-specific dissertations. The analyses have identified differences in normalised frequency of *must*, *have to*, and *should* between Chinese EFL students and British students, and highlighted a preference among Chinese students for using root sense of these modals to suggest actions, in contrast to British students who tend to use them to assess propositions. The study also revealed disciplinary variations in the use of these modals between BM and EL, complementing previous studies that compared

modal use across other disciplines in SS and AH. These differences, both between student groups and between disciplines, add to the existing knowledge of modality in academic writing. The study also shifts the focus of modality in Chinese EFL academic writing towards the examination of discipline-specific dissertations rather than argumentative essays on general topics. It is hoped that this thesis has contributed to the development of a more comprehensive profile of the modals in EFL student academic writing and that it will inspire further research in similar areas.

References

Abdessalem, H. (2020). Writer-author presence and responsibility in attribution and averral: A model for the analysis of academic discourse. *Arab Journal of Applied Linguistics*, *5*(1).

Alfaifi, A., Atwell, E., & Hedaya, I. (2014). Arabic Learner Corpus (ALC) v2: A new written and spoken corpus of Arabic learners. In *Proceedings of the Learner Corpus Studies in Asia and the World (LCSAW) 2014* (pp. 31-01). Kobe, Japan.

Anthony, L. (2020). *AntConc* (Version 3.5.9) [Computer Software]. Tokyo, Japan: Waseda University. Available from https://www.laurenceanthony.net/software

Bai, Y. J. (2015). Jiyu Yingyu Yuliaokude Zhongguoxuesheng Yingyuqingtaidongci Xideyanjiu (A Study on Modal Verbs Acquisition of Chinese Students — A corpus study based on CLEC). *Qiqihaerdaxue xuebao* (Journal of Qiqihar University: Phi & Soc Sci), (5), 122-124.

Balakrishnan, N., Voinov, V., & Nikulin, M. S. (2013). *Chi-squared goodness of fit tests with applications*. Academic Press.

Becher, T. & Trowler, P. R. (2001). *Academic tribes and territories: intellectual enquiry and the cultures of disciplines* (2nd ed.). Open University Press.

Bhalla, V., & Klimcikova, K. (2019, August). Evaluation of automatic collocation extraction methods for language learning. In *Proceedings of the Fourteenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 264-274).

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finnegan, E. (1999). *Longman Grammar of Spoken and Written English*. Longman.

Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (2016). On the (non) utility of Juilland's D to measure lexical dispersion in large corpora. *International Journal of Corpus Linguistics*, 21(4), 439-464.

Blanchard, D., Tetreault, J., Higgins, D., Cahill, A., & Chodorow, M. (2014). *ETS Corpus of Non-Native Written English LDC2014T06*. Linguistic Data Consortium.

Bley-Vroman, R. (1983). The comparative fallacy in interlanguage studies: the case of systematicity, *Language Learning* 33, 1–17.

BNC Consortium. (2007).*The British National Corpus, XML Edition*. Oxford Text Archive. http://hdl.handle.net/20.500.12024/2554.

Brezina, V. (2018). *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

Brezina, V., Gablasova, D., Harding, L., & Bottini, R. (2022). Linguistic demands of EMI in Higher Education: A corpus-based analysis of reading and writing in EMI university settings in China, Italy and Thailand. *Work in progress*. Retrieved from https://www.research.lancs.ac.uk/portal/en/upmprojects/linguistic-demands-of-emi-in-higher-education-a-corpusbased-analysis-of-reading-and-writing-in-emi-university-settings-in-china-italy-and-thailand(39b46186-40c4-483d-a259-7c9c0f0e989d).html.

Bu, J. (2011). A study of pragmatic transfer in suggestion strategies by Chinese learners of English. *Studies in Literature and Language*, *3*(2), 28.

Cai, W.T. (2010). Tanhanyumotaicide fenbuyuquan shizhiduiyingguanxi (On the Correspondence between the Distribution and Interpretation of Chinese Modal Words). *Zhongguoyuwen* (Chinese), (3), 208-221.

Cappelle, B., Depraetere, I., & Lesuisse, M. (2019). The necessity modals *have to*, *must*, need to, and *should*: Using n-grams to help identify common and distinct semantic and pragmatic aspects. *Constructions and Frames*, *11*(2), 220-243.

Centre for English Corpus Linguistics. (2024, February 28). *Learner corpora around the world*. Université catholique de Louvain. https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html.

Chen, W. (1978). *A Brief Introduction to Grammar*. Shanghai Education Press.

Chen, Z. (2012). Expression of Epistemic Stance in EFL Chinese University Students' Writing. *English Language Teaching*, *5*(10), 173-179.

Cheng, X. T. & Qiu, J. (2007). Zhongguoxuesheng Yingyuzuowenzhong Qingtaidongcide Shiyongqingkuang——yixiangjiyu yuliaokudeyanjiu (The Use of Modal Verbs in Chinese EFL Learners' Compositions: A Corpus-based Study). *Waiyudianhuajiaoxue* (CAFLE), (6), 9-15.

Chinese Ministry of Education. (2020, April 07). *Jiaoyubu Bangongting Guanyu Yinfa 2020 Nian Zhongxiaoxue Jiaoxue Yongshu Mulu de Tongzhi* (Notice from the Ministry of Education on Issuing the 2020 Catalogue of Textbooks for Primary and Secondary Schools). Retrieved from http://www.moe.gov.cn/srcsite/A26/moe_714/202004/t20200417_444236.html.

CKIP. (1993). *Zhongwen Cilei Fenxi* (Analysis of Chinese Lexical Categories) (3rd ed.). Academic Sinica.

Coates, J. (1983). *The Semantics of the Modal Auxiliaries*. Croom Helm.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Erlbaum.

Collins, P. (2009). *Modals and quasi-modals in English*. Rodopi.

Cook, V. (1997). Monolingual bias in second language acquisition research. *Revista Canaria de Estudios Ingleses*, 34, 35–50.

Cooke, R. A., & Szumal, J. L. (2000). Using the Organizational Culture Inventory to understand the operating cultures of organizations. In N. M. Ashkanasy, C. P. M. Wilderom, & M. F. Peterson (Eds.), *Handbook of organizational culture and climate* (pp. 147–163). Sage.

Council of Europe. (2020). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume*. Council of Europe. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4.

Davies, M. (2004). *British National Corpus* (from Oxford University Press). Available online at https://www.english-corpora.org/bnc/.

Davies, M. (2008-). *The Corpus of Contemporary American English* (COCA). Retrieved from https://www.english-corpora.org/coca/.

Davies, M. (2010). *The Corpus of Historical American English (COHA)*. Available online at https://www.english-corpora.org/coha/.

De Haan, F. (2012). The relevance of constructions for the interpretation of modal meaning: The case of *must*. *English Studies*, *93*(6), 700-728.

Depraetere, I., & Reed, S. (2020). Mood and modality in English. *The handbook of English linguistics*, 207-227.

Deshors, S. C. (2016). *Multidimensional perspectives on interlanguage: Exploring may and can across learner corpora*. UCL Presses Universitaires De Louvain.

Dontcheva-Navratilova, O., Adam, M., Povolná, R., Vogel, R., & Dontcheva-Navratilova, O. (2020). Persuasion in Academic Discourse: Metadiscourse as a Means of Persuasion in Anglophone and Czech Linguistics and Economics Research Articles. *Persuasion in Specialised Discourses*, 121-158.

Dudley-Evans, T. (1999). The dissertation: A case of neglect. *Issues in EAP writing research and instruction*, 28-36.

Durrant, P., & Mathews-Aydınlı, J. (2011). A function-first approach to identifying formulaic language in academic writing. *English for Specific Purposes*, *30*(1), 58-72.

Ehrman, M. (1966). *The meanings of the modals in present-day American English*. Mouton & Co.

Ellis, R. (1994). *The study of second language acquisition*. Oxford University Press.

Ellis, R. (2006). Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge. *Applied linguistics*, 27(3), 431-463.

Ewer, J. (1979). *The modals in formal scientific discourse: Function, meaning and use*. Santiago, Chile, University of Chile, Department of English Research Report Mimeograph.

Farrokhi, F., & Emami, S. (2008). Hedges and boosters in academic writing: native vs. non-native research articles in applied linguistics and engineering. *Journal of English Language Pedagogy and Practice*, 1(2), 62-98.

Feinstein, A. R., & Cicchetti, D. V. (1990). High agreement but low kappa: I. The problems of two paradoxes. *Journal of clinical epidemiology*, 43(6), 543-549.

Fellbaum, C. (1998, ed.) *WordNet: An Electronic Lexical Database*. Cambridge, MIT Press.

Firth, J. R. (1957). A synopsis of linguistic theory 1930–55. In *Studies in linguistic analysis* (pp. 1–32). The Philological Society.

Francis, W. N., & Kucera, H. (1979). Brown corpus manual. *Letters to the Editor*, *5*(2), 7.

Frey, J. C., König, A., & Fišer, D. (2020). Creating a learner corpus infrastructure: Experiences from making learner corpora available. In *ITM Web of Conferences* (Vol. 33, p. 03006). EDP Sciences.

Furmaniak, G. (2020). On the (con)textual properties of must, have to and shall: An integrative account. In P. Hohaus & R. Schulze (Eds.), *Re-assessing modalising modal expressions: Categories, co-text and context* (pp. 281-310). John Benjamins.

Gabrielatos, C. (2010). *A corpus-based examination of English if-conditionals through the lens of modality: Nature and types* (Doctoral thesis, Lancaster University).

Gabrielatos, C., & Sarmento, S. (2006). Central modals in an aviation corpus: Frequency and distribution. *Letras de Hoje*, 41(2), 215-240.

Gao, M. (1986). *On Chinese Grammar*. Beijing Commercial Press.

Gao, X. (2023). A cross-sectional investigation of the use of modal verbs in Chinese EFL learners' English writing. *Porta Linguarum: revista internacional de didáctica de las lenguas extranjeras*, (40), 41-55.

Gardner, S., & Holmes, J. (2010). From section headings to assignment macrostructures in undergraduate student writing. In *Thresholds and Potentialities of Systemic Functional Linguistics: Multilingual, Multimodal and Other Specialised Discourses* (pp. 268-290). EUT Edizioni Università di Trieste.

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

Gilquin, G. (2000/2001). The integrated contrastive model: Spicing up your data. *Languages in contrast*, *3*(1), 95-123.

Gilquin, G. (2022). One norm to rule them all? Corpus-derived norms in learner corpus research and foreign language teaching. *Language Teaching*, 55(1), 87-99.

Giltrow, J. (2005). Modern conscience: Modalities of obligation in research genres. *Text-Interdisciplinary Journal for the Study of Discourse*, *25*(2), 171-199.

Granger, S. (1996). From CA to CIA and back: An integrated contrastive approach to bilingual and learner computerized corpora. *Languages in contrast: Text-based cross-linguistic studies*, 37-51.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In Granger, S. (Ed.), *Learner English on Computer* (pp. 3-18). Addison Wesley Longman.

Granger, S. (2012). How to use foreign and second language learner corpora? In A. Mackey & S.M. Gass (Eds.), *Research methods in second language acquisition: A practical guide* (pp. 7-29). Wiley-Blackwell.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7-24.

Granger, S., Dupont, M., Meunier, F., Naets, H. & Paquot, M. (2020) *The International Corpus of Learner English. Version 3*. Presses universitaires de Louvain.

Granger, S., Gilquin, G., & Meunier, F. (Eds.). (2015). *The Cambridge handbook of learner corpus research*. Cambridge University Press.

Granger, S., Swallow, H., & Thewissen, J. (2022). *The Louvain error tagging manual (Version 2.0)*. Louvain: Centre for English Corpus Linguistics, Université catholique de Louvain. Retrieved from https://cdn.uclouvain.be/groups/cms-editors-cecl/Granger%20et%20al._Error%20 tagging%20manual_v2.0_2022.pdf

Greenbaum, S., & Nelson, G. (1996). The international corpus of English (ICE) project. *World Englishes*, *15*(1), 3-15.

Groom, N. (2000). Attribution and averral revisited: Three perspectives on manifest intertextuality in academic writing. In P. Thompson (Ed.), *Patterns and perspectives: Insights into EAP writing practice* (pp. 14-25). The University of Reading.

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

Gwet, K. L. (2008). Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1), 29-48.

Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters*. Advanced Analytics, LLC.

Halliday, M. A. K. (1970). Functional diversity in language as seen from a consideration of modality and mood in English. *Foundations of Language*, 4, 225-242.

Halliday, M.A.K. & Matthiessen, C. (2004). *An introduction to functional grammar* (3rd ed.). Oxford University Press.

Hammarfelt, B. (2019). Discipline. In B. Hjørland & C. Gnoli (Eds.), *ISKO Encyclopedia of Knowledge Organization* (pp. 1-22). Retrieved March 30, 2023, from

https://www.diva-portal.org/smash/record.jsf?pid=diva2%3A1361084&dswid=9095

Han, Z. (2004). *Fossilization in adult second language acquisition*. Multilingual Matters.

Hasselgren, A. (1994). Lexical teddy bears and advanced learners: A study into the ways Norwegian students cope with English vocabulary. *International journal of applied linguistics*, *4*(2), 237-258.

Hermerén, L. (1978). *On Modality in English*. Gleerup.

Hilpert, M. (2016). Change in modal meanings: Another look at the shifting collocates of may. *Constructions and Frames*, *8*(1), 66-85.

Hilpert, M., & Flach, S. (2021). Disentangling modal meanings with distributional semantics. *Digital Scholarship in the Humanities*, *36*(2), 307-321.

Hilpert, M., & Perek, F. (2015). Meaning change in a petri dish: Constructions, semantic vector spaces, and motion charts. *Linguistics Vanguard*, 1(1), 339-350.

Hinkel, E. (1995). The use of modal verbs as a reflection of cultural values. *TESOL quarterly*, *29*(2), 325-343.

Ho, D. (2018). *Notepad++* (Version 7.5.6) [Computer software]. https://notepad-plus-plus.org.

Hsieh, C. L. (2005). Modal verbs and modal adverbs in Chinese: An investigation into the semantic source. *UST Working Papers in Linguistics*, *1*(1), 31-58.

Hu, C., & Li, X. (2015). Epistemic Modality in the Argumentative Essays of Chinese EFL Learners. *English Language Teaching*, *8*(6), 20-31.

Huang, X. Y. K. (2009). Multiple-modal constructions in Mandarin Chinese: A cartographic approach and an MP perspective. In *Proceedings of the 21 st North American Conference on Chinese Linguistics (NACCL-21)* (Vol. 2, pp. 524-540).

Huddleston, R. (1971). *The sentence in written English: a syntactic study based on an analysis of scientific texts*. Cambridge University Press.

Huddleston, R., Pullum, G. K. (2002). *The Cambridge Grammar of the English Language*. Cambridge University Press.

Hyland, K. (2002a). Authority and invisibility: Authorial identity in academic writing. *Journal of pragmatics*, 34(8), 1091-1112.

Hyland, K. (2002b). Directives: Argument and engagement in academic writing. *Applied linguistics*, 23(2), 215-239.

Hyland, K. (2004). *Disciplinary discourses, Michigan classics ed.: Social interactions in academic writing*. University of Michigan Press.

Hyland, K. (2005). Stance and engagement: A model of interaction in academic discourse. *Discourse studies*, *7*(2), 173-192.

Hyland, K. (2018). *Metadiscourse: Exploring interaction in writing*. Bloomsbury Publishing.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of second language writing*, *6*(2), 183-205.

Hyon, S. (1996). Genre in three traditions: Implications for ESL. *TESOL quarterly*, *30*(4), 693-722.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. *Learner corpus studies in Asia and the world, 1*, 91-118.

Jenkins, L. (1972). *Modality in English syntax* (Doctoral thesis, Massachusetts Institute of Technology)

Johansson, S., Leech, G., & Goodluck, H. (1978). *Manual of Information to Accompany the Lancaster-Oslo/Bergen Corpus of British English, for Use with Digital Computers*. Department of English, University of Oslo.

Kecskes, I., & Kirner-Ludwig, M. (2017). "It would never happen in my country I *must* say": A corpus-pragmatic study on Asian English learners' preferred uses of *must* and *should*. *Corpus Pragmatics*, *1*, 91-134.

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

Koutsantoni, D. (2004). Attitude, certainty and allusions to common knowledge in scientific research articles. *Journal of English for Academic Purposes*, *3*(2), 163-182.

Krashen, S. (1982). *Principles and practice in second language acquisition*. Pergamon Press.

Krippendorff, K. (2012 [1980]). *Content analysis: An introduction to its methodology*. Sage.

Larsson, T., Paquot, M., & Plonsky, L. (2020). Inter-rater reliability in learner corpus research: Insights from a collaborative study on adverb placement. *International Journal of Learner Corpus Research*, *6*(2), 237-251.

Lee, D. Y. (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.

Lee, S. H. S. (2010). Command strategies for balancing respect and authority in undergraduate expository essays. *Journal of English for Academic Purposes*, 9(1), 61-75.

Levshina, N., & Heylen, K. (2014). A radically data-driven Construction Grammar: Experiments with Dutch causative constructions. In R. Boogaart, T. Colleman, & G. Rutten (Eds.), *Extending the scope of construction grammar* (pp. 17-54). De Gruyter Mouton.

Lew, R., Frankenberg-Garcia, A., Rees, G., Roberts, J. C., & Sharma, N. (2018, July). ColloCaid: A real-time tool to help academic writers with English collocations. In *Proceedings of the XVIII EURALEX International Congress*.

Li, C. N., & Thompson, S. A. (1981). *Mandarin Chinese: A functional reference grammar*. University of California Press.

Li, R. (2004). *Modality in English and Chinese: A typological perspective*. Universal-Publishers.

Li, W. (2016). The cultural ID in the modal system: A contrastive study of English abstracts written by Chinese and native speakers: Can modality differences be an important indicator of the China English variety?. *English Today*, *32*(4), 6-11.

Li, X. (2020). *Modal Verb Treatment in EFL Textbooks: A Quantitative and Qualitative Approach* (Doctoral thesis, The Chinese University of Hong Kong).

Liang, M. C. (2008). Zhongguodaxuesheng yingyubiyuzhongde qingtaixulieyanjiu (A Corpus-based Study of Modal Sequences in Chinese Tertiary EFL Learners' Written Production). *Waiyujiaoxueyuyanjiu: waiguoyuwenshuangyuekan* (Foreign Language Teaching and Research: Bimonthly), 40(1), 51-58.

Lin, T. H. J. (2012). Multiple-modal constructions in Mandarin Chinese and their finiteness properties1. *Journal of Linguistics*, *48*(1), 151-186.

Liu, N., Feng, Z., & Wang, Q. (Eds.). (2024). *Education in China and the World: Achievements and Contemporary Issues*. Springer Nature.

Liu, Y., Pan, W., & Gu, W. (1983). *Practical Modern Chinese Grammar*. Foreign Language Teaching & Research Press.

Long, Z. Y. (2013). Yingyuzhuanyexuesheng yilunwenzhong qingtaixulie shiyongpinlyude nianjibiaozheng (A Study of the Developmental Features of Modal Sequences in Chinese English Majors' Argumentative Writing). *Zhongguo waiyujiaoyu: jikan* (Foreign Language Education in China: Quarterly), (1), 3-14.

Lyons, J. (1977). *Semantics (Vol. 2)*. Cambridge University Press.

Lyons, J. (1995). *Linguistic Semantics: An introduction*. Cambridge University Press.

Ma, G. and Lu, X. J. (2007). Jiyuzhongguoxuexizhe yingyuyuliaokude qingtaidongciyanjiu (the Analysis on Modal Verbs——based on Chinese Learner English Corpus ST 6). *Waiyudianhuajiaoxue* (CAFLE), (3), 17-21.

Ma, J. (1898). *Ma's Grammar. Shanghai: Commercial Press*. Commercial Press.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge university press.

Merriam, S. B., & Tisdell, E. J. (2015). *Qualitative research: A guide to design and implementation*. John Wiley & Sons.

Michigan Corpus of Upper-level Student Papers. (2009). *Ann Arbor*, The Regents of the University of Michigan.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. Retrieved from https://arxiv.org/abs/1301.3781.

Millán, E. L. (2008). Epistemic and Approximative Meaning Revisited: The use of hedges boosters and approximators when writing research in different disciplines. *English as an additional language in research publication and communication*, 61, 65.

Ministry of Education of China. (2018). *China's standards of English language ability*. https://cse.neea.edu.cn/html1/folder/1505/249-1.htm

Nesi, H. (2011). BAWE: An introduction to a new resource. In A. Frankenberg-Garcia, L. Flowerdew & G. Aston (Eds.), *New Trends in Corpora and Language Learning* (pp. 213-228). Continuum.

Nesi, H., & Gardner, S. (2006). Variation in disciplinary culture: University tutors' views on assessed writing tasks. *British studies in applied linguistics*, *21*, 99.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge University Press.

Nesi, H., & Gardner, S. (2018). The BAWE corpus and genre families classification of assessed student writing. *Assessing Writing*, *38*, 51-55.

Nesi, H., Gardner, S., Thompson, P., & Wickens, P. (2008). *British Academic Written English Corpus*. Oxford Text Archive. http://hdl.handle.net/20.500.12024/2539

Nuyts, J. (2001). Subjectivity as an evidential dimension in epistemic modal expressions. *Journal of Pragmatics*, 33, 383-400.

Oktavianti, I. N. (2019). Necessity and Obligation Modals in English Academic Discourse: A Corpus-Based Analysis. *IJELTAL (Indonesian Journal of English Language Teaching and Applied Linguistics)*, *4*(1), 47-59.

Palmer, F. (1990 [1979]). *Modality and the English modals*, 2nd edition. Longman.

Palmer, F. (2001). *Mood and Modality*, second ed. Cambridge University Press.

Panocová, R. & Lukáš L. (2019). Epistemic Modal Markers in Two Domains of Academic Research Papers in English. *Brno Studies in English*. 45 (2): 121–38.

Papafragou, A. (1998). The acquisition of modality: Implications for theories of semantic representation. *Mind & language*, *13*(3), 370-399.

Paquot, M. (2007). Towards a productively-oriented academic word list. In J. Walinski, K. Kredens, & S. Gozdz-Roszkowski (Eds.), *Corpora and ICT in language studies. PALC 200*5 [Lodz Studies in Language 13], (pp. 127–140). Frankfurt am Main, Peter Lang.

Parkinson, J. (2020). Stance and modals of obligation and necessity in academic writing. *Register Studies*, 2(1), 102-130.

Parkinson, J. (2022). A comparison of use of modal auxiliaries of obligation and necessity in science writing by ESL and L1 students. In *Multifunctionality in English* (pp. 117-133). Routledge.

Peacock, M. (2014). Modals in the construction of research articles. *Ibérica*, (27), 143-164.

Perek, F. (2014). Rethinking constructional polysemy: The case of the English conative construction. In D. Glynn & J. Robinson (Eds.), *Polysemy and Synonymy: Corpus Methods and Applications in Cognitive Linguistics* (pp. 61-85). John Benjamins Publishing Company.

Perek, F. (2015). *Argument structure in usage-based construction grammar*. John Benjamins Publishing Company.

Perek, F. (2016). Using distributional semantics to study syntactic productivity in diachrony: A case study. *Linguistics*, 54(1), 149-188.

Perek, F. (2018). Recent change in the productivity and schematicity of the way-construction: A distributional semantic analysis. *Corpus Linguistics and Linguistic Theory*, 14(1), 65-97.

Perek, F. (2021). *Distributional semantic models for English verbs and nouns*. OSF. https://doi.org/10.17605/OSF.IO/N324F.

Perkins, M.R. (1983). *Modal Expressions in English*. Frances Pinter.

Peters, A., & Bembridge, G. (2016). Structural Problems of Multiple Modal Constructions: Views from Mandarin, Rural Chesapeake English and Jamaican Creole. In *Proceedings of the 2016 annual conference of the Canadian Linguistic Association* (pp. 1-14).

Pienemann, M. (1989). Is language teachable? Psycholinguistic experiments and hypotheses. *Applied Linguistics*, 10(1), 52-79.

Piqué-Angordans, J., Posteguillo, S., & Andreu-Besó, J. V. (2001). A pragmatic analysis framework for the description of modality usage in academic English contexts. *ELIA*, 2, 213-224.

Piqué-Angordans, J., Posteguillo, S., & Andreu-Besó, J. V. (2002). Epistemic and deontic modality: a linguistic indicator of disciplinary variation in academic English. *LSP and professional communication (2001-2008)*, 2(2).

Princeton University. (2010). *WordNet*. Retrieved from http://wordnet.princeton.edu.

Quirk, R., Greenbaum, S., Leech, G., Svartvik, J. (1985). *A Comprehensive Grammar of the English Language*. Longman.

Rizomilioti, V. (2006). Exploring epistemic modality in academic discourse using corpora. In *Information Technology in languages for specific purposes: Issues and prospects* (pp. 53-71). Springer US.

Rozumko. A. (2017). Adverbial Markers of epistemic modality across disciplinary discourses: a contrastive study of research articles in six academic disciplines. *Studia Anglica Posnaniensia*. 52(1): 73–101.

Sameri, M., & Tavangar, M. (2013). Epistemic Modality in Academic Discourse: A Cross-Linguistic and Cross. *The Iranian EFL Journal*, 15(1), 127-147.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Springer.

Selinker, L. (1972). INTERLANGUAGE. *International Review of Applied Linguistics in Language Teaching*, *10*(1-4), 209-232. https://doi.org/10.1515/iral.1972.10.1-4.209.

Selinker, L. (1989). CA/EA/IL: The earliest experimental record. *IRAL: International Review of Applied Linguistics in Language Teaching*, 27(4), 267.

Selinker, L. (2014). Interlanguage 40 years on: Three themes from here. In Z. Han & E. Tarone (Eds.), *Interlanguage* (pp. 221-246). John Benjamins.

Simpson, P. (1990). Modality in literary-critical discourse. In W. Nash (Ed.), *The writing scholar: Studies in academic discourse* (pp. 63-94). Sage.

Sinclair, J. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. M. (1988). Mirror for a text. *Journal of English and Foreign Languages*, 1, 15-44.

Squires, G. (2005). Art, science and the professions. *Studies in Higher Education*, *30*(2), 127-136.

Sun, Q. (2018). *A Corpus Based Study: Modals in the English Textbooks in China* (Unpublished master's dissertation, University of Birmingham).

Takimoto, M. (2015). A corpus-based analysis of hedges and boosters in English academic articles. *Indonesian Journal of Applied Linguistics*, 5(1), 95-105.

Tang, L. L. (2013). Zhongguoyingyuxuexizhe qingtaidongci yuyifazhande yuliaokukaocha (A Corpus-based Study of the Acquisition of Modal Verbs'

Semantic Meanings by Chinese English- major Learners). *Dangdaiwaiyuyanjiu* (Contemporary Foreign Languages Studies), (6), 32-36.

Tang, T, & Tang, Z. (1997). Huayu qingtaici xulun (Introduction to Chinese modal expressions). In World Chinese Education Association (ed.), *Diwujie Shijie Huayuwen Jiaoxue Yantaohui Lunwenji: Yuwen Fenxi* (Proceedings of the Fifth World Chinese Teaching Conference: Linguistic Analysis). pp.175-197. World Chinese Publishing.

The British National Corpus, version 2 (BNC World). (2001). Distributed by Oxford University Computing Services on behalf of the BNC Consortium. Available from http://www.natcorp.ox.ac.uk/.

Thompson, P. (2001). *A pedagogically-motivated corpus-based examination of PhD theses: Macrostructure, citation practices and uses of modal verbs* (Unpublished doctoral thesis, University of Reading).

Tiee, H. (1985). Modality in Chinese. In Nam-Kil Kim, and Henry Hung-Yeh Tiee (Eds.), *Studies in East Asian Linguistics*. pp.84-96. Department of East Asian Languages and Cultures, University of Southern California.

Treebank of Learner English (TLE). Accessed via: https://cbmm.mit.edu/publications/treebank-learner-english-tle

Tsang, C. L. (1981). *A semantic study of modal auxiliary verbs in Chinese*. (Doctoral thesis, Stanford University).

Turner, S. (2000). What are disciplines? And how is interdisciplinarity different. *Practising interdisciplinarity*, 46-65.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of artificial intelligence research*, 37, 141-188.

University College London. (n.d.). *Survey of English Usage*. Retrieved from https://www.ucl.ac.uk/english-usage/about/index.htm.

Van der Auwera, J., & Plungian, V. A. (1998). *Modality's semantic map*. Linguistic Typology. 2, 1.79–124.

Van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-SNE. *Journal of machine learning research*, 9(11).

Vazquez, I., & Giner, D. (2008). Beyond mood and modality: Epistemic modality markers as hedges in research articles: A cross disciplinary study. *Revista Alicantina de Estudios Ingleses*, 21, 171-190.

Vičič, P., & Petek, K. J. (2016). The role of modal verbs in research papers in the field of logistics. *Scripta Manent*, 11(1), 21-41.

Vincent, B. (2020). The expression of obligation in student academic writing. *Journal of English for Academic Purposes*, *44*, 100840.

Vincent, B. (2015). *Modality and the V wh pattern* (Doctoral thesis, University of Birmingham).

Wang, M., Malmasi, S., & Huang, M. (2015, June). The jinan chinese learner corpus. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 118-123).

Warchał, K. (2007). Necessity and obligation in written academic discourse: *MUST* in research articles. *PASE Papers*, 1, 226-247.

Warchał, K. (2008). Claiming authority: Modals of obligation and necessity in academic written English. The case of *SHOULD*. *Linguistica e Filologia*, 27, 21-37.

Warchał, K. (2010). Taking stance across languages: High-value modal verbs of epistemic necessity and inference in English and Polish linguistics research articles. *Linguistica Silesiana*, 31, 123-136.

Wei, M. (2016). Language ideology and identity seeking: Perceptions of college learners of English in China. *Journal of Language, Identity & Education*, 15(2), 100-113.

Wen, Q. F., Wang, L. F., & Liang, M. C. (2009). *Spoken and written English corpus of Chinese learners (SWECCL)*. Foreign Language Teaching and Research Press.

Whitley, R. (2000). *The intellectual and social organization of the sciences*. Oxford University Press.

Wray, A. (2012). What do we (think we) know about formulaic language? An evaluation of the current state of play. *Annual review of applied linguistics*, *32*, 231-254.

Xiao, Y. (2017). Chinese EFL learners' acquisition of modal verbs: A corpus-based study. *International Journal of English Linguistics*, *7*(6), 164-170.

Yang, H. Z., Gui, S. C. & Yang, D. F. (2005). *Jiyu CLEC yuliaokude zhongguoxuexizhe yingyufenxi* (Corpus-based Analysis of Chinese Learner English). Shanghai Foreign Language Education Press.

Yang, X. (2018). A Corpus-Based Study of Modal Verbs in Chinese Learners' Academic Writing. *English Language Teaching*, 11(2), 122-130.

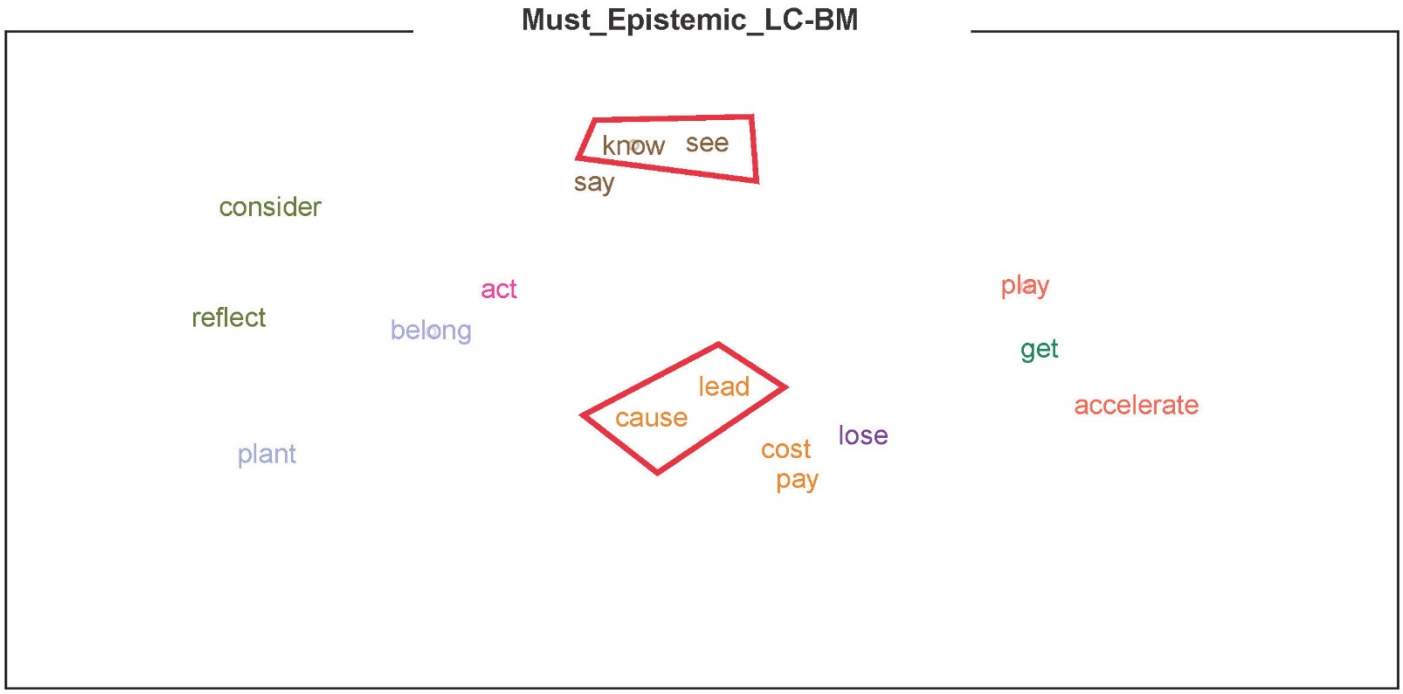
Yang, A., Zheng, S. Y., & Ge, G. C. (2015). Epistemic modality in English-medium medical research articles: A systemic functional perspective. *English for Specific Purposes*, 38, 1-10.

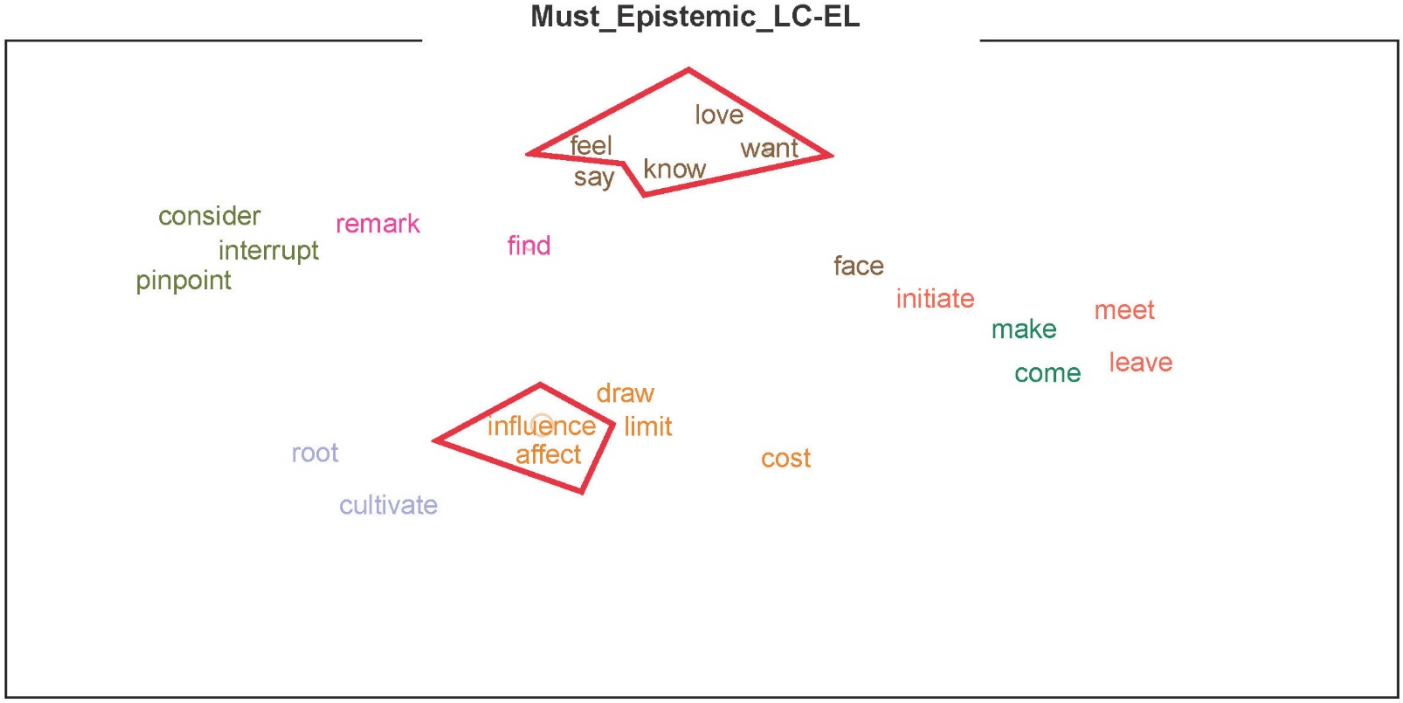
Ye, X. (2021). EFL Learning motivation differences of Chinese junior secondary school students: A mixed-methods study. *Education 3-13*, 49(2), 203-216.

Zou, Y. (2018). *First person pronouns in academic discourse by novice writers in China* (Doctoral thesis, University of Birmingham).

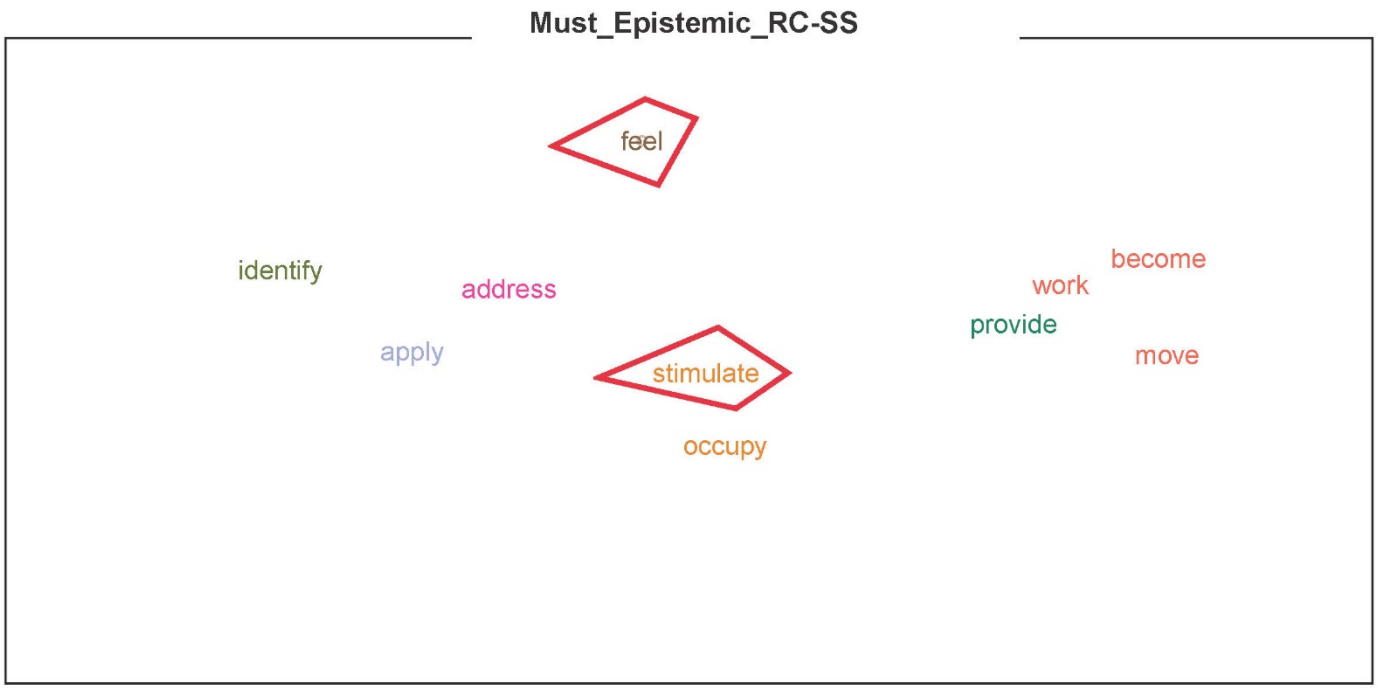
Appendices

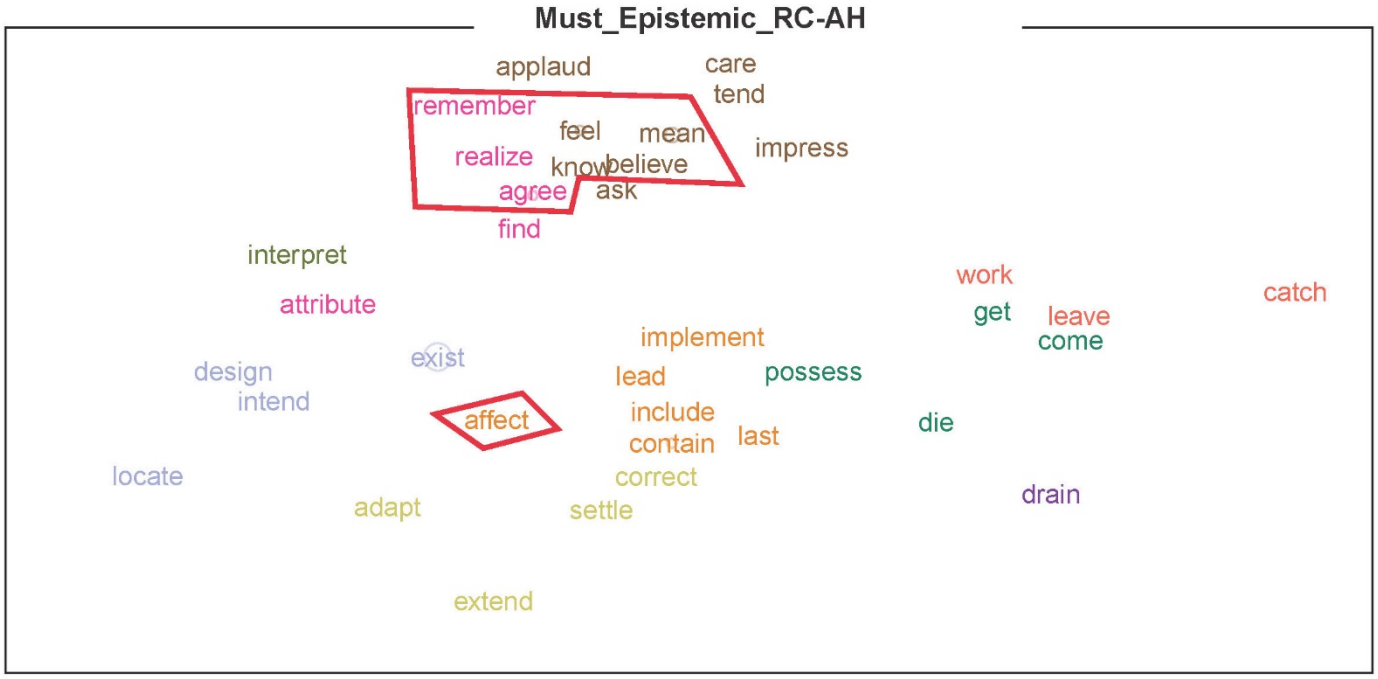
Appendix A: Distributional semantic plots of the verb collocates of epistemic *must* in the learner corpus



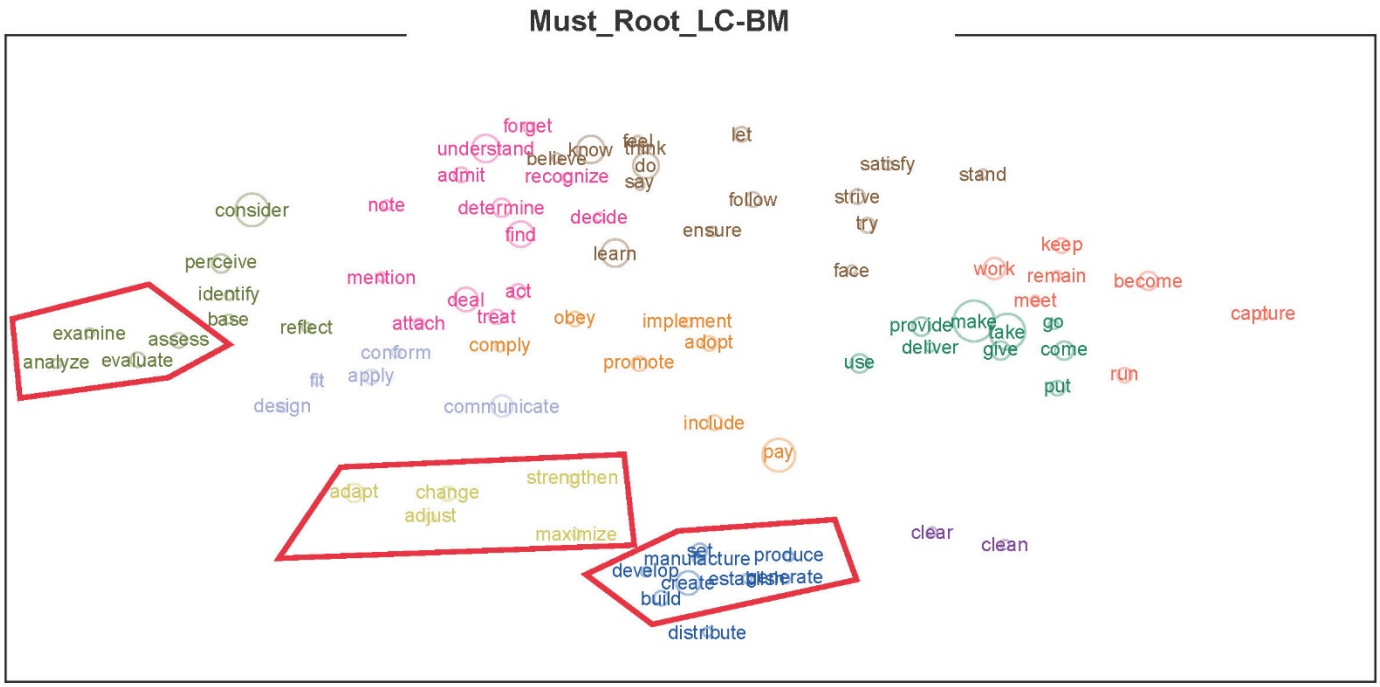


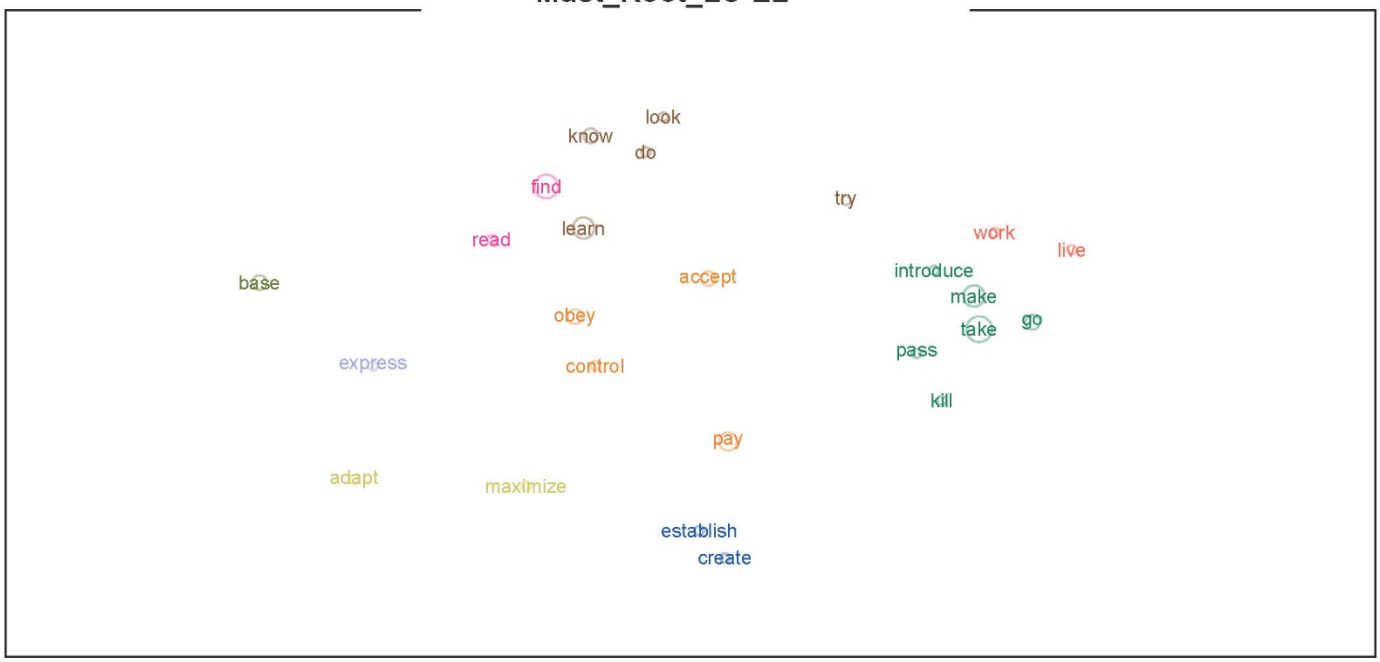
Appendix B: Distributional semantic plots of the verb collocates of epistemic *must* in the reference corpus



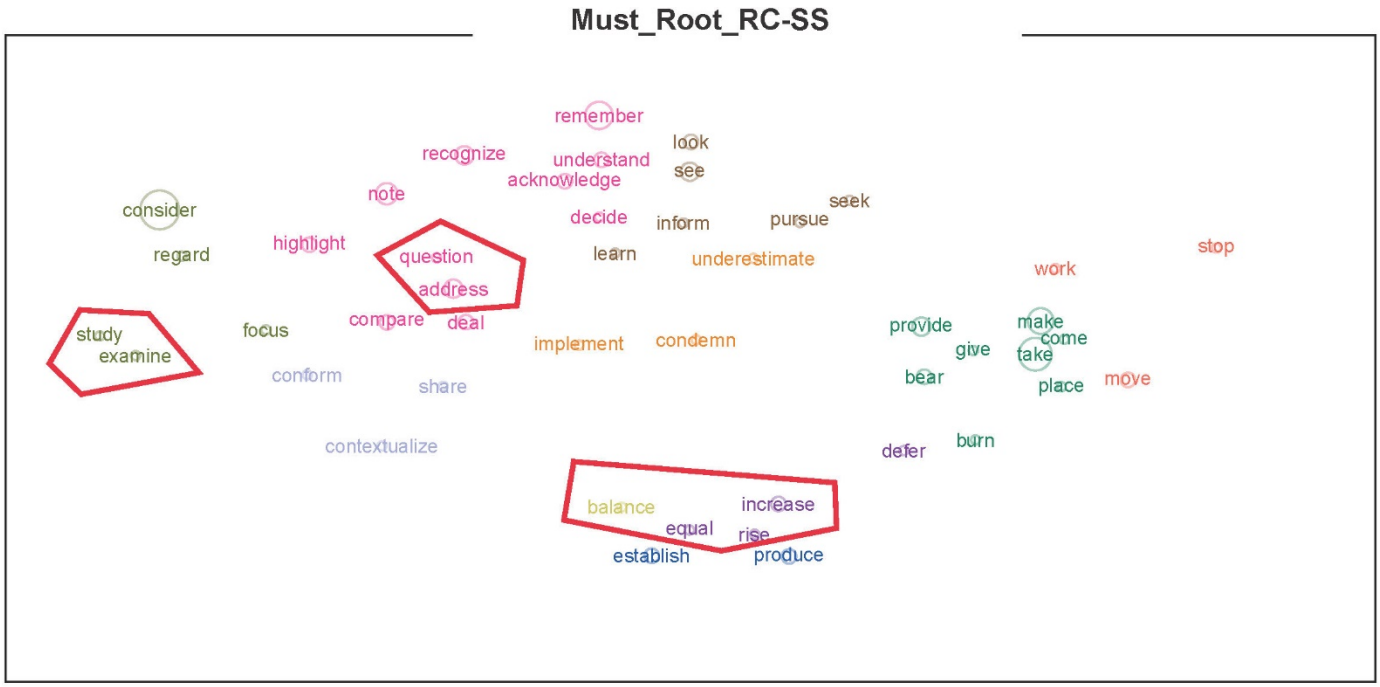


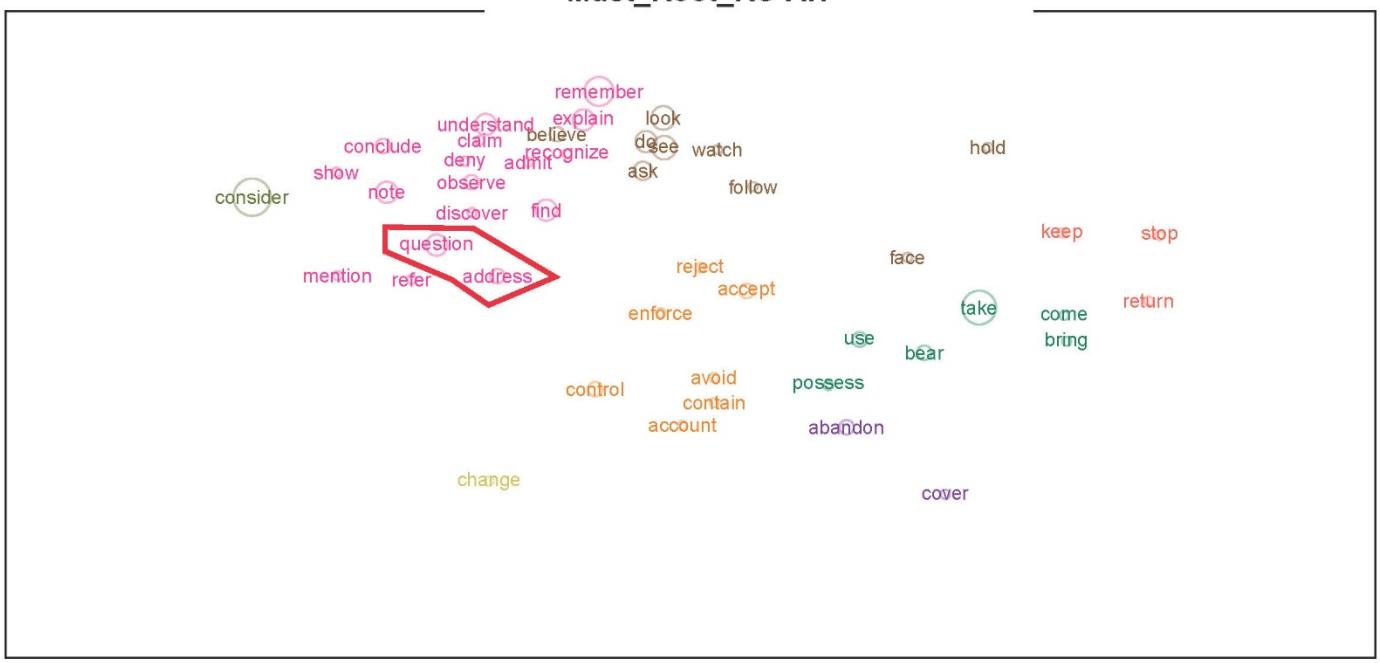
Appendix C: Distributional semantic plots of the verb collocates of root *must* in the learner corpus



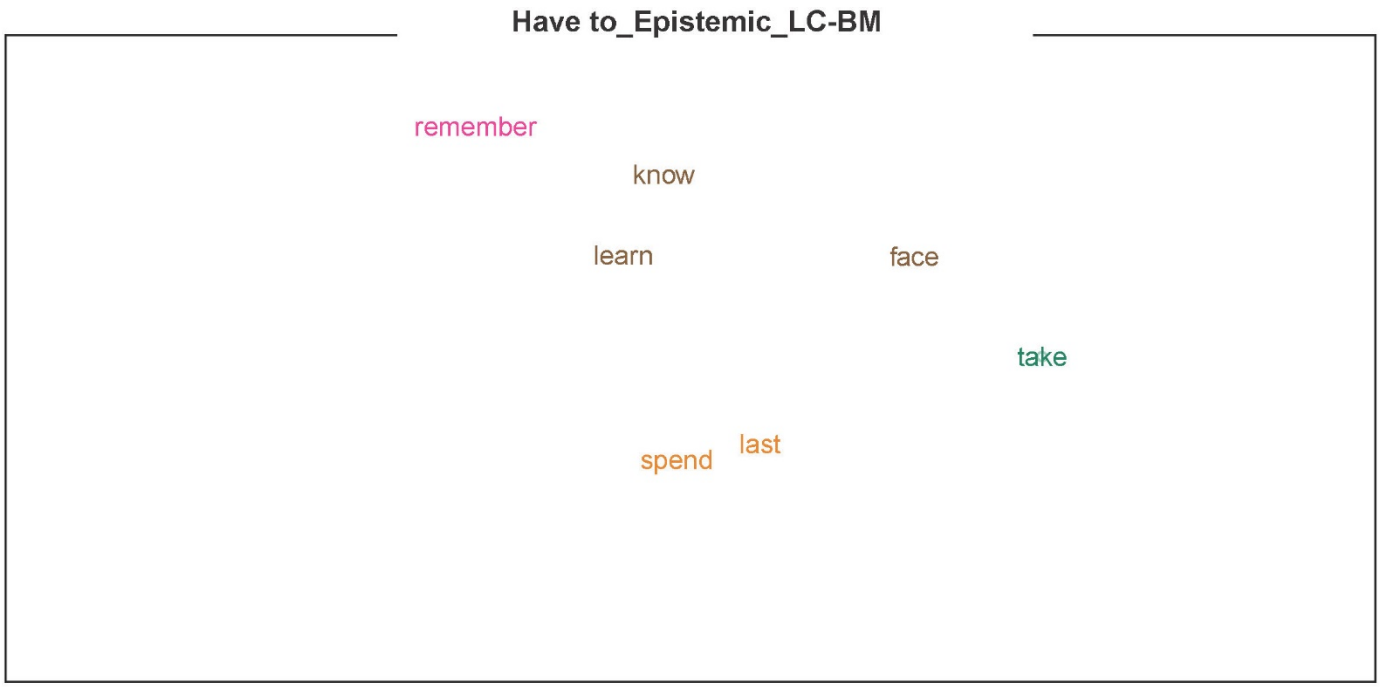


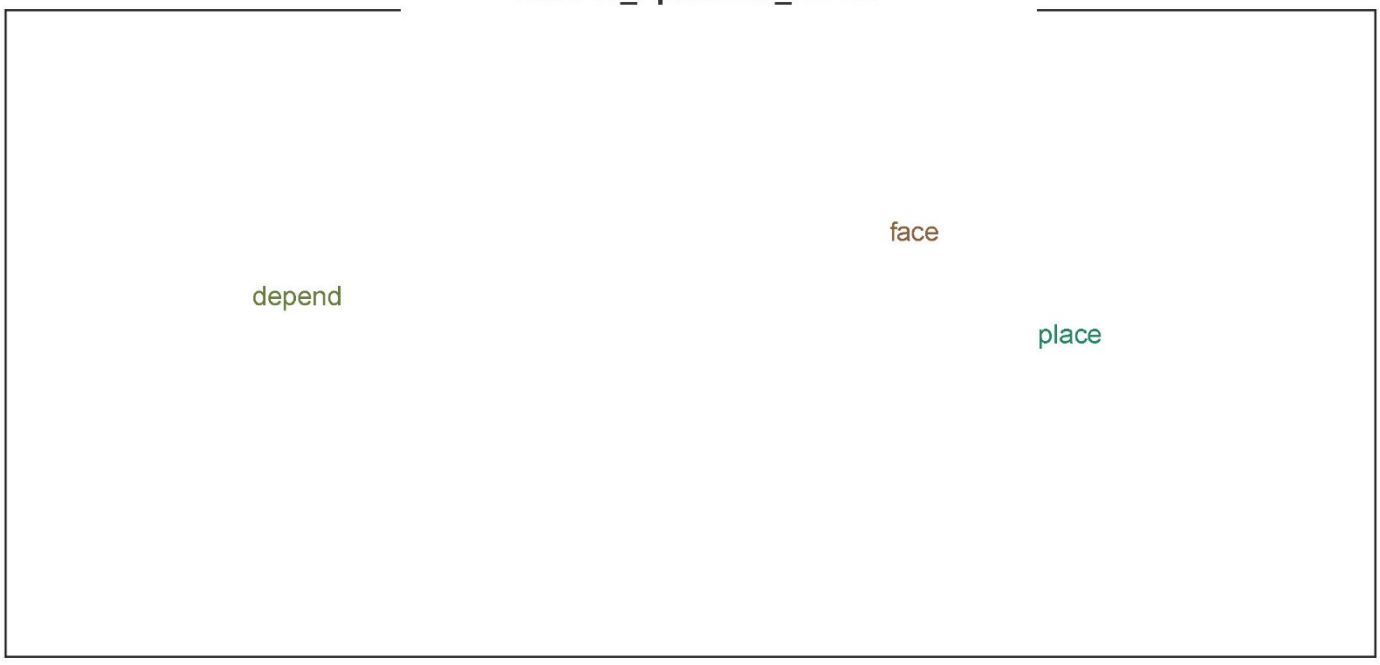
Appendix D: Distributional semantic plots of the verb collocates of root *must* in the reference corpus



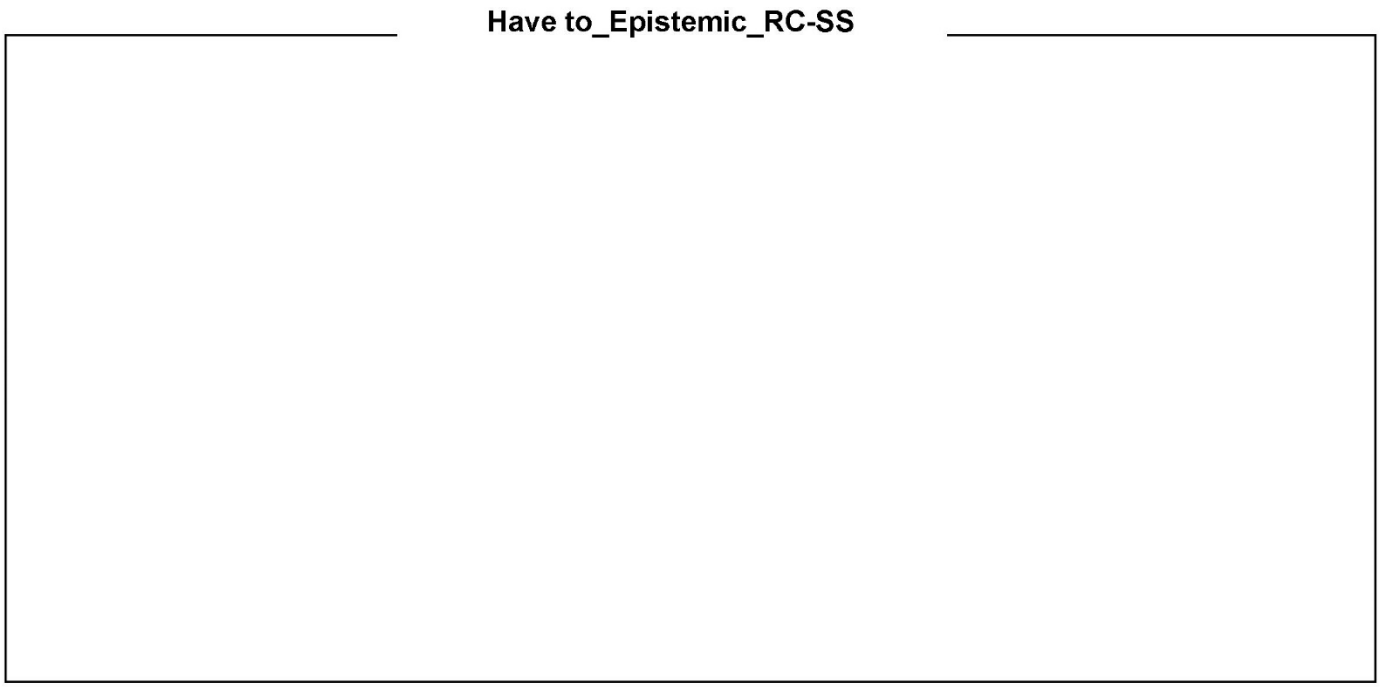


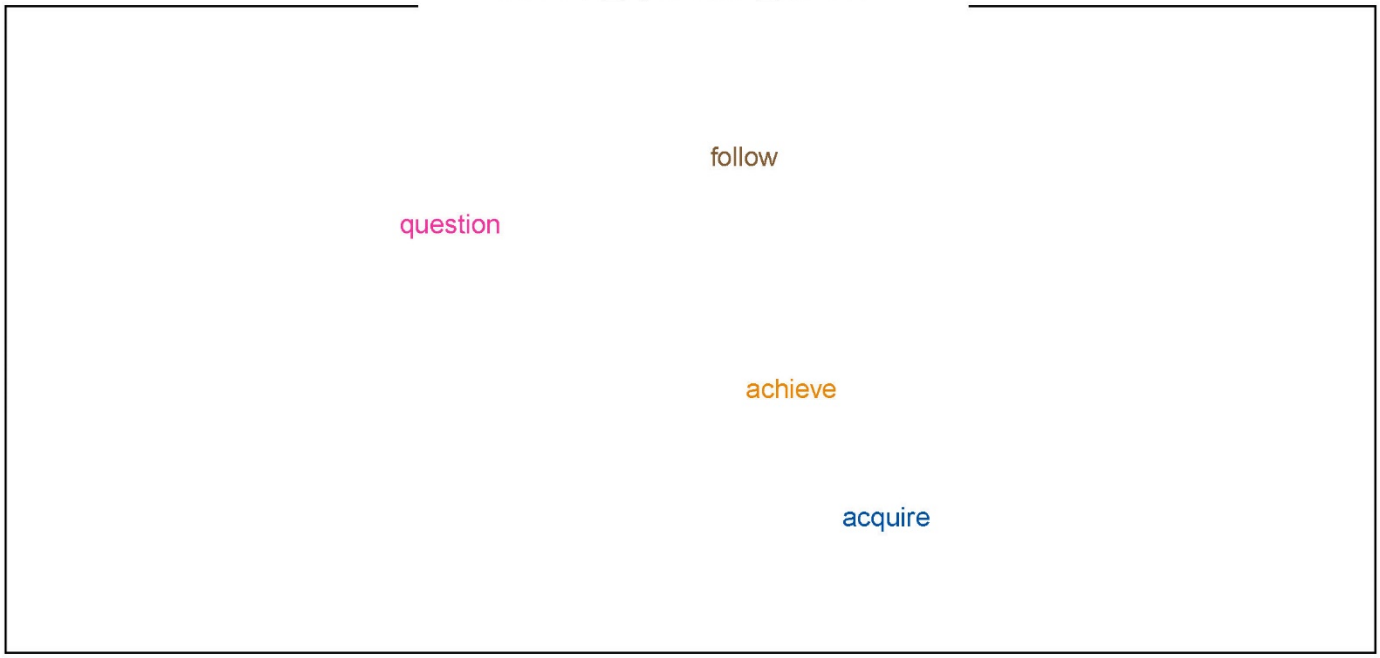
Appendix E: Distributional semantic plots of the verb collocates of epistemic *have to* in the learner corpus



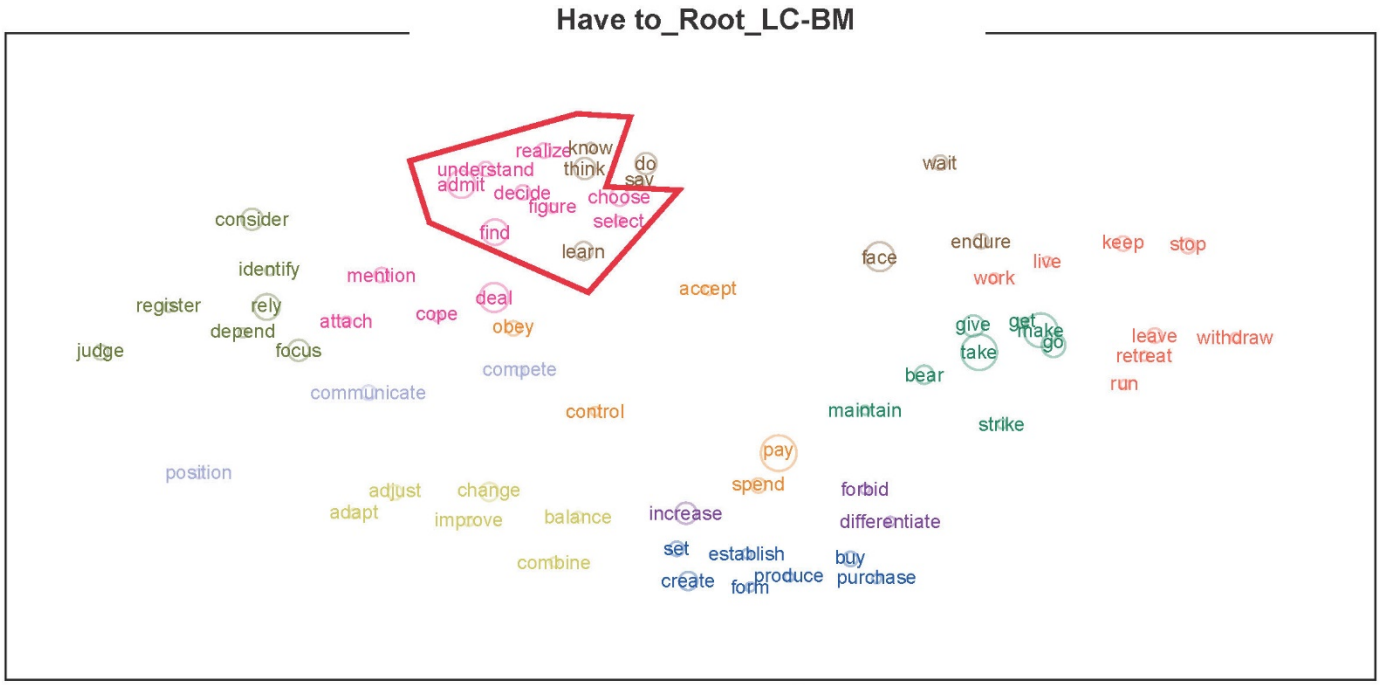


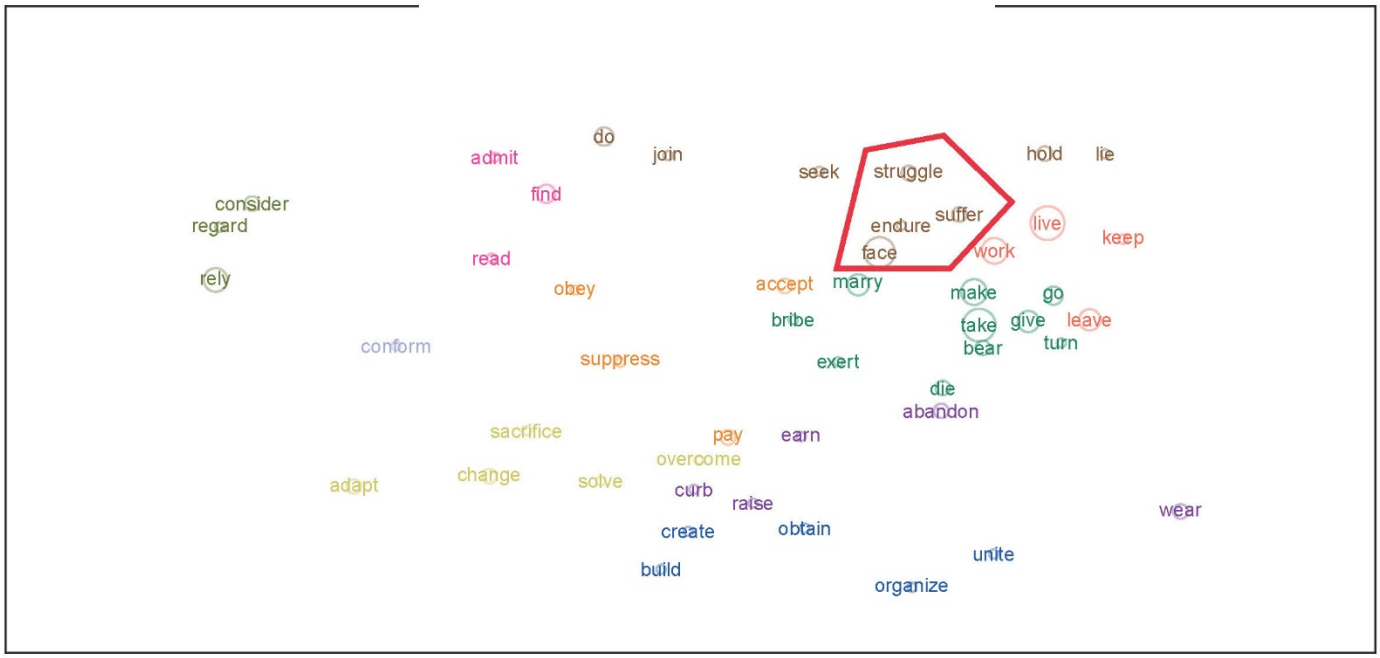
Appendix F: Distributional semantic plots of the verb collocates of epistemic *have to* in the reference corpus



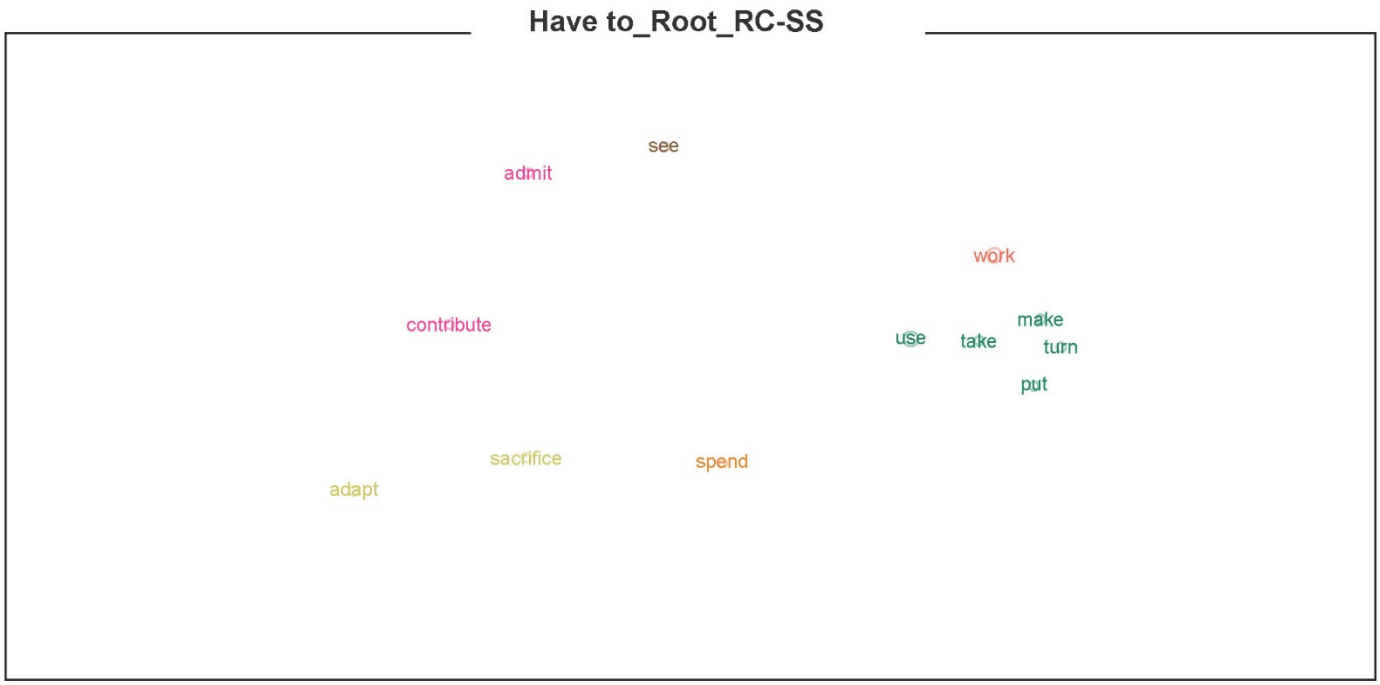


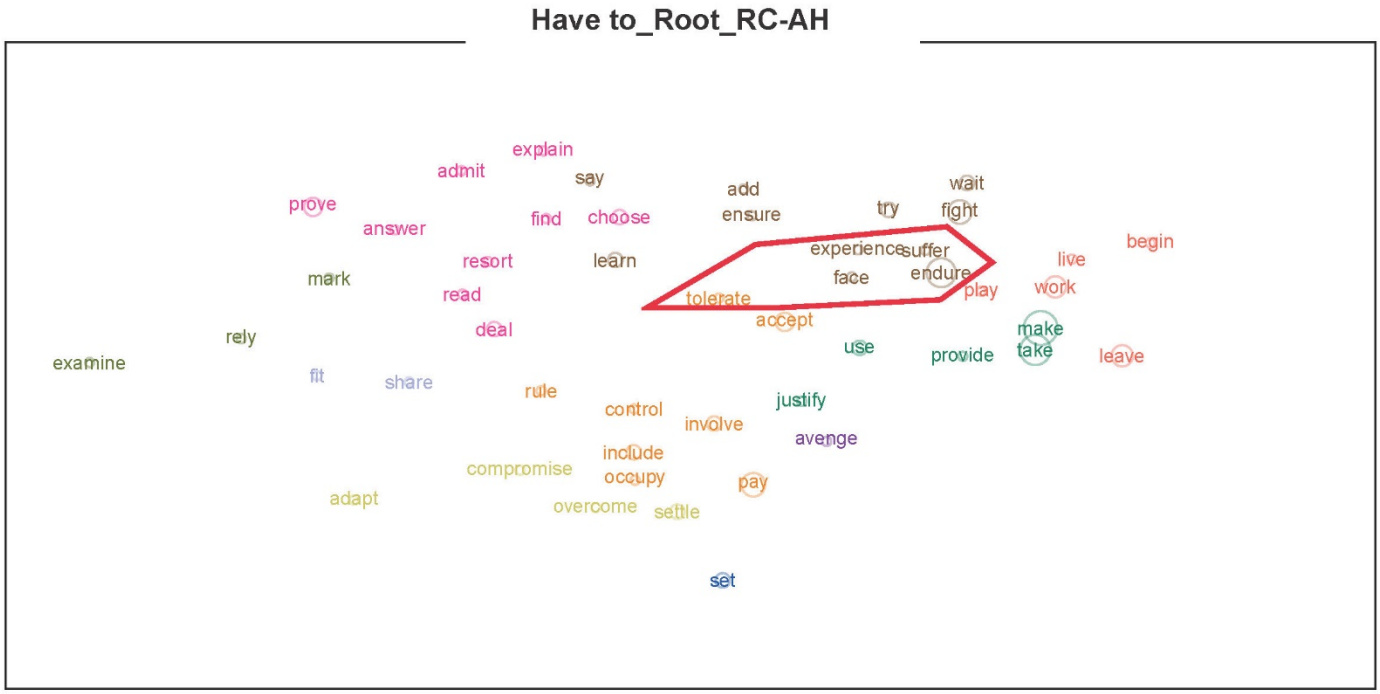
Appendix G: Distributional semantic plots of the verb collocates of root *have to* in the learner corpus



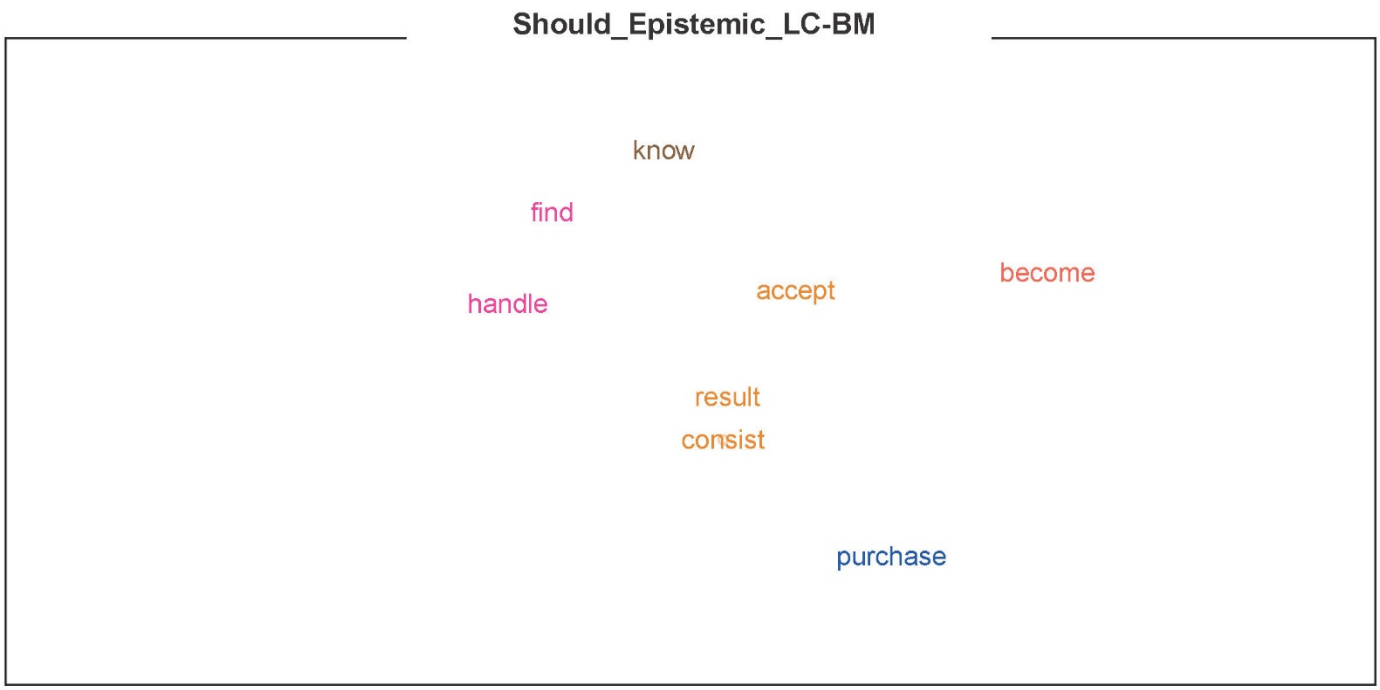


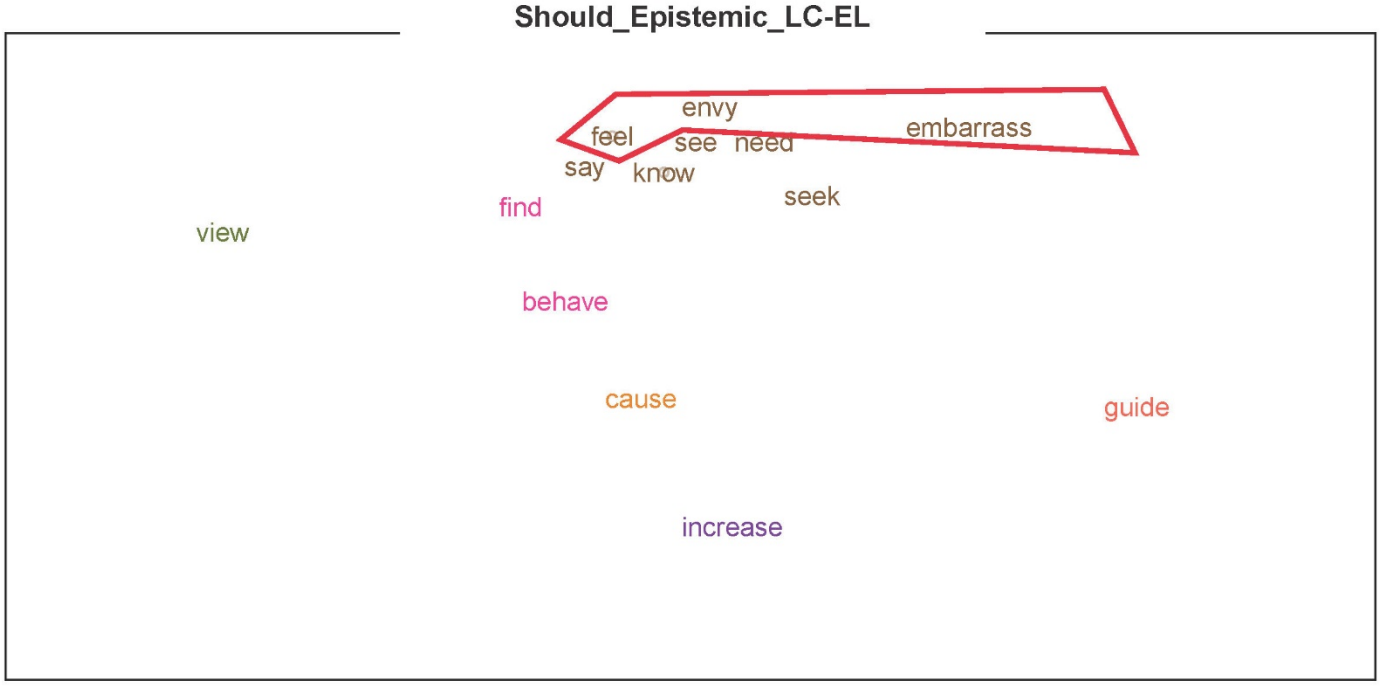
Appendix H: Distributional semantic plots of the verb collocates of root *have to* in the reference corpus



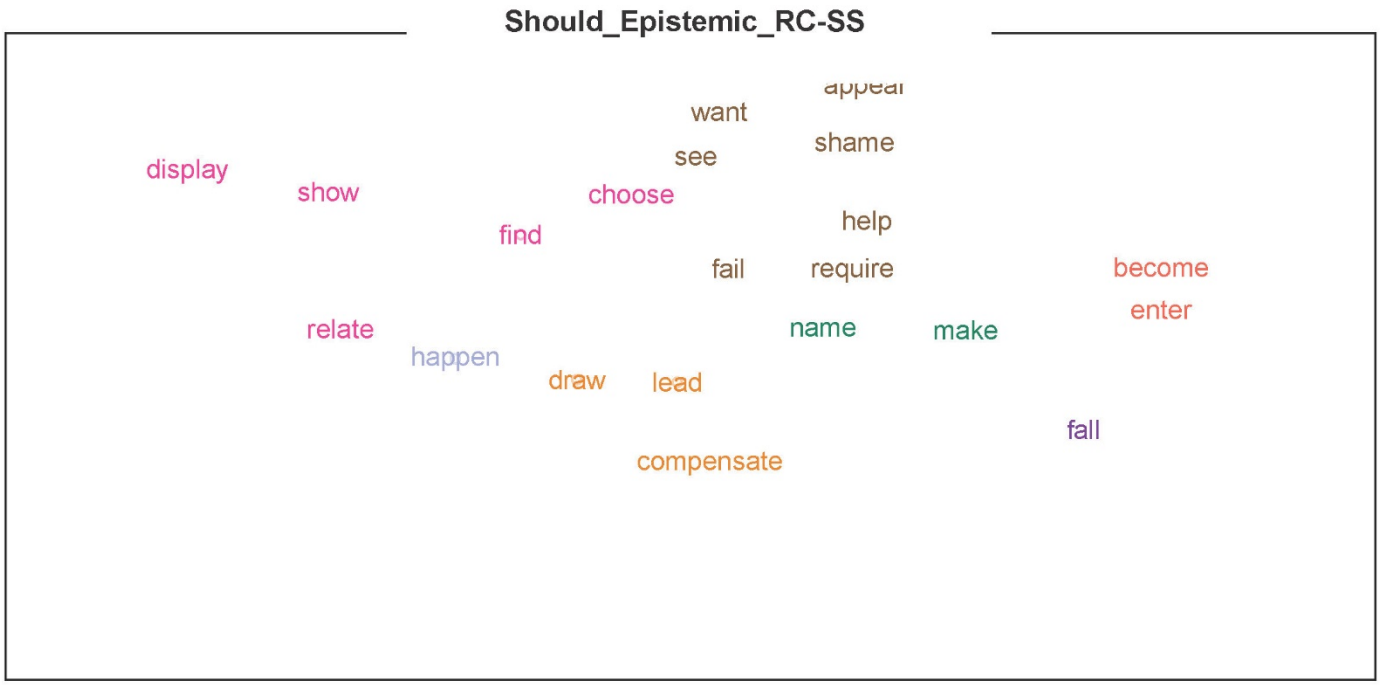


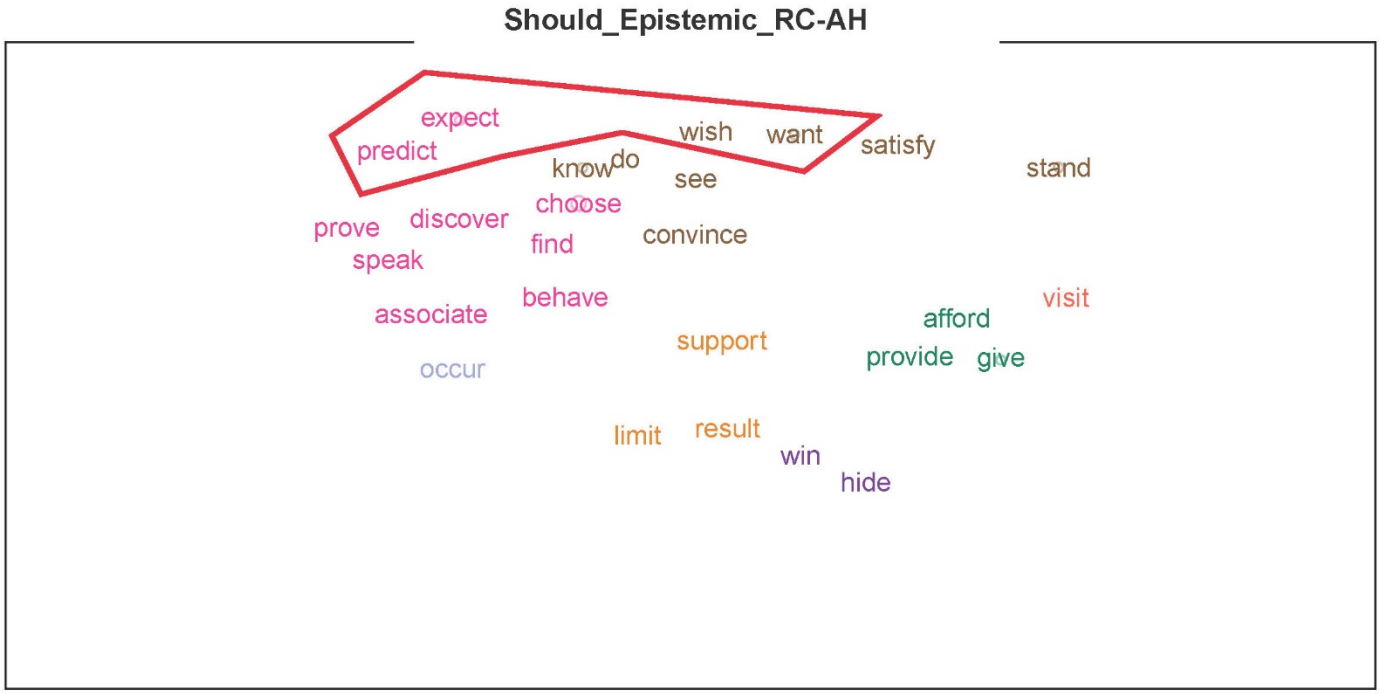
Appendix I: Distributional semantic plots of the verb collocates of epistemic *should* in the learner corpus



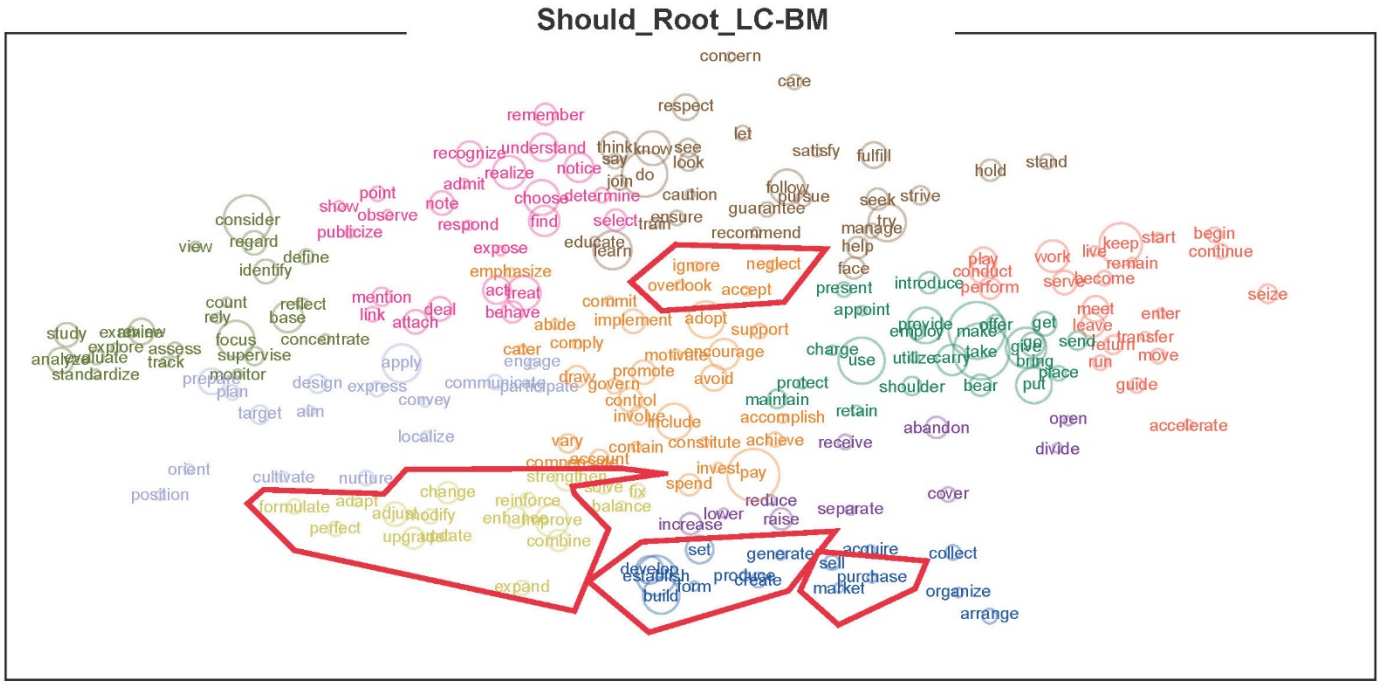


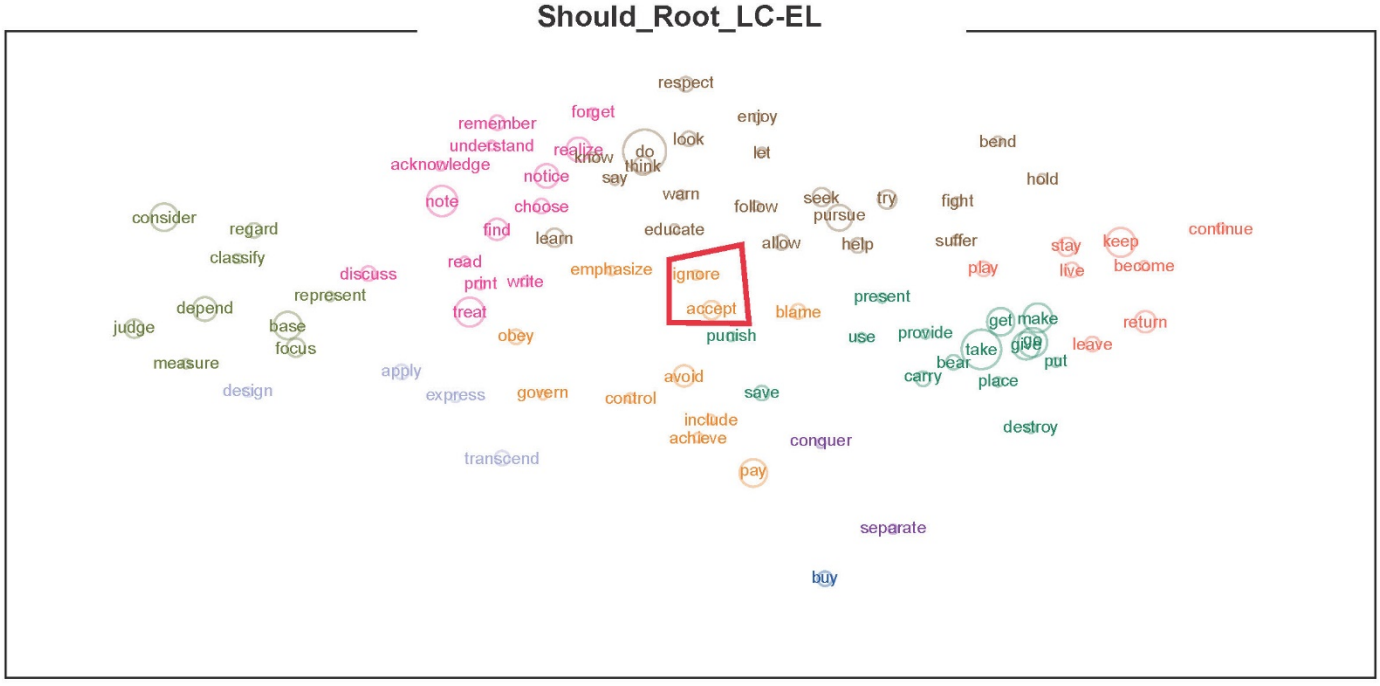
Appendix J: Distributional semantic plots of the verb collocates of epistemic *should* in the reference corpus





Appendix K: Distributional semantic plots of the verb collocates of root *should* in the learner corpus





Appendix L: Distributional semantic plots of the verb collocates of root *should* in the reference corpus