Sensitivity to Differences between Speech and Writing: Hong Kong Students' Use of Syntactic Features in English

CHUI, Sze Yan

A Thesis Submitted in Partial Fulfilment

of the Requirements for the Degree of

Doctor of Philosophy

in

Applied English Linguistics

The Chinese University of Hong Kong
December 2010

ProQuest LLC.
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106 - 1346

Thesis/Assessment Committee

Professor George Braine (Chair)

Professor Gerald Nelson (Thesis Supervisor)

Professor Carmen Lee (Committee Member)

Professor Andy Kirkpatrick (External Examiner)

ABSTRACT

This thesis investigates Hong Kong secondary school students' sensitivity to differences between spoken and written English by examining their use of 67 syntactic features. A model specifying how native speakers vary their use of syntactic features across speech and writing, Biber (1988), has been adopted as the theoretical framework. Fifty-two oral presentations delivered by Form 6 students and 52 public examination essays written by Form 7 students, both of which total about 10,000 words, have been analysed. The students' performance is compared with native speakers' performance in similar spoken and written production on the level of syntactic features and the level of textual dimensions.

Analysis on most heavily overused and underused syntactic features shows that, when compared with native speakers, Hong Kong students favour the use of present tense, tentative style, simple noun phrase structure and explicit clause-relation marking in oral presentations. They prefer using present tense constructions, adopting pronouns for nominal functions and using an interactive tone in written essays. In both the learner speech data and the learner writing data, quite a number of the overused and underused syntactic features are closely related to the differences between speech and writing found in Biber (1988), hinting at Hong Kong students' inadequate sensitivity to mode differences in English.

Analysis on textual dimensions suggests that Hong Kong students' oral presentations differ from comparable native-speaker oral production by being more written-like in terms of the use of explicit/situation-dependent reference and the inclusion of abstract/non-abstract information. Hong Kong students' written essays deviate from comparable native-speaker written production by being more spoken-like in terms of the involved/informational focus and the use of explicit/situation-dependent reference. Moreover, both the learner speech data and the learner writing data bear some resemblance to common native-speaker genres in the opposite mode.

The present study clearly demonstrates that Hong Kong students have limited sensitivity to the conventional usage of syntactic features in spoken and written English. The teaching profession should help the students develop better sociolinguistic competence when teaching grammar, speaking and writing. Future research on second language acquisition should also focus more on the learners'

sociolinguistic development so that second language learners' communicative ability can be better understood.

論文摘要

本項研究旨在檢視本港中學生的英語特徵，以及他們對口語及書面語之差異之認識。調查資料包括五十二篇中六學生發表的英語口頭報告及五十二篇中七學生在公開考試寫的英語文章。本研究以 Biber (1988) 作爲理論基礎，調查學生對六十七種句法特徵的使用頻率，並將之與以英語爲母語人士作出比較。

研究的結果顯示，本地中學生在英語口頭報告中，比以英語爲母語人士更愛以現在時態及不確定語調討論事情，也較愛使用連接詞和結構較簡單的名詞短語。寫作方面，他們也比以英語爲母語人士更喜歡使用現在時態，同時較常以代名詞取代名詞，並加入更多第二人稱句法以拉近與讀者之間的距離。不論在英語口頭報告中還是在英語寫作中，本港中學生與以英語爲母語人士之間使用頻率差別最大的句法特徵，部份與 Biber (1988) 內所提及的英語口語及書面語之差異有關。由此可見，他們未必能夠全面掌握英語口語及書面語之常規用法。

根據 Biber (1988) 理論模型的分析顯示，本港中學生的英語口頭報告在兩個範疇中，即不依賴情景的指稱與情景界定的指稱的使用方面，以及信息表達的抽象程度方面，都比以英語爲母語人士所作的演說更像書面語。同樣，他們的英語文章亦在兩個範疇中，即內容是着重人際關係還是信息交流方面，以及不依賴情景的指稱與情景界定的指稱的使用方面，比以英語爲母語人士所寫的文章更像口語。

總括而言，本項研究清楚顯示香港學生對英語口語及書面語之差異之認識有限，筆者建議教育界應在教授英語文法、說話及寫作技巧時，加強學生對英語口語及書面語之差異之了解。今後的第二語言習得研究可以把重點放在學習者學習掌握英語常規用法的過程方面，從而加深對學習者的英語交際能力的認識。

ACKNOWLEDGEMENT

I would like to express my gratitude to all those who helped me during my three-year PhD study.

First of all, I would like to sincerely thank my supervisor, Prof. Gerald Nelson. This thesis would not have been possible without his professional guidance and continuing support. There were times of frustration during various stages of research implementation and thesis writing. It was his advice and patience that helped me overcome all the obstacles and complete this thesis.

I am very grateful to my thesis committee members, Prof. George Braine and Prof. Carmen Lee, who constantly encouraged me and gave me useful comments. I am also indebted to Prof. Peter Skehan, a former committee member, for his invaluable advice at the early stage of my study.

I would like to thank my friend, Francine Chan, who kindly helped me to contact her colleagues and arrange the data collection of my learner speech data, and two other teachers in her school, Ms Chan and Ms Yip, who kindly allowed me to collect data in their English classes. Special thanks are due to the 64 Form 6 students who participated in the study.

Last but not least, I would like to express deep gratitude to the Department of English of the Chinese University of Hong Kong. When I joined the department as an undergraduate student in 2000, I could never imagine that I would go far and get a doctorate one day. I am very grateful to all the teachers, both linguistics ones and literature ones, who taught me during the eight years I studied in the department. It was their teaching that fostered my love of the English language and my interest in English linguistics.

TABLE OF CONTENTS

CHAPTER 3 METHODOLOGY

CHAPTER 4 RESULTS AND DISCUSSION: SPEECH DATA OF HONG KONG LEARNERS

APPENDICES

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1 INTRODUCTION

1.1 Background

Communicative language use involves knowledge of different kinds. In his model of communicative language ability, Bachman (1990) proposes various components of language competence (see Figure 1). Two major components are organizational competence, which governs the relationships between linguistic signals and their referents, and pragmatic competence, which concerns the dynamics between language users and contexts (Bachman, 1990, p. 89). Under pragmatic competence, there is the component of sociolinguistic competence, which refers to "knowledge of the sociolinguistic conventions for performing language functions appropriately in a given context" (Bachman, 1990, p. 90). One specific type of such knowledge that Bachman (1990) mentions is the sensitivity to differences in register. In other words, knowing how speakers of the same speech community vary their language use in different registers can facilitate one's language comprehension and production. Following Halliday, McIntosh and Strevens (1964), Bachman (1990) characterises register in terms of the field of discourse, the mode of discourse and the style of discourse (p. 95). In particular, the mode of discourse refers to "the medium or mode of the language activity" (Halliday, McIntosh & Strevens, 1964, p. 91). Hence, sensitivity to the mode of discourse, i.e. language users' awareness of the differences between speech and writing, represents a part of the sociolinguistic competence that is required for communicative language use.

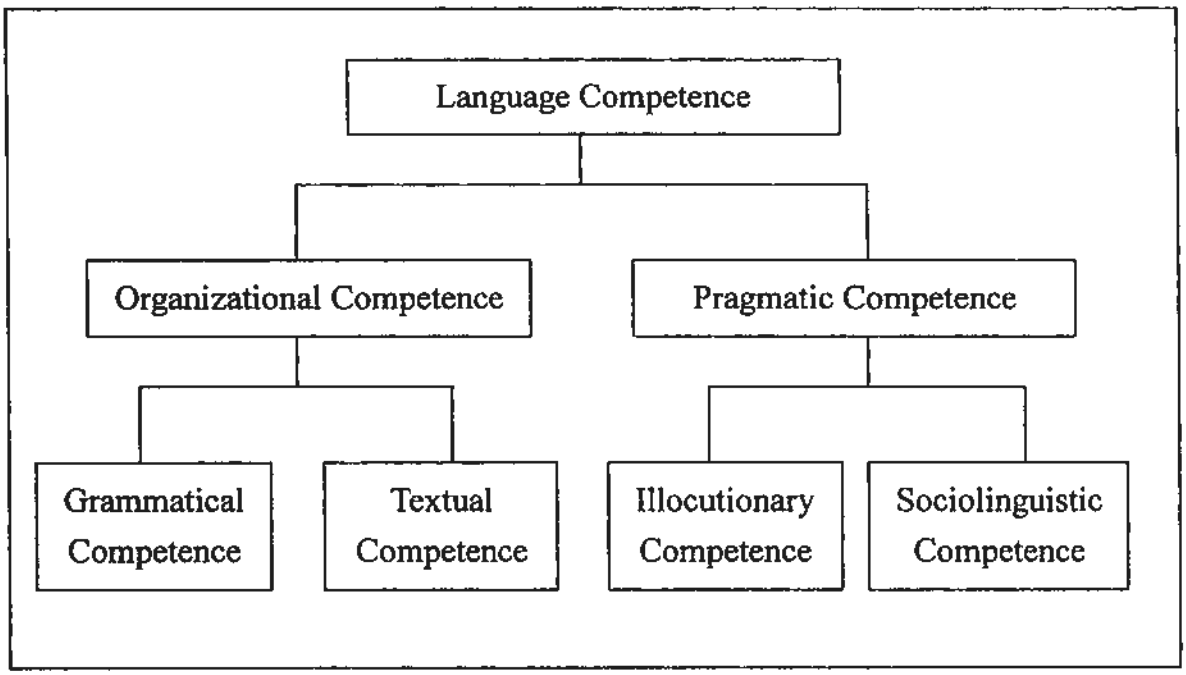

Figure 1. Model of language competence (Bachman, 1990, p. 87)

Research on second language learners' English proficiency tends to emphasise development in grammatical competence. Other components of language competence are much less investigated. However, grammatical competence alone cannot guarantee effective communication. Development in other components is equally important to second language learners. It is thus desirable to conduct more studies on these under-researched areas. Ellis (1994) observes that although the pragmatic aspects of learner language are attracting more attention in the second language acquisition research field, most studies concentrate on the acquisition and the production of speech acts (p. 159). Therefore, relatively little is known about other facets of second language learners' pragmatic competence, notably sociolinguistic competence.

1.2 Rationale

Spoken English and written English are partially overlapping yet partially distinct. Their subtle relationship vexes many second language learners of English. Research evidence has shown that spoken-like features are not uncommon in second language learners' writing (e.g. Chao, 2003; Cobb, 2003; McCrostie, 2008; Petch-Tyson, 1998). Nevertheless, the few existing studies on second language learners' ability to cope with mode differences are rather limited in scope: Most of them focus on identifying spoken-like usage in the form of involvement features in the learners' writing. Other linguistic differences between spoken and written English have been largely ignored. Apparently, no attempt has been made to investigate written-like features in speech. These limitations in methodology suggest that there is much room for improvement in research on second language learners' sensitivity to mode differences.

Studies on Hong Kong learners' English occasionally contain comments on the learners' confusion over spoken and written language use. Nonetheless, these comments are only inserted as proposed explanations for some interlanguage phenomena. Hong Kong students' ability to produce English appropriate for the mode of discourse is seldom a central topic for investigation. In view of this research gap, the present study was conducted.

1.3 Scope of the Present Study

The present study is dedicated to one under-researched area in second language learners' communicative competence: sociolinguistic competence. The investigation focuses on second language learners' sensitivity to differences between spoken English and written English in language production. The second language learners

under investigation are senior students in Hong Kong secondary schools. The data consists of samples of their spoken English and written English, both of which total approximately 10,000 tokens. The appropriateness of the students' language use for a specific mode of discourse was evaluated by a comparison with the target language norm of British English. The comparison was made with reference to the theoretical framework of Biber (1988), which specified how native speakers varied their use of syntactic features among different genres in spoken and written English.

1.4 Significance

The significance of the present study can be explained from three perspectives. First, from the theoretical perspective, the present study contributes to the understanding of second language learners' sociolinguistic competence, a neglected area in second language acquisition literature. Second, from the methodological perspective, the present study investigates second language learners' awareness of mode differences systematically by drawing on a theoretical model of native-speaker usage. Differences between spoken and written English manifest themselves in a wide range of linguistic features. Observations on a single feature cannot generate conclusive evidence of the appropriateness of interlanguage. Taking into account a relatively large number of syntactic features, the present study can provide a comprehensive picture of the learners' sensitivity to spoken and written English. Finally, from the pedagogical perspective, the present study informs educational practitioners of Hong Kong students' English proficiency. The study not only presents a survey of Hong Kong students' use of syntactic features in speech and writing, but also evaluates how well the students' usage conforms to native-speaker conventional usage. Better understanding Hong Kong students' strengths and

weaknesses, the teaching profession can act accordingly.

1.5 Research Questions

The major research question guiding the present study is: Does Hong Kong students' English produced in one mode exhibit characteristics typically found in English in another mode? Two identical sets of research questions have been devised to investigate learner language data produced in the spoken mode and that produced in the written mode. They are presented separately as Research Question 1 and Research Question 2 in this section so that specific reference can be made in the rest of this thesis:

1. Does Hong Kong students' spoken English exhibit characteristics typically found in English writing?
    1.1. How are Hong Kong students' oral presentations different from comparable native-speaker oral production (prepared speeches), in terms of the use of syntactic features?
    1.2. Do the differences in the use of syntactic features observed in Question 1.1 suggest that Hong Kong students' oral presentations exhibit written-like characteristics?
    1.3. How are Hong Kong students' oral presentations different from comparable native-speaker oral production (prepared speeches), in terms of the three textual dimensions identified in Biber's (1988) model?
    1.4. Is there any similarity between Hong Kong students' oral presentations and common native-speaker written genres, in terms of the three textual dimensions identified in Biber's (1988) model?

2. Does Hong Kong students' written English exhibit characteristics typically found in English speech?

    2.1. How are Hong Kong students' written essays different from comparable native-speaker written production (written essays), in terms of the use of syntactic features?

    2.2. Do the differences in the use of syntactic features observed in Question 2.1 suggest that Hong Kong students' written essays exhibit spoken-like characteristics?

    2.3. How are Hong Kong students' written essays different from comparable native-speaker written production (written essays), in terms of the three textual dimensions identified in Biber's (1988) model?

    2.4. Is there any similarity between Hong Kong students' written essays and common native-speaker spoken genres, in terms of the three textual dimensions identified in Biber's (1988) model?

In both sets of questions, the first two sub-questions focus on the microscopic analysis on Hong Kong students' use of syntactic features and the last two sub-questions draw the readers' attention to the macroscopic analysis on textual dimensions of the learner language data.

1.6 Organisation of the Thesis

This thesis consists of six chapters. The present chapter, Chapter 1, orientates the readers towards the thesis by providing some background information of the study. The rationale and the signification of the study are stated. The research questions are introduced and the structure of the thesis is presented.

Chapter 2 summarises relevant studies from different research fields composing the theoretical background of the present study. It reviews literature on the differences between speech and writing, research on language learners' ability to tackle mode differences and studies on syntactic features in Hong Kong students' English.

Chapter 3 offers information on methodological issues necessary for the interpretation and the evaluation of the research findings. It not only gives details on the theoretical framework adopted in the present study, but also explains how the study was implemented by discussing the research design, the language data, and the data processing and analysis procedures.

Chapter 4 and Chapter 5 report findings generated from the analysis on Hong Kong students' oral presentations and the analysis on Hong Kong students' written essays respectively. Figure 2 and Figure 3 illustrate the structure of the two chapters. In each chapter, the analysis on syntactic features and the analysis on textual dimensions are divided into two sections. In each section, the results are discussed in relation to the corresponding research questions. Overall, findings in these two chapters can help determine whether Hong Kong students show confusion over spoken and written language use.

Chapter 6 ends this thesis by drawing a conclusion from the findings in Chapter 4 and Chapter 5. Different implications on pedagogy and research are considered. Limitations of the study are noted and directions for future research are suggested.

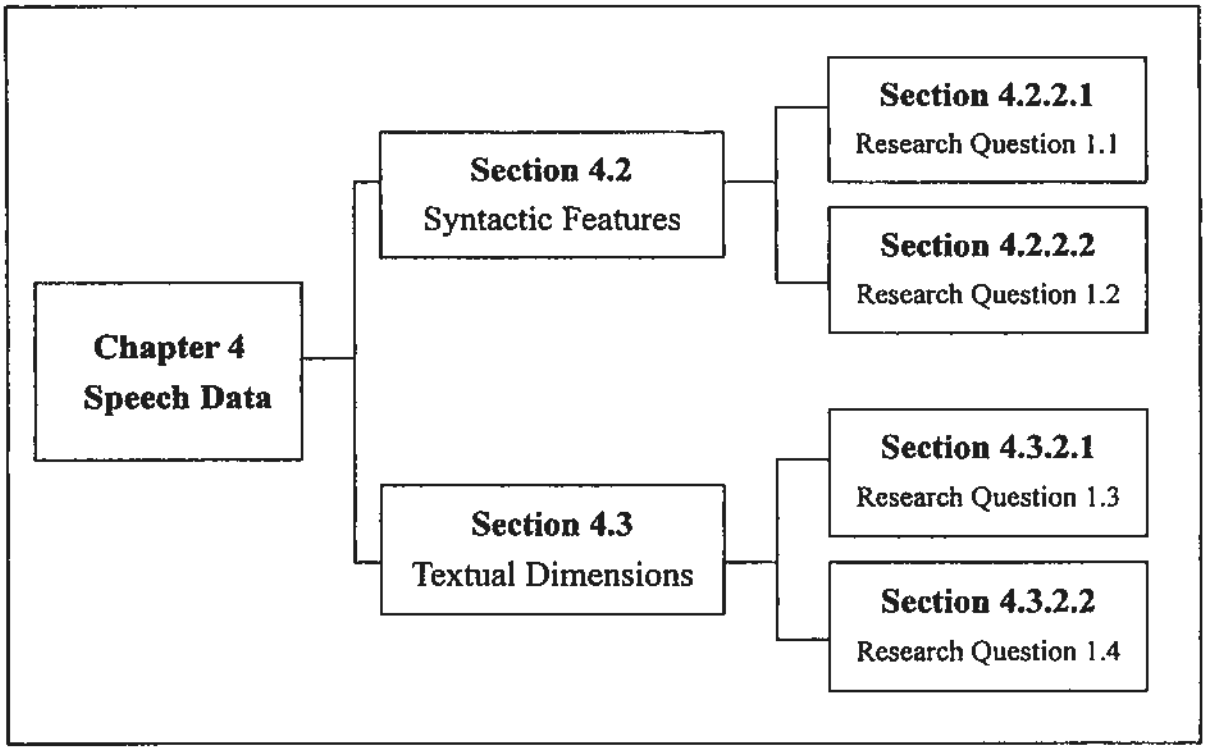

Figure 2. Organisation of Chapter 4

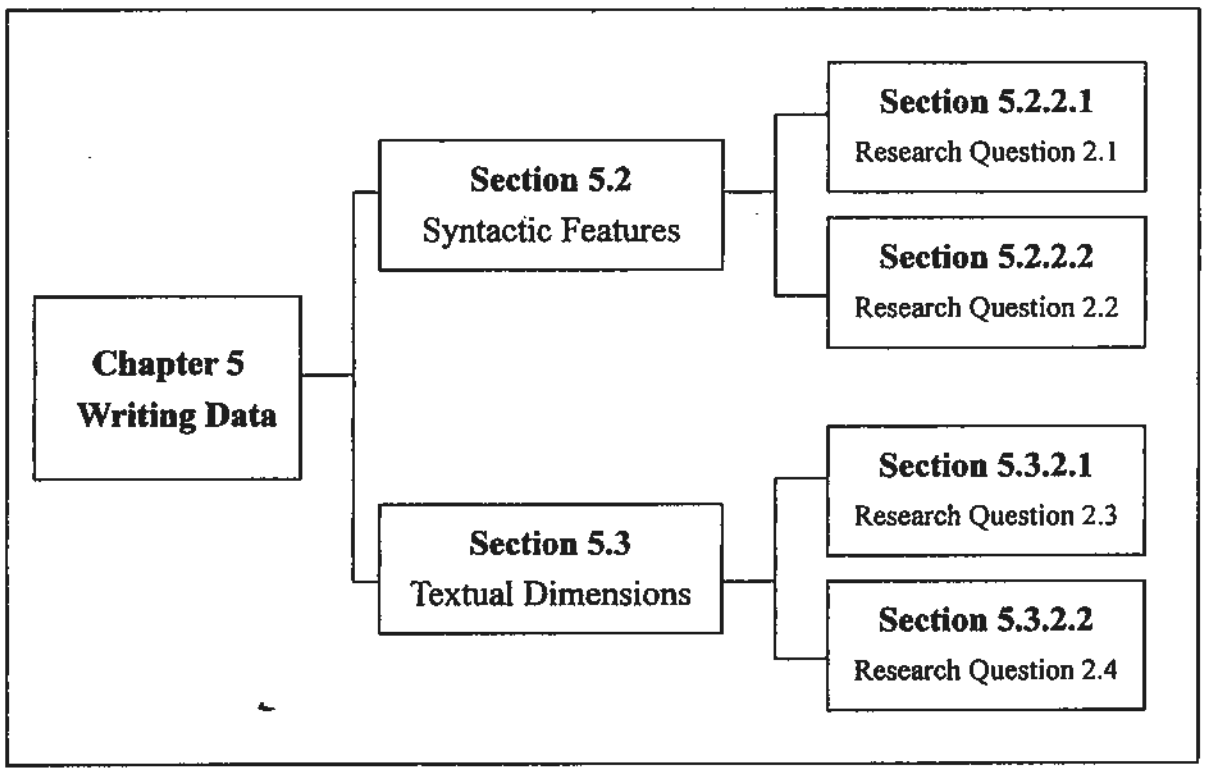

Figure 3. Organisation of Chapter 5

CHAPTER 2 LITERATURE REVIEW

2.1 Introduction

After a brief introduction of the whole thesis, this chapter provides a review of relevant literature so that the readers can understand more about the research background of the present study. This chapter is divided into three parts. The first part (Section 2.2) introduces the research tradition on mode differences. Some early studies investigating differences between spoken and written English are reviewed. The limitations of these early studies are discussed and the methodological trend of more recent studies is also noted. The second part (Section 2.3) examines research focusing on language learners' ability to cope with mode differences. Studies of first language learners are cited to show that developing this aspect of sociolinguistic competence is not an easy task even for native speakers. Several studies on non-Chinese and Chinese second language learners of English are reported so as to illustrate how sensitivity to mode differences has been researched. The last part (Section 2.4) offers some information about Hong Kong students' use of syntactic features. The emphasis is placed on research findings involving overuse and underuse patterns of syntactic features generated from a comparison of Hong Kong learners' data with comparable native speakers' data.

2.2 Differences Between Speech and Writing

There has been a relatively long history of research into the differences between speech and writing. According to DeVito (1967), people living 2000 years ago already noted the differences between the two modes and early quantitative studies can date back to the 1920s (p. 354). However, it was not until the 1950s and the

1960s that this research field began to attract scholars' attention (Roberts & Street, 1997, p. 168). Studies of various interests were carried out by researchers from different disciplines like anthropology, literary studies, linguistics, education and psychology (Akinnaso, 1982, p. 97; Chafe & Danielewicz, 1987, p. 83). Given the vast amount of literature accumulated, it is simply impossible for the review in this section to be exhaustive. Therefore, the following discussion is limited to quantitative studies on the syntactic differences between spoken and written English. In Section 2.2.1, some early studies are described so as to exemplify the popular methodology used to investigate mode differences from the 1960s to the 1980s. The findings of these studies are summarised to show how the early researchers characterised differences between speech and writing. In Section 2.2.2, limitations associated with the methodology of the early studies are identified. Issues concerning the participants, the language data and the syntactic features are raised. In Section 2.2.3, methodological development in the past 20 years is discussed. Some studies on mode differences published since the late 1980s are cited to demonstrate how the limitations of early studies have been overcome. A relatively new research method, multi-dimensional analysis, is also introduced.

2.2.1 Early Studies: 1960s – 1980s

This section reviews a few early studies on the differential use of syntactic features in spoken and written English. For each study, details of research design are included and key findings are reported. It is hoped that after reading this section, the readers can get a sense of the methodology employed by the early researchers, as well as the theoretical understanding of mode differences during that period. Owing to the scope of the thesis, the review in this section is bound to be selective. The

following studies have been chosen to exemplify different general qualities associated with speech and writing and they are organised according to these different qualities in this section. For a comprehensive review of early research on differences between spoken and written English, readers can refer to Akinnaso (1982) and Biber (1988).

2.2.1.1 Fragmentation/ integration.

O'Donnell (1974) investigated syntactic density in speech and writing by comparing some oral and written data produced by the same person. He used data from a television interview and some newspaper articles, both involving some publicly expressed opinions on general issues. Realising the lack of a comparable structural unit in speech and writing, O'Donnell adopted T-units as the unit of comparison in his study. A T-unit represents an independent clause, plus any dependent clauses attached to it (O'Donnell, 1974, p. 103). His data consists of 100 T-units from the participant's speech and an equal number of T-units from his writing. The results show that on average the T-units in writing (24.97 words) are longer than the T-units in speech (17.92 words). There are also more T-units with dependent clauses in writing (68 out of 100) than in speech (56 out of 100). The researcher regards the greater T-unit length associated with writing as an indicator of higher syntactic density. Furthermore, O'Donnell's study shows that while speech contains more nominal clauses, writing contains more adjectival and adverbial clauses. There are also more nonfinite verbs, passive constructions, auxiliaries and attributive adjectives found in writing. As O'Donnell acknowledges in his conclusion, the fact that only one participant was involved seriously affects the external validity of his study (p. 109). However, the value of his study lies in his effort in seeking a suitable

syntactic unit for comparison across speech and writing, as the use of T-units can be one way to deal with the absence of sentence boundaries in spoken discourse (Akinnaso, 1982, p. 107).

Chafe (1982) relates the differences in the use of certain linguistic structures to the different production conditions in speech and writing. One of his claims is that writing is more integrated and speech is more fragmented because language users usually take much more time to write than to speak (pp. 36-38). He enumerates a number of devices signaling integration and argues that they are more common in writing by referring to statistics from his approximately 10,000-word samples of dinner-table conversations and 12,000-word samples of academic papers, both collected from 14 subjects consisting of faculty members and graduate students. The integration devices that Chafe (1982) discusses include nominalizations, *of* prepositional phrases for genitive subjects/ objects, present and past participles (including those used as nouns and adjectives, as well as those used in reduced relative clauses), attributive adjectives, conjoined noun, adjective or verb phrases, series, sequences of prepositional phrases, *to* and *that* complement clauses, and relative clauses. All these features have been found to exhibit a higher frequency in the researcher's written data than in his spoken data. Regarding the oral mode, Chafe points out that fragmentation can be perceived from the lack of connectives and the use of conjunctions (*and, but, so, because*) at the beginning of idea units. This widely cited study represents an important attempt to quantify a relatively large number of linguistic features in accounting for the differences between speech and writing. The use of idea units, i.e. units with a coherent intonation contour separated by pauses (Chafe, 1982, p. 37), also offers another possibility in handling the lack of clear sentence boundaries in speech. Nevertheless, Chafe's (1982) comparison is

based on informal spoken discourse and formal written discourse. In other words, the two types of data differ not only in medium, but also in style. Consequently, the different distributions of linguistic features reflect both mode differences and stylistic differences. Akinnaso (1982) even suspects that Chafe's (1982) findings "result more from the maximization of contrasts in the data base than from differences in modality" (p. 108).

Beaman (1984) focused her investigation of syntactic complexity on the use of coordination and subordination. In order to control differences in variables other than the mode, she compared some unplanned, informal, spontaneous discourse produced by the same group of participants in both spoken and written modes. Altogether 20 spoken narratives (12,594 words) and 20 written ones (7,072 words) were used. Her detailed analysis of coordination and subordination of these two types of discourse yielded some interesting findings. First, there are more coordinated sentences in the written stories (38 per 100 sentences) than in the spoken ones (25 per 100 sentences). Second, there are low percentages of subordinated sentences in both the spoken stories (13 per 100 sentences) and the written stories (12 per 100 sentences). Third, there are more finite nominal subordinate clauses (*that*-clauses, interrogative clauses and nominal relative clauses) in the spoken stories (10.9 per 1,000 words) than in the written ones (3.8 per 1,000 words), but there are more nonfinite nominal subordinate clauses (*to*-infinitive clauses and *-ing* clauses) in the written stories (30.4 per 1,000 words) than in the spoken ones (17.5 per 1,000 words). Fourth, there are more adjectival relative clauses in the spoken stories (11.7 per 1,000 words) than in the written stories (6.9 per 1,000 words). Finally, there are more adverbial subordinate clauses in the written stories (13.0 per 1,000 words) than in the spoken stories (8.4 per 1,000 words). Beaman's (1984) conclusion is that regarding the use of

subordination, "spoken narrative is on the whole just as complex as, if not more complex in some respects, than written narrative" (p. 78). She also agrees with Halliday that "the types of complexities involved in the two modalities are different" (Beaman, 1984, p. 78). Although the third and the last findings in Beaman (1984) confirm O'Donnell's (1974) results, her second finding disagrees with O'Donnell (1974). Given the larger amount of data involved in Beaman (1984), one may readily believe in her finding and reject O'Donnell's (1974). But it should be reminded that the nature of the discourse used in the two studies is different and a direct comparison may not be applicable. One room for improvement for Beaman (1984) is the presentation of data. The statistics are sometimes given as a percentage over the total number of sentences and sometimes given as a normalised frequency. Readers may find this confusing.

Chafe and Danielewicz (1987) further discuss the properties of spoken and written language by comparing some conversations, lectures, letters and academic papers produced by 20 university professors and graduate students. They suggest that, owing to the limited capacity of short-term memory, the average length of idea units produced by speakers is shorter than that produced by writers (Chafe & Danielewicz, 1987, p. 86). Therefore, speakers tend to use fewer linguistic features that increase the size of idea units such as prepositional phrases (especially sequences of prepositional phrases), nominalizations, attributive adjectives and nouns, phrasal coordination and participles. Moreover, speakers prefer simple clausal coordination to other structures which may complicate the relationships between different clauses. Nevertheless, the researchers remind the readers of the fact that some of the differences they discuss exist "because of differences in the speaking and writing processes themselves" but others exist "because of the varied contexts, purposes, and

subject matters of both spoken and written language" (Chafe & Danielewicz, 1987, p. 87). As a result, the findings of Chafe and Danielewicz (1987) should be treated with caution.

2.2.1.2 Involvement/ detachment.

DeVito's (1966) comparison of oral discussions (9,000 words) and written articles (9,000 words) produced by the same 10 university faculty members examined two features relevant to the concept of involvement: self-reference words and consciousness of projection terms, i.e. terms indicating the speakers' observation. His results show that both of these are significantly more common in speech than in writing. In speech, there are 26.4 self-reference terms and 8.5 consciousness of projection terms per 900 words, but in writing there are only 15.4 self-reference terms and 1.5 consciousness of projection terms per 900 words. Nonetheless, the two categories are vaguely defined in DeVito (1966). What has and what has not been counted in his study remains a puzzle to the readers.

In Chafe's two studies mentioned in Section 2.2.1.1, the researchers also discuss differences between speech and writing in terms of involvement and detachment. Chafe (1982) considers writing more detached and speech more involved because language users can usually interact with their audience when speaking, but not when writing (p. 45). He noticed that a few linguistic features showing involvement with the audience were more common in his informal spoken data. These devices include first person references (i.e. references to the speaker), references to the speaker's mental processes, devices monitoring information flow (e.g. *well, I mean,* and *you know*), emphatic particles (e.g. *just* and *really*), devices for fuzziness (e.g. *and so on, something like,* and *sort of*), and direct quotes. In contrast, devices for signaling

detachment like passive voice and nominalizations were found to be more frequently used in formal writing. In Chafe and Danielewicz (1987), the researchers suggest that the presence of audience in speech makes speakers produce language with more involvement with audience, selves and the concrete reality (p. 105). This involvement is evident in the speakers' use of interactional responses, the phrase *you know*, first person pronouns and place and time adverbials. Chafe and Danielewicz (1987) think that, whereas speakers tend to produce involved language, writers usually produce detached language which is relatively abstract and timeless (p. 108). Writers use more often abstract subjects, passive constructions and academic hedges indicating probability. All these devices indicate the writers' attempt to distance themselves from their discussion.

2.2.1.3 Abstractness.

DeVito (1967) studied the differences between speech and writing by examining their levels of abstraction as represented by a formula deriving from the frequencies of finite verbs, definite articles and abstract nouns. This formula, a simplified version, has the advantages of clearly defined variables and high correlation with the original formula involving 16 different linguistic features (p. 355). DeVito (1967) used data from written articles totaling 8,000 words and an equal amount of oral discussions based on these articles. Both types of discourse were produced by 10 university faculty members. The findings suggest that speech is significantly less abstract than writing in terms of the scores obtained from the formula. In particular, there are significantly more finite verbs and significantly fewer abstract nouns in speech than in writing. The researcher speculates that the higher frequency of occurrence in English makes finite verbs more readily available

to speakers and the direct nature of finite verbs also suits the production conditions in speaking (Devito, 1967, pp. 357-358). In contrast, abstract nouns are usually longer words and they are less commonly used, hence less readily available to speakers (Devito, 1967, p. 358). As DeVito (1967) notes in his article, the three features under investigation do not represent all factors affecting the level of abstraction in discourse (p. 357). The small subject size and the choice of academics as subjects also limit the generalisability of his research, so his results are better viewed as indicative only.

2.2.1.4 Planning.

Ochs (1979) draws some generalisations about the features of relatively unplanned discourse, as opposed to planned discourse. Firstly, in unplanned discourse, language users rely more on context but less on syntax in getting meaning across (Ochs, 1979, p. 62). The evidence cited includes the omission of referents and the low density of conjunctions frequently found in spontaneous speech. Secondly, people employ more morphosyntactic structures that they acquired early in their life (Ochs, 1979, p. 68), which is exemplified in their greater use of the determiner-plus-noun structures than the noun-plus-relative-clause structures, as well as their greater use of active over passive voice and present over past tense, in spontaneous speech. Thirdly, it is more common to find repetition and repair in unplanned discourse (Ochs, 1979, p. 70). Finally, there is a higher degree of similarity in the form and content of unplanned discourse (Ochs, 1979, p. 72).

When discussing the above four characteristics, Ochs (1979) mainly uses illustrative examples to convince readers of her ideas. Few statistics have been provided. Although the points that Ochs makes are based on the distinctions between

planned and unplanned discourse, she does imply in her article that writing is more planned than speech:

> On the other hand, we can find extreme examples of total discourse planning when we deliver or listen to a speech that has been written down in advance and has been read aloud. (Ochs, 1979, p. 56)
>
> Furthermore, written discourse may be more plannable than spontaneous spoken discourse. In writing, the communicator has more time to think out what he is going to say and how it will be said. Additionally, the writer can rewrite and reorganize the discourse a number of times before it is eventually communicated. (Ochs, 1979, p. 58)

It is clear from these two quotes that Ochs considers speech based on some written sources to be more planned and that she thinks the production conditions in writing allow more planning to occur. The implication underlying these two quotes explains why this article is often cited in discussion on mode differences. In fact, the data in Ochs (1979), which consists of (a) some child-child, child-adult, and adult-adult communication, (b) some informal conversations of adult native and non-native speakers, and (c) some unplanned oral narratives and planned written narratives produced by the same groups of participants (p. 54), also reinforces the impression that the contrasts Ochs sets up should be relevant to the differences between speech and writing. However, readers should bear in mind that Ochs (1979) never equates the spoken-written distinction with the unplanned-planned continuum, as she acknowledges the lack of unplanned written discourse and planned spoken discourse in her data set (p. 55).

2.2.2 Limitations of Early Studies

Section 2.2.1 has introduced some quantitative studies published from the 1960s to the 1980s. Their methodological approach, which is also the major approach in literature, is for the researchers to collect some spoken and written discourse and then count the frequencies of certain linguistic features in the two types of materials. However, this quantitative, product-oriented approach does not represent the only way to carry out research on mode differences. Some researchers investigated the issue in a qualitative manner. For example, Tannen (1982) qualitatively analysed some spoken and written narratives produced by the same participants. In addition to studying the end products, another possibility for qualitative research is to investigate the language production process. For example, Halpern (1984) conducted a study by examining the adaptations needed to convert raw transcripts of speech into edited written records.

Reviewing the literature on differences between speech and writing, Akinnaso (1982) has observed that "in spite of the vast amount of data that have been accumulated on the relationship between spoken and written language, no agreement has been reached either on the exact nature or the extent of the differences (and similarities) between the two" (p. 99). In fact, a number of problems have been identified by researchers working in the field. First, many studies do not control variables other than the mode differences. As a result, their findings may reflect not only differences between speech and writing, but also differences in other factors (Akinnaso, 1982, pp. 109-110; Beaman, 1984, p. 51; Tannen, 1982, p. 5). For example, Chafe (1982) compares informal spoken language (dinner-table conversations) with formal written language (academic papers). Besides differences on the medium, these two types of discourse differ in terms of the level of formality,

the degree of planning and the variety of topics. Therefore, the syntactic differences observed may not be true differences between speech and writing. Few studies can eliminate other influences by adopting an experimental-like design like Drieman (1962) did.

Second, the use of a relatively small amount of data and the choice of data limit the external validity of the research. Usually only one genre is selected to represent the spoken or the written language (Biber, 1988, pp. 52-53). For example, dinner-table conversations, oral discussions, spoken narratives and television interviews have been chosen as samples of spoken language in Chafe (1982), DeVito (1966), Beaman (1984) and O'Donnell (1974) respectively. The researchers' use of only one genre to represent the language in a particular mode ignores the wide range of linguistic possibilities existing within each mode. There is in fact more than one type of spoken or written language. As pointed out in Tannen (1985), the most researched genres in the early studies, casual conversation and expository prose, "typify but do not exhaustively characterize spoken and written discourse" (p. 124). Furthermore, the number of subjects involved and the size of the language data sampled in each study are rather limited (Biber, 1988, p. 52). For instance, DeVito's (1966) spoken data from 10 subjects already makes up 9,000 words. As the researcher needed to analyse both spoken and written data, the resultant data size, 18,000 words, constituted huge workload for the researcher when personal computer was not yet popular. At the same time, many studies (e.g. Chafe, 1982; Chafe & Danielewicz, 1987; DeVito, 1966, 1967) have chosen university students and professors as the target group of investigation (Akinnaso, 1982, pp. 110-111). The heavy reliance on language data produced by people in university represents a serious bias in the data source.

Third, a relatively small number of linguistic features have been put under investigation (Biber, 1988, p. 54). Some studies (e.g. DeVito, 1966, 1967) counted fewer than ten features. Studies like Chafe (1982) and Chafe and Danielewicz (1987) which counted around 20 features can already be regarded as large-scale ones and these large-scale studies are not common. As suggested in Biber (1988), sociolinguistic research has demonstrated the complexity in natural language variation (p. 54). It seems unlikely that investigations based on a few linguistic features can adequately capture the differences between spoken and written English.

Finally, terms proposed to characterise the different qualities of spoken and written discourse are often overlapping and confusing. They are problematic in the sense that they only represent characteristics of the prototypes, but not all the oral or written genres. They are also problematic in the sense that they have been proposed and used by different researchers, which makes their compatibility a question for the readers. For example, what O'Donnell (1974) refers to as *syntactic density* is actually similar to *syntactic complexity* in Beaman (1984). The concept is again highly relevant to the idea of *integration* mentioned in Chafe (1982), because in order to pack more information into a spoken or written unit, language users tend to make the resultant syntactic structure more complicated. In another article, Chafe (1985), Chafe invented yet another term *idea unit expansion* to refer to what he previously called *integration*. Unfortunately, there is no clear definition for these different terms. The sets of syntactic features that are categorised under these different labels cannot help differentiate the terms, because similar but not identical features were counted in these different studies. Readers who want to compare results across studies are left to judge for themselves whether the different terms refer to the same thing or not.

2.2.3 Methodological Development Since the Late 1980s

Early researchers working on mode differences were not totally unaware of the limitations discussed in the previous section; they were simply unable to solve the problems. Nevertheless, the situation has begun to change since the 1980s. With the wide-spread use of computer and the development of corpus linguistics, the limitations can be gradually overcome. One important contribution has come from the advent of large language corpora. The compilation of corpora can grant researchers access to substantial materials of various types produced by different population groups, solving the second problem identified in Section 2.2.2. As a result, more recent studies on speech and writing can be implemented on a larger scale, with a much larger amount of data, as well as more varied types of data. For example, Greenbaum and Nelson's (1999) investigation on the use of elliptical clauses in speech and writing made use of 54 spoken texts from four different categories, totaling 115,000 words, and 28 written texts from eight different categories, totaling 62,000 words, from the British component of the International Corpus of English (ICE-GB). Moreover, the use of large corpora enables researchers to investigate some relatively infrequent syntactic constructions. For example, Johansson's (1993) study on the use of *whose* and *of which* with non-personal antecedents in written and spoken English has found that there are only 16 occurrences of *whose* per million words and 5 occurrences of *of which* per million words in the spoken component of the Birmingham Corpus. In Johansson and Geisler (1998), a frequency of 27 per million words has been found concerning the use of *which* preposition stranding structures in the Lancaster-Oslo-Bergen Corpus (the LOB Corpus), a corpus of written English. Features like these will not occur at all, if the researchers just work on data of a few thousand words. Consequently, the availability of large English

corpora has greatly facilitated research into the differences between speech and writing.

What is more significant is that advances in computer technology and corpus linguistics have prompted the introduction of a new methodology to the investigation of differences between spoken and written English. Drieman (1962) has commented that regarding research on mode differences, "[o]nly the study of extensive spoken and written texts, obtained from subjects of all age-groups and social levels and about diverse themes, will yield reliable results" (p. 55). The new methodology, multi-dimensional analysis, is capable of fulfilling Drieman's (1962) requirement. This new approach, advocated by Douglas Biber, allows researchers to analyse a large number of linguistic features in a large amount of language data, hence helping the researchers to take into account the various situational variables in human language use. It is a method that can overcome the first three limitations associated with early studies on mode differences discussed in Section 2.2.2.

The multi-dimensional approach is grounded in sociolinguistic theory on linguistic variation which suggests that systematic variation exists across groups of co-occurring linguistic features rather than on an individual marker alone (Biber, 1988, pp. 21-22). These co-occurring patterns are taken to represent some underlying functional dimensions (Biber, 1988, p. 13). This methodology can be considered innovative in three ways. Firstly, while most early studies compare speech and writing from only one aspect, the multi-dimensional analysis can capture differences on various dimensions (Biber, 1988, p. 24). It is evident from the short review in Section 2.2.1 that the qualities that early researchers proposed to characterise spoken and written English focus on different aspects. For example, fragmentation/ integration and involvement/ detachment are concepts at different levels: The former

is related to the surface structure of language, but the latter is concerned with the relationship between the language users and the audience (Tannen, 1982, p. 17). Therefore, taking into consideration differences from various aspects, the multi-dimensional approach can better account for the mode differences. Secondly, the methodology shows differences between speech and writing in terms of "continuous scales of variation rather than dichotomous poles" (Biber, 1988, p. 24). This change in perspective allows researchers to reveal subtle differences among genres within the same production mode. Finally, whereas most previous research determines features distinguishing speech and writing in advance, the multi-dimensional analysis identifies patterns of co-occurring linguistic features through empirical studies (Biber, 1988, p. 24). The results of the new approach are thus more capable of reflecting differences between speech and writing.

The multi-dimensional analysis was first fully documented in Biber's (1988) study on variation across spoken and written English (see Chapter 3 Section 3.2 for details). Since then, Biber and his colleagues have applied the same method to investigate various issues concerning mode differences. For example, Biber (1995) adopted the multi-dimensional approach to compare four languages, English, Nukulaelae Tuvaluan, Korean and Somali, by using data from their spoken and written genres. Biber (2001) conducted a multi-dimensional analysis of some eighteenth-century spoken and written English registers in a corpus called ARCHER. From his studies, Biber has realised the importance of register to the distribution of linguistic features. Consequently, he urges for research on register variation and suggests that such research can produce more fruitful results for the investigation of mode differences (Biber & Vásquez, 2008, p. 536). His recent book, Biber (2006), includes a multi-dimensional analysis of spoken and written genres appearing

specifically in the university context. In addition to Biber and his colleagues, other researchers have also employed the multi-dimensional method when studying spoken and written English. Louwerse, McCarthy, McNamara and Graesser (2004) conducted a multi-dimensional study by using corpus data nearly identical to Biber (1988). Whereas Biber (1988) counted primarily lexico-grammatical features, they counted some features of cohesion at the discourse level. Crossley and Louwerse (2007) also carried out a multi-dimensional analysis by counting bigrams, i.e. co-occurrence of two neighbouring words, in some spoken and written English corpora. It appears that since the late 1980s, i.e. after the publication of Biber (1988), the multi-dimensional approach has become the most influential method in researching mode differences.

In Section 2.2, the history of research on mode differences has been briefly introduced. Some studies published from the 1960s to the 1980s have been reviewed and the limitations of these early studies have been discussed. More recent methodological development in this research field has also been noted. To end this section, the following summary of research findings from multi-dimensional studies is provided:

1. Some dimensions are strongly associated with spoken and written differences; other dimensions have little or no relation to speech and writing. [For example, the dimensions found to be relevant to mode differences in Biber (1988) are Dimension 1 Involved versus Informational Production, Dimension 3 Explicit versus Situation-dependent Reference, and Dimension 5 Abstract versus Non-abstract Information.]

2. There are few, if any, absolute linguistic differences between spoken and written registers.
3. However, there are strong and systematic linguistic differences between stereotypical speech and stereotypical writing, that is, between conversation and written informational prose.
4. The spoken and written modes differ in their linguistic potential: They are not equally adept at accommodating a wide range of linguistic variation. (Biber & Vásquez, 2008, p. 537)

The last point in the above summary can be better understood in relation to the following quotation from Biber and Conrad (2009):

> spoken genres are surprisingly similar to one another in their typical linguistic characteristics, regardless of differences in communicative purpose, interactiveness, and pre-planning; but in contrast ... written registers have a wide range of linguistic diversity. Thus, there is a genuine difference between speech and writing, but it is one-directional: speakers are highly constrained in the kinds of discourse that they are (normally) able to produce. ... In contrast, writers have a much wider range of possibilities for the kinds of discourse that they can produce. (p. 262)

The information from Biber and Vásquez (2008) and Biber and Conrad (2009) has summarised the state-of-the-art knowledge concerning differences between the two modes of language production. In the present study, Biber (1988) has been adopted as the theoretical model explaining how native speakers of English vary their use of syntactic features in speech and writing. Details about the implementation of the multi-dimensional analysis and the research findings in Biber (1988) are elaborated in Chapter 3 Section 3.2.

2.3 Language Learners' Sensitivity to Mode Differences

Even though there is hardly any absolute difference between spoken and written English, competent language users are able to use both appropriately and to discern violations of the norm of usage. Language learners are thus expected to acquire this aspect of sociolinguistic competence. However, this is by no means an easy task. The variation across speech and writing poses difficulties to first language learners and second language learners alike. At first glance, it seems impossible for native speakers of English to experience problems in handling linguistic differences associated with the two modes. Nonetheless, as many have commented, spoken English and written English are better regarded as two distinct language systems. In order to become sociolinguistically competent, first language learners, as well as second language learners, have to be able to "recognise and be capable of responding to functional and conventional differences between speech and writing" (Winch & Gingell, 1994, p. 161). Consequently, the problems encountered by first language learners should not be overlooked. In this section, some studies on language learners' ability to cope with mode differences are reviewed. Section 2.3.1 focuses on research investigating first language learners. A model which explains how native speakers develop appropriate usage in writing is introduced. Two examples of studies examining oral features in first language learners' writing are also provided. Section 2.3.2 discusses research on second language learners. Studies investigating both non-Chinese and Chinese students are reported and their main findings are presented.

2.3.1 L1 Learners

Studies that are relevant to first language learners' sensitivity to mode differences are those researching the learners' literacy development. One major obstacle to native speakers' development of writing skills is interference from spoken English. First language learners acquire their mother tongue at home when they are still young children. What they acquire is the ability to communicate orally with those around them. Very often, they do not learn to write until they are sent to school. By then their speech development is well ahead of their writing development. Problems may arise if they cannot adjust their language to the new mode of communication. For example, they have to learn how to employ words and punctuation marks to express meanings which they used to convey through intonation and stress (Gumperz, Kaltman & O'Connor, 1984, p. 5; Krauthamer, 1999, p. 2).

Kroll (1981) has proposed a four-phase developmental model to explain native-speaking children's transition to proficient writers. In Phase 1, preparation, children begin by learning about the technical aspect of writing including handwriting and spelling. At this stage, their writing development is still in its infancy, but their oral English is already quite well developed (pp. 39, 41). In Phase 2, consolidation, the children's spoken English and written English are integrated. Forms from these two modes of production appear to be similar (pp. 39, 47). In Phase 3, differentiation, the children become more aware of the differences between speech and writing. They begin to master usage typical of English writing (p. 39). Finally, in Phase 4, systematic integration, the children have grown into competent language users who can use different linguistic features appropriately in different spoken and written contexts (pp. 39, 53). In Kroll's (1981) opinion, Phase 2 and

Phase 3 are most important to the teaching of writing in schools and the progression from Phase 2 to Phase 3 is not easy (pp. 40, 47).

Since the present study focuses on investigating second language learners, it is not my intention to provide a comprehensive review of literature on first language learners. However, two examples are cited to demonstrate that first language learners do encounter problems in acquiring the sociolinguistic competence necessary for using English appropriately in writing. These two studies are Krauthamer (1999) and Szpara and Wylie (2007). Krauthamer (1999) observed some spoken language features in her native-speaking students' writing. By consulting some literature on mode differences and on writing errors, the teacher researcher identified a list of features characterising speech, which she called *SLIPs* (Spoken Language Interference Patterns), and further classified them into erroneous ones and non-erroneous ones. Erroneous SLIPs are those oral usage patterns deviating from Standard Written English and hence are unacceptable in writing. Examples include inflectional deletions, double negatives and colloquialism (p. 49). Non-erroneous SLIPs include features like contractions, first person and second person pronouns, and questions (p. 49). In her book, Krauthamer argues that many students borrow usage from spoken English into written English. Nevertheless, her data, which consists of statistics from some writing error studies and her analysis of errors in one student essay, is not adequate in supporting her argument.

A line of research of first language literacy is dedicated particularly to native speakers of non-standard dialects. The greater distance between the English that they speak and the written English that are socially acceptable implies the greater effort they have to put in to produce language suitable for the written mode. Colombi and Schleppegrell (2002) even suggest that the challenge that speakers of non-standard

dialects face is no less than the challenge that second language learners face in developing target-language literacy (p. 5). An example of studies examining dialectal interference on native speakers' writing is Szpara and Wylie (2007). The researchers investigated the question whether African American and European American test takers showed linguistic differences in their written English. They analysed data from 18 texts produced by African Americans and 14 texts produced by European Americans in a national teacher assessment. All features deviating from Edited American English were identified and further classified. The researchers found that features associated with African American English, e.g. zero copula and the absence of *-s* morpheme in regular plural nouns, and Speech Code Errors, e.g. subject-verb discord and use of homophones, were present, but not frequent in their samples. They suggest that the formal, academic nature of their data may explain the low frequency of these non-standard features (p. 259). On the whole, a larger number of these features have been found in the African American teachers' writing than in the European American teachers' writing (p. 260). It can be shown in Szpara and Wylie (2007) that even adult native speakers may still produce oral-language features in writing occasionally and that those who speak a dialect different from the standard language produce more such features. Results from this study clearly indicate the difficulty in developing sensitivity to mode differences.

Despite the wealth of literature discussing how spoken language affects first language learners' writing, one should not take the relationship for granted. For one thing, writing errors are not necessarily caused by negative transfer from spoken English. As Johnstone (2002) points out, although the spoken variety that native speakers acquire at home may affect their writing performance, one must not ignore the effects of other factors, e.g. problematic educational policy and unsatisfactory

teacher quality (p. 173). Furthermore, the relationship between speech and writing is not necessarily uni-directional: Writing may affect speaking, too. Miller (1994) observes that research has shown differences between middle-class speakers' speech and lower-class speakers' speech. He speculates that middle-class speakers receive more education, which in turn leads to more exposure of written language, and hence their speech is more written-like (pp. 4301-4303).

2.3.2 L2 Learners

The previous section has shown that developing sensitivity to mode differences is not as simple a task to first language learners as one expects. Research on first language learners tends to investigate how speech, their more dominant language system, affects writing. For second language learners living in an English-speaking society, the development in oral English proficiency may precede the development in written English. Consequently, their case should be similar to that of the first language learners: It is highly likely that their spoken English interferes with their written English. However, for second language learners who do not live in an English-speaking environment and who receive English input mainly through formal education at school, i.e. what some researchers refer to as foreign language learners, the situation is far more complex. Very often, these learners learn both speaking and writing skills simultaneously. It is less predictable whether their speech influences writing or vice versa.

In this section, research on second language learners' sensitivity to mode differences is reviewed. Although comments about the learners' confusion over spoken and written usage can be found in research reports occasionally, these comments are merely presented as a possible interpretation for research findings (e.g.

Hyland & Milton, 1997, p. 192). There are, in fact, not many studies aiming at investigating how well second language learners handle differences between spoken and written English. In this section, only studies with a focus on mode differences in research design or data analysis are reported.

2.3.2.1 Non-Chinese Learners.

In an attempt to determine whether frequency counts of linguistic features can accurately reflect improvement in second language learners' writing, Shaw and Liu (1998) analysed writing from 164 students speaking 15 different mother tongues. The researchers asked the students to write an essay on the same topic before and after a summer course. Their pre-course essays total approximately 34,000 words and their post-course essays 37,000 words. Some of the features that Shaw and Liu (1998) counted were adopted from Biber (1988) and one of their research questions was to determine whether the students' writing changed from being more spoken-like to more written-like. Their major findings indicate that there is a significant decrease in the use of contractions and subordinate clauses of cause in the post-course essays. There is also a significant increase in the use of passives and connectors. Shaw and Liu (1998) suggest that the students' writing has become less spoken-like by the end of the course.

Hinkel (2002) examined 68 syntactic, lexical, clausal and rhetorical features, many of which were similar to features investigated in Biber (1988), in 1,457 English expository and argumentative essays produced by university students in America. She compared the use of these features by native speakers of American English and the use of these features by different groups of advanced second language learners including Chinese, Japanese, Korean, Vietnamese, Indonesian and

Arabic students. While Hinkel (2002) mainly discusses her findings from the contrastive rhetoric perspective and speculates how the different mother tongues affect the students' use of features in English, she does consider how sensitivity to mode differences is related to her findings. The researcher comments that:

> Both NS and NNS prose gravitates toward syntactically and lexically simple constructions usually marked by features of conversational and spoken genre, demonstrated by high frequency rates of vague nouns, universal pronouns, private verbs, *be*-copula verbs, amplifiers, and emphatics. (Hinkel, 2002, p. 74)

Even though these spoken-like features exist in both native-speaking students' and second language learners' essays, they occur much more frequently in the second language learners' writing (Hinkel, 2002, p. 74).

By using the framework of Biber (1988), van Rooy (2008) compared Black South African English with Standard English in order to provide evidence for the status of Black South African English as an English variety. He analysed 158 essays from the Tswana Learner English Corpus (TLE) and 110 essays from the Louvain Corpus of Native English Essays (LOCNESS). Both sets of data contain approximately 81,000 words written by university students. Van Rooy's (2008) analysis of individual features shows that some overused features in Black South African English including present tense, place adverbials, second person pronouns, causative adverbial subordinators, conditional adverbial subordinators and hedges indicate a more informal and involved style (p. 283). His analysis of textual dimension scores suggests that both corpora exhibit a higher level of involvement and informality than academic prose (p. 287). Although there is a lower Dimension 1 score, which represents a lower level of involvement, in the TLE data than in the LOCNESS data, the researcher notes that there are significantly more second person

pronouns in the TLE data than in the LOCNESS data and speculates that the TLE data is more interactive in nature (p. 287). In van Rooy's (2008) opinion, the Black South African English in TLE "shows a number of correspondences with spoken language, setting it apart from LOCNESS, but it does not become speech-like, it remains a written register in style, just systematically different from LOCNESS in a number of noteworthy respects" (p. 288). Van Rooy (2008) does not discuss Dimension 3 and Dimension 5 scores in relation to mode differences, but for the readers' reference, his results show smaller Dimension 3 and Dimension 5 scores in the TLE data than in the LOCNESS data. According to Biber's (1988) model, the TLE data contains more situation-dependent reference and less abstract information than the LOCNESS data does.

Whereas the first three studies introduced in this section, Shaw and Liu (1998), Hinkel (2002) and van Rooy (2008), discuss second language learners' sensitivity to mode differences only as a part of their analysis, the rest of the studies are devoted completely to investigating this aspect of sociolinguistic competence. Petch-Tyson (1998) is a study examining second language learners' sensitivity to one specific type of involvement: writer/reader visibility, i.e. "the way the participation of writer and reader are explicitly coded in the discourse" (p. 108). The writer/reader visibility devices that Petch-Tyson (1998) counted are first person pronouns, second person pronouns, fuzziness words, emphatic particles and reference to situation of writing/reading. Her data consists of argumentative essays in four different sub-corpora (French, Dutch, Swedish and Finnish) of the International Corpus of Learner English (ICLE) and in a corpus of native American English. Each set of language data contains more than 50,000 words. Her findings show that the Swedish learners use the most writer/reader visibility signals. The four groups of learners, in

descending order of writer/reader visibility in their writing, are the Swedish, the Finnish, the Dutch, and the French. At the same time, the four non-native groups all use many more writer/reader visibility signals than the native-speaking American students do. Petch-Tyson (1998) points out that the non-native learners' overuse of writer/reader visibility devices may make their writing "conform less to the conventions" (p. 116) and may generate a negative impression on the readers. However, the researcher is more interested in the differences observed among the different cultural groups than in the differences observed between the native speakers and the second language learners.

Cobb (2003) and McCrostie (2008) are both replications of Petch-Tyson (1998). Cobb's (2003) data is made up of more than 400 expository essays produced by advanced second language learners of English in Québec. His study focuses only on the learners' use of first person and second person pronouns. Results show that there are a total of 6.47 first person and second person pronouns per 100 words in the Québec learners' writing. This figure is much higher than the figure for European French learners (2.04 per 100 words) and the figure for native speakers (0.89 per 100 words) reported in Petch-Tyson (1998). Another study, McCrostie (2008), involves an examination of writer/reader visibility devices in 333 argumentative essays written by first-year and second-year English majors in Japan. His findings demonstrate that both groups of Japanese students employ many more writer/reader visibility devices than the native-speaking students in Petch-Tyson (1998) do. However, when compared with the first-year students, the second-year students use far fewer writer/reader visibility devices, especially first person and second person pronouns. McCrostie (2008) suggests that the university writing courses are one factor leading to the dramatic decrease in the students' use of writer/reader visibility

devices (p. 106).

Ädel (2008) proposes four possible reasons for over-involvement in second language learners' writing: (a) the learners' unsatisfactory knowledge of register, (b) interference from the learners' first-language cultural conventions, (c) general writing strategies adopted by the learners, and (d) different conditions associated with the writing task (pp. 40-41). In Ädel (2008), the researcher investigated how the last factor could affect second language learners' use of involvement features in writing. Her study involves a comparison of timed and untimed essays and a comparison of essays with and without access to secondary sources. Her data includes (a) a 87,000-word corpus of timed essays from the Swedish subcorpus of ICLE (SWICLE), (b) a 70,000-word corpus of untimed essays from SWICLE, and (c) a 50,000-word corpus of untimed essays from the Uppsala Student Essay corpus. The first two sets of the data represent essays with no access to secondary sources and the last set represents essays with such access. Ädel (2008) has generated some interesting findings: There are significantly more first person singular pronouns in timed essays than in untimed essays. There are significantly more disjuncts, questions and exclamations in essays without access to secondary sources than in essays with such access. The researcher concludes that "the overuse of involvement features cannot just be attributed to lack of register awareness" (Ädel, 2008, p. 46) and that "given the right circumstances, Swedish learners are able to produce relatively successfully pieces of argumentative discourse" (Ädel, 2008, p. 46).

In a lexicographical project, Gilquin and Paquot (2008) examined some academic discourse functions such as comparing and contrasting, exemplifying and expressing personal opinion in native and non-native English. They analysed 350 lexical items in the 15-million-word written academic component of the British

National Corpus (BNC), the 10-million-word spoken component of BNC, and the 1.5-million-word ICLE with untimed argumentative essays from 14 different subcorpora. Gilquin and Paquot (2008) observed "a strong tendency among learners, regardless of their mother tongue, to use spoken-like features in their written production" (p. 45). They gave four examples to illustrate that the learners overused items common in spoken English: (a) the adverbs *perhaps* and *maybe*; (b) amplifying adverbs including *really, of course, certainly, absolutely* and *definitely*; (c) phrases showing writer visibility, e.g. *it seems to me* and *I would like/want/am going to talk about*; and (d) the sentence-initial *and*. More importantly, Gilquin and Paquot (2008) discuss four possible explanations for the learners' use of spoken-like features in writing (pp. 51-57). The first explanation attributes the spoken-like features to the learners' transfer of oral strategies to their writing, but this explanation applies only to those who learn English through oral interaction (i.e. ESL students), but not those who learn English through formal instruction (i.e. EFL students). The second explanation concerns the transfer of register knowledge from the learners' mother tongues. The third explanation suggests that the learners' use of spoken-like features is a result of the teaching practice, e.g. the presentation of different connectors as synonyms with no differentiation of register. The last explanation is developmental factors: The learners, as novice writers, have not yet fully developed their knowledge about writing. As their writing proficiency increases, their use of spoken-like features will decrease. Gilquin and Paquot (2008) believe that the above explanations are not mutually exclusive and future research should not ignore the possible interaction among these different factors (pp. 51, 57).

2.3.2.2 Chinese Learners.

In Chao's (2003) study on Taiwanese students' writing, he examined influence from spoken language as one of his analyses. The researcher collected English argumentative essays from four levels of Taiwanese students: 191 Grade 10 students, 175 Grade 12 students, 104 second-year English majors and 47 third-year English majors. He counted the following features of oral language use in the students' writing: self-directed questions, questions directed to readers, questions using *you*, declarative sentences ending with a question mark, exclamations, imperatives, interjections, spoken-like narration and spoken connectors (*because, so, and, but*) used at the beginning of a sentence. Chao's (2003) results show that the spoken-like features are used significantly differently across the four groups of students. In particular, there is a consistent, decreasing trend in the use of spoken connectors from the Grade 10 students to the third-year English majors. Whereas the use of spoken connectors correlates negatively with the scores of the essays, other spoken-like features as a whole correlate positively with the scores. The researcher suggests that second language learners' writing development parallels first language literacy development and that his findings match the prediction from Kroll's (1981) model (Chao, 2003, p. 160).

Hyland (2004) investigated the use of dialogic features in Hong Kong students' academic writing. He compared a 630,000-word corpus consisting of 64 final-year project reports written by university students from eight departments with a 1.3-million-word corpus consisting of 240 research articles published in the corresponding academic fields. He found that there were fewer engagement features in the student reports (Total = 23.9 per 10,000 words) than in the published articles (Total = 53.5 per 10,000 words). The same pattern existed in each of the engagement

features: questions, inclusive pronouns, second person pronouns, directives, shared knowledge and asides. He also noticed that students from different disciplines employed engagement features differentially. In general, students majoring in Information System employed the most engagement features (35.9 per 10,000 words) and those majoring in Marketing employed the fewest (13.2 per 10,000 words). The researcher explains that the observed differences largely reflect the different writing styles valued in different disciplines (p. 11). Hyland's (2004) finding that Hong Kong students use fewer first person plural pronouns and second person pronouns in writing is very different from the finding from Petch-Tyson (1998), Hinkel (2002), Cobb (2003) and McCrostie (2008) that second language learners of English tend to employ more first person and second person pronouns in writing. One reason for this discrepancy is that Hyland's (2004) data represents some technical writing in the students' academic fields, whereas the other four studies are based on expository and argumentative essays of a more general nature.

Littlewood and Li (2006) conducted a study to determine how sensitive Chinese learners of English were to mode differences in English. The researchers gave the participants, 100 non-English majors in Hong Kong as well as 50 English majors and 50 non-English majors in Mainland China, 12 pairs of sentences and asked them to decide which sentence in each pair was taken from spoken language and which from written language. The 12 matched pairs of sentences were derived from Li's written handout for audience in an academic conference and her oral presentation in the same conference. In order to ensure that the linguistic forms used appealed to native speakers' intuition of the two modes, Littlewood had applied the same test to a group of native speakers. He found that the native speakers were able to judge correctly whether the sentences came from a written source or a spoken source. In Littlewood

and Li (2006), the Hong Kong students, the English majors in Mainland and the non-English majors in Mainland obtained an average score of 10.6, 10.4 and 9.7 out of 12 respectively. The researchers suggest that the scores reflect the amount of authentic English exposure available to the different groups of students: Hong Kong, as a westernised city and a former British colony, allows local students more opportunities to get English exposure, so Hong Kong students are generally more familiar with English than their mainland counterparts. Due to the nature of their study, English majors in Mainland are exposed to more English materials than other students in Mainland. In addition to the score distribution patterns among the three groups of learners, Littlewood and Li (2006) also examined the score distribution patterns of Hong Kong participants with different English grades in the public examination. There is a general tendency that students with poorer results in the public examination perform less well in the judgement test, but the trend that students with better results perform better cannot be clearly shown, because of the relatively small number of participants obtaining higher grades in the public examination.

2.3.2.3 Observations on studies of L2 learners.

Eleven different studies have been reviewed in Section 2.3.2.1 and Section 2.3.2.2. Seven of them (Ädel, 2008; Cobb, 2003; Gilquin & Paquot, 2008; Hyland, 2004; Littlewood & Li, 2006; McCrostie, 2008; Petch-Tyson, 1998) have been deliberately designed to research second language learners' sensitivity to mode differences. The other four represent studies investigating second language writing, contrastive rhetoric or a new English variety, and the analysis on sensitivity to mode differences is merely a secondary focus. Regarding the seven studies devoted to

examining second language learners' sensitivity to mode differences, five studies (Ädel, 2008; Cobb, 2003; Hyland, 2004; McCrostie, 2008; Petch-Tyson, 1998) have investigated linguistic features related to the concept of involvement.

Taking into account all the 11 studies researching sensitivity to mode differences as the primary or the secondary focus, university students appears to be the most popular target group of investigation. Nearly all researchers have chosen to study university students. The only exceptions are Cobb (2003), whose subjects are applicants for university programmes, and Chao (2003), whose data involves writing from both high school students and university students. All studies except Littlewood and Li (2006) have examined spoken-like features in second language learners' writing and seven of these studies (Ädel, 2008; Chao, 2003; Cobb, 2003; Gilquin & Paquot, 2008; Hinkel, 2002; McCrostie, 2008; Petch-Tyson, 1998) have adopted expository and argumentative essays as the data for analysis. Overall, only three studies (Chao, 2003; McCrostie, 2008; Shaw & Liu, 1998) address the developmental aspect of the learners' sensitivity to mode differences and only two studies (Ädel, 2008; Gilquin & Paquot, 2008) provide a relatively detailed discussion on factors contributing to the spoken-like nature of second language learners' writing.

It can be observed from the literature review in Section 2.3.2.1 and Section 2.3.2.2 that current knowledge on second language learners' sensitivity to mode differences is rather limited. Researchers interested in this area tend to focus narrowly on involvement features but ignore other linguistic differences between spoken and written English. They only consider how spoken-like the students' writing is, but not how written-like the students' speech is. The researchers' source of language data is mainly restricted to expository and argumentative essays written by

university students. Therefore, in order to understand more about this particular aspect of second language learners' sociolinguistic competence, more empirical studies which address the neglected areas in literature have to be carried out.

2.4 Use of English Syntactic Features by Hong Kong Learners

Section 2.2 and Section 2.3 have reviewed literature from the two research areas most relevant to the topic of investigation of the present study: Section 2.2 has discussed research on linguistic differences between spoken and written English and Section 2.3 has reported studies examining second language learners' sensitivity to mode differences. Since the present study investigates Hong Kong second language learners' sensitivity to mode differences through their use of syntactic features, it is desirable to provide the readers with some information about Hong Kong students' use of various syntactic features. In this section, studies on Hong Kong students' use of English syntactic features are reviewed. Section 2.4.1 summarises some studies focusing on error analysis, which is the dominant research approach in the literature. Section 2.4.2 introduces studies adopting the same corpus-based approach as the present study and examining the students' overuse and underuse patterns. Emphasis is placed on studies counting features similar to those in the present study.

2.4.1 Studies on Errors

Most studies on Hong Kong students' use of syntactic features, especially the early studies, tend to focus on the error patterns in the learners' English (e.g. Bunton, 1991; A. Y. W. Chan, 2004a, 2004b; B. Chan, 1991; Green, 1991; Webster & Lam, 1991; Webster, Ward & Craig, 1987; V. Yip & Matthews, 1991). The researchers often use written data, the sources of which may include sentence translation tasks

(e.g. A. Y. W. Chan, 2004b), free compositions (e.g. A. Y. W. Chan, 2004a; B. Chan, 1991; Green, 1991) and academic essays (e.g. V. Yip & Matthews, 1991). In addition to these primary data sources, researchers sometimes consult secondary sources such as teacher questionnaires (e.g. Webster, Ward & Craig, 1987) and reference books on common writing errors (e.g. Bunton, 1991; Webster, Ward & Craig, 1987). The target groups of investigation range from junior secondary school students (e.g. A. Y. W. Chan, 2004a, 2004b), senior secondary school students (e.g. Bunton, 1991; A. Y. W. Chan, 2004a, 2004b; B. Chan, 1991; Green, 1991) to tertiary students (e.g. A. Y. W. Chan, 2004a, 2004b; V. Yip & Matthews, 1991).

Quite a number of studies on Hong Kong students' syntactic errors explain the students' problems in terms of negative transfer from the students' mother tongue. For example, Webster, Ward and Craig (1987) suggest that the learners' errors in tense and aspect are likely to be caused by the differences between the Cantonese verb system and the English one (p. 69). A. Y. W. Chan (2004b) considers interference from Chinese to be the main reason for Hong Kong students' syntactic errors such as the omission of the copular *be* after modal verbs, the use of *very* before verbs, and the substitution of *have* for *be* in existential constructions (p. 66). Besides research articles, there are also a few books which list Hong Kong students' common grammatical errors in a dictionary-like format (e.g. Boyle & Boyle, 1991; Bunton, 1989). All these different books and articles on errors provide information about how Hong Kong students' use of syntactic features deviates from the prescriptive rules of the English language.

2.4.2 Studies on Overuse and Underuse

Recent studies on Hong Kong students' use of syntactic features (e.g. Bolton, Nelson & Hung, 2002; Fung & Carter, 2007; Hyland & Milton, 1997; Milton, 2001) have begun to compare the students' usage with native speakers' usage in similar language corpora. These studies present a new perspective for understanding Hong Kong students' English by examining the students' overuse and underuse of syntactic features. Since the present study also includes the same kind of analysis, the results of these studies are particularly relevant to the present study. In the rest of this section, findings of these studies are reported.

Fung and Carter (2007) is the only study in this section that is based on spoken data. It compares the use of discourse markers in a Hong Kong learner corpus, a 14-thousand-word corpus of group discussions from 49 Form 6 students, with a native speaker corpus, a 460-thousand-word pedagogical sub-corpus from the Cambridge and Nottingham Corpus of Discourse in English (CANCODE). Its results reveal a discrepancy between the two groups' usage. In general, the learners use less frequently the discourse markers that the native speakers usually use (pp. 426-427). In particular, the Hong Kong students overuse *I think, yes, but,* and *because*, but underuse *and, right, yeah, well, so, now, sort of, you know, actually, see, say,* and *cos* (p. 429). According to Fung and Carter (2007), the learner group usually employs discourse markers to perform referential and structural functions, but less so often for interpersonal function (p. 435). In other words, the students tend to use discourse markers to mark textual relationships, e.g. cause and contrast, and discourse structure, e.g. transitions and sequences, but they are less likely to use discourse markers for affective and social functions, e.g. marking shared knowledge and attitudes.

Hyland and Milton (1997) investigated the use of epistemic modality in two 0.5-million-word corpora of public examination scripts, one produced by Hong Kong secondary school leavers and one by their British counterparts. The researchers counted 75 common epistemic devices including modal verbs (e.g. *could* and *may*), adjectives (e.g. *always* and *certain*), nouns (e.g. *claim* and *doubt*), lexical verbs (e.g. *appear* and *predict*) and adverbs (e.g. *almost* and *possibly*). The findings show that both groups of students employ a similar number of devices to express epistemic modality, but the non-native students use a narrower range of items (pp. 188-189). Whereas the British students show equal preference for adverbs and modal verbs, the Hong Kong students use twice as many modal verbs as adverbials (pp. 189-190). Half of the epistemic devices in the Hong Kong student data mark certainty; in contrast, only one-third of the devices in the native student data are used for such a purpose (p. 193). According to Hyland and Milton (1997), this finding confirms the view that second language learners of English tend to make stronger assertions in writing than native speakers do (p. 193). Further analysis suggests that Hong Kong students of a higher proficiency level employ epistemic devices in a more similar manner to the native students' usage. They rely less on modal verbs and employ fewer certainty markers than students of a lower proficiency level (pp. 190-191, 195).

Observing some methodological problems in literature, Bolton, Nelson and Hung (2002) conducted a study on Hong Kong students' use of connectors by making the following decisions: (a) devising the list of connectors directly from a corpus of academic writing, (b) calculating the ratio of occurrence in terms of frequency per sentence, and (c) comparing the student data with a corpus of academic writing. Their data involves 20 samples of Hong Kong undergraduate

students' writing (around 46,000 words) in the Hong Kong component of the International Corpus of English (ICE-HK), 20 samples of British undergraduate students' writing (around 43,000 words) in the British component of the International Corpus of English (ICE-GB) and 40 samples of published academic writing in ICE-GB. It is found that Hong Kong students use a smaller number of different connectors when compared with expert academic writers (p. 175). At the same time, they overuse some connectors. *So*, *and*, *also*, *thus*, *but*, *therefore*, *moreover*, *then*, *on the other hand*, and *in fact* are among the most seriously overused connectors by Hong Kong students (p. 177).

The last study reviewed in this section, Milton (2001), is a large-scale study set out to investigate Hong Kong students' written interlanguage. It is very similar to the present study in the sense that a large number of linguistic features are considered. Whereas the present study examines syntactic features, Milton (2001) focuses on grammatical classes. Milton (2001) involves a comparison of a corpus of Hong Kong students' writing, which consists of 0.5 million tokens of grade-D and grade-E essays in a public examination (AS-level Use of English Examination), with a corpus of British students' writing, which contains 0.5 million tokens of grade-A and grade-B essays in a public examination (A-level General Studies Examination). The researcher tagged both sets of data by using the CLAWS7 (Constituent Likelihood Automatic Word-tagging System) part-of-speech tagger and determined cases of significant overuse and underuse by calculating log likelihood. He reported 43 cases of overuse and 55 cases of underuse. Some of the overused and underused word classes that may be relevant to the syntactic features counted in the present study are listed in Table 1.

Table 1

Some of the Overused and Underused CLAWS Word Classes in Hong Kong Students' Writing (Milton, 2001, pp. 6-8)

Overuse	Underuse
Plural common nouns	Singular common nouns
Base form of lexical verbs	Past participles of lexical verbs
Infinitives	*-s* form of lexical verbs
you	General prepositions
they	*of*
Modal auxiliaries	Past tense of lexical verbs
we	*with*
General adjectives	Adverbs introducing apposition
I	Prepositional adverbs/ particles
them	Singular proper nouns
Common nouns neutral for number	Plural organizational nouns
us	Plural temporal nouns
not, n't	Locative adverbs
for	Singular locative nouns
Singular temporal nouns	Superlative adjectives
me	Comparative adjectives
Subordinating conjunctions	Singular indefinite pronouns
themselves	Adverbs of time
General adverbs	*he, she*
Comparative general adverbs	*him, her*

As evident in the review in Section 2.4, there is very little research on Hong Kong students' use of syntactic feature in spoken English, but there is a reasonable amount of research on their use of syntactic features in written English. Some of the early publications on Hong Kong students' use of syntactic feature rely solely on the teachers' intuition and findings are discussed not with reference to statistics but with reference to illustrative examples. Despite the fact that many studies have adopted the more traditional approach and focused on the learners' errors, there emerges a

new trend that researchers investigate how learners' usage deviates from native speakers' usage in terms of overuse and underuse patterns. These studies on overuse and underuse further promote our understanding of Hong Kong students' use of syntactic features and their findings are especially relevant to the present study.

2.5 Summary

Differences between spoken and written English have been extensively researched. Interesting qualities have been identified for both speech (e.g. fragmentation and involvement) and writing (e.g. integration and detachment). Due to the methodological limitations of early research, no clear conclusion on mode differences can be drawn. However, thanks to the advance in computer technology and the development of corpus linguistics, relatively recent studies begin to overcome the methodological obstacles. In particular, studies adopting the multi-dimensional approach have generated some promising results. Research on native speakers' literacy development and writing skills throws light on the problems that first language learners encounter in developing sensitivity to mode differences. The dominance of oral communication may make it difficult for native-speaking writers, especially the beginner writers, to conform to the written norms when necessary. Regarding research on second language learners, although comments about the learners' confusion over spoken and written usage can be found occasionally, studies which investigate the topic of sensitivity to mode differences as a central theme are limited. Existing studies tend to examine spoken-like features in the learners' writing and the spoken-like features investigated are often restricted to those related to involvement. There is a need for research extending beyond this narrow scope of investigation. Studies on Hong Kong students' use of English

syntactic features are based primarily on written data. Very little is known about their use of syntactic features in spoken English. Whereas early studies on the learners' use of syntactic features only analyse grammatical errors, some recent studies pay attention to how students' usage deviates from native speakers' usage in terms of overuse and underuse. This kind of corpus-based analysis, though still a minority in relevant literature, is particularly important in promoting understanding of relatively advanced Hong Kong learners' English.

CHAPTER 3 METHODOLOGY

3.1 Introduction

As suggested in Chapter 2, research on second language learners' sensitivity to mode differences is limited in scope: Most studies focus on the identification of involvement features in university students' writing. However, the variation across speech and writing is very complex. Simple observation on a few linguistic features cannot possibly generate conclusive evidence. Furthermore, second language learners' confusion about mode differences may manifest itself not only in spoken-like features in writing, but also in written-like features in speech. In view of the limitations of existing studies, the present study aims at determining second language learners' sensitivity to mode differences by examining the use of 67 syntactic features in Hong Kong students' spoken and written English. In this chapter, details of the research methodology are unfolded. The description begins with a summary of Biber's (1988) model on differences between spoken and written English, which is the theoretical framework underlying the present study. Next, there is an outline of research design. What follows is a profile of language data employed in the present study, together with an account of the data processing and analysis procedures. The chapter ends with some methodological issues noticed during the pilot study.

3.2 Theoretical Framework

To determine whether Hong Kong learners use English in a manner appropriate for a specific mode, a model which specifies native-speaker usage is required. In the present study, Biber's (1988) multi-dimensional model has been adopted for this

purpose. Although D. Lee (2005) has treated Biber (1988) with skepticism, Biber (1988) appears to be the best available model for the present study. As mentioned in Chapter 2 Section 2.2, the multi-dimensional approach adopted in Biber (1988) helps overcome the methodological limitations common in early studies on mode differences. When compared with Biber's other multi-dimensional studies on spoken and written English such as Biber (2001), which focuses on eighteenth-century English, and Biber (2006), which examines academic English in the university context, Biber (1988), which addresses a wide range of genres in modern British English, better serves the purpose of the present study. Moreover, Biber (1988) is based on relatively frequent syntactic features. This is an advantage over two other multi-dimensional models on contemporary spoken and written English, Louwerse et al. (2004), which draw on discourse features, and Crossley and Louwerse (2007), which consider collocational features, because both of them demand much longer texts for analysis. The descriptions of these two models are also less detailed than Biber (1988). Given the above reasons, Biber (1988) has been chosen for the theoretical native-speaker model in the present study.

Biber's (1988) model was constructed from an extensive survey of the frequencies of 67 syntactic features in 23 common English genres, the texts of which totalled 960,000 words. Most of his written data was sampled from the Lancaster-Oslo-Bergen Corpus of British English (the LOB Corpus) and all of his spoken data was obtained from the London-Lund Corpus of Spoken English. In other words, his model represents how native speakers of British English use syntactic features in oral and written communication. In the rest of this section, information about Biber's (1988) methodology and the resultant model is provided so as to set the scene for the subsequent discussion of research findings in this thesis.

3.2.1 Method

The research procedures of Biber (1988) can be summarised as follows: First, all linguistic features potentially relevant to the distinction between spoken and written English were shortlisted from a review of literature. Second, texts were sampled to represent various communicative situations in both speech and writing. Third, the sampled texts were tagged for part-of-speech information by a computer programme. Fourth, automatic searching for the selected features was performed and their frequencies per text were calculated. Fifth, a factor analysis was carried out to identify co-occurring linguistic features and the resultant factors were interpreted as different textual dimensions according to the common communicative functions of the co-occurring features. Finally, dimension scores for each text were calculated from the frequencies of the co-occurring features and different genres were compared with reference to their average dimension scores (Biber, 1988, pp. 63-64).

The 23 genres of texts sampled in Biber (1988) have been listed in Table 2. There are 17 genres of written discourse. Except the two letter genres, all were taken from the LOB Corpus (Biber, 1988, pp. 66-67). These 15 genres are comprised of 324 texts, each of which contains 2,000 words. The texts were sampled from different domains including newspapers (e.g. press reportage), government institutions (e.g. official documents), educational settings (e.g. academic prose) and leisure reading (e.g. general fiction). The personal letters and the professional letters were compiled by Douglas Biber and William Grabe respectively. There are altogether 16 letters in Biber's (1988) written data. In addition to the 17 written genres, six genres of spoken discourse were included in Biber's (1988) analysis. The 141 transcripts of these six genres were all taken from the London-Lund Corpus of Spoken English (Biber, 1988, pp. 66-67). Their lengths range from 400 to 2500

words (Biber, 1988, p. 208). The genres represent communicative situations in which the audience is present in front of the speakers (e.g. face-to-face conversation) and situations in which the audience is not present (e.g. broadcast), as well as situations involving individual conversational partners (e.g. telephone conversation) and situations involving a large group of audience (e.g. prepared speeches).

Table 2

Genres of Texts Used in Biber (1988) (p. 67)

Written Genres	1. press reportage
	2. editorials
	3. press reviews
	4. religion
	5. skills and hobbies
	6. popular lore
	7. biographies
	8. official documents
	9. academic prose
	10. general fiction
	11. mystery fiction
	12. science fiction
	13. adventure fiction
	14. romantic fiction
	15. humour
	16. personal letters
	17. professional letters
Spoken Genres	18. face-to-face conversation
	19. telephone conversation
	20. public conversations, debates and interviews
	21. broadcast
	22. spontaneous speeches
	23. prepared speeches

Table 3

Syntactic Features Counted in Biber (1988) (pp. 73-75)

Tense and aspect markers	1. past tense
	2. perfect aspect
	3. present tense
Place and time adverbials	4. place adverbials
	5. time adverbials
Pronouns and pro-verbs	6. first person pronouns
	7. second person pronouns
	8. third person personal pronouns (excluding *it*)
	9. pronoun *it*
	10. demonstrative pronouns
	11. indefinite pronouns
	12. pro-verb *do*
Questions	13. direct *wh*-questions
Nominal forms	14. nominalizations
	15. gerunds
	16. total other nouns
Passives	17. agentless passives
	18. *by*-passives
Stative forms	19. *be* as main verb
	20. existential *there*
Subordination features	21. *that* verb complements
	22. *that* adjective complements
	23. *wh* clauses
	24. infinitives
	25. present participial clauses
	26. past participial clauses
	27. past participial *whiz* deletion relatives
	28. present participial *whiz* deletion relatives
	29. *that* relative clauses on subject position
	30. *that* relative clauses on object position
	31. *wh* relatives on subject position
	32. *wh* relatives on object position
	33. pied-piping relative clauses
	34. sentence relatives
	35. causative adverbial subordinators
	36. concessive adverbial subordinators
	37. conditional adverbial subordinators
	38. other adverbial subordinators
Prepositional phrases, adjectives, and adverbs	39. total prepositional phrases
	40. attributive adjectives
	41. predicative adjectives
	42. total adverbs
Lexical specificity	43. type/token ratio
	44. mean word length

Table 3 (contd)

Lexical classes	45. conjuncts
	46. downtoners
	47. hedges
	48. amplifiers
	49. emphatics
	50. discourse particles
	51. demonstratives
Modals	52. possibility modals
	53. necessity modals
	54. predictive modals
Specialized verb classes	55. public verbs
	56. private verbs
	57. suasive verbs
	58. *seem* and *appear*
Reduced forms and dispreferred structures	59. contractions
	60. subordinator *that* deletion
	61. stranded prepositions
	62. split infinitives
	63. split auxiliaries
Coordination	64. phrasal coordination
	65. independent clause coordination
Negation	66. synthetic negation
	67. analytic negation

Table 3 shows the 67 syntactic features counted in Biber (1988). They belong to 16 different categories. In general, Biber (1988) defines these syntactic features with reference to Quirk, Greenbaum, Leech and Svartvik (1985). One may notice that not all features are grammatical ones; some, e.g. Feature 43 type/token ratio and Feature 44 mean word length, are lexical features. For convenience, the term *syntactic features* is used across this thesis to refer to the 67 linguistic features in general.

Due to the vast amount of language data and the large number of features concerned, the counting of syntactic features in Biber (1988) could not be made purely manually. A computer was used to automate the searching process (Biber, 1988, pp. 64, 211). Although the use of computer could increase efficiency of the

searching process and ensure reliability of the search results, the computer was not as flexible as the human brain. As a result, the search of the features could not be exhaustive. Appendix A outlines the basic principles involved in the computer searching algorithms. The table is a summary of the information in Biber's (1988) Appendix II. Due to the constraints of the computer search, the identification of syntactic features was restricted to instances that were typical and rule-governed. For example, regarding the search for demonstrative pronouns (Feature 10) in written data, only those followed by a verb, a clause punctuation mark, a *wh* pronoun or the word *and* could be found. Other instances (e.g. *I saw this yesterday*) were ignored. Since a lot of time was required to edit the computer search results, the researcher tried to minimise the chance of his disambiguating multi-functional words, which could be very time-consuming. For example, regarding the search for place adverbials (Feature 4), words like *in* and *on*, which were commonly found in place adverbials (e.g. *in the room* and *on the table*) but which were also commonly used for other functions, were excluded from the algorithm altogether. From the above discussion, it becomes evident that Biber's (1988) search method did not include every instance of the syntactic features. Therefore, it is necessary to summarise the algorithms so that, when interpreting the results of the present study, the readers can understand what has and what has not been counted in each feature. However, in view of its length, the summary is put in Appendix A so as not to obstruct the flow of the texts in the thesis.

3.2.2 Major Components in the Model

The major components of Biber's (1988) model are six textual dimensions identified from the factor analysis of frequencies of the 67 syntactic features. (Originally seven factors were proposed in his factor analysis, but the last factor was dropped due to the relatively low factor loadings of its features.) The six textual dimensions were interpreted with reference to the common communicative functions of the co-occurring syntactic features as well as the distribution of the different genres along the dimensions. These six dimensions are:

- Dimension 1 Involved versus Informational Production
- Dimension 2 Narrative versus Non-narrative Concern
- Dimension 3 Explicit versus Situation-dependent Reference
- Dimension 4 Overt Expression of Persuasion
- Dimension 5 Abstract versus Non-abstract Information
- Dimension 6 On-line Informational Elaboration

In Biber's (1988) model, these underlying qualities of English discourse are quantified and presented as continua.

To understand each textual dimension in Biber (1988), the syntactic features grouped under the dimension as well as the distribution of genres along the dimension have to be taken into consideration (pp. 101-114, 128-160). The syntactic features are the surface linguistic devices demonstrating the underlying quality; the distribution of genres contextualises the quality in terms of real-life communicative situations. Tables 4-9 and Figures 4-9 are reproduced from Biber (1988) to show the syntactic features loaded on each dimension and the distribution of genres in each dimension respectively. Dimension 1 is the textual dimension with the most syntactic features loaded on it: There are 23 positive features and five negative features. The

positive end of Dimension 1 represents not only an affective and interactional language style, but also a style of real-time language use characterised by generalised content, fragmented structures and reduced forms. Typical examples of English genres with very high Dimension 1 scores are telephone conversations and face-to-face conversations. In contrast, the negative end of Dimension 1 indicates highly informational language, as well as carefully edited language with precise vocabulary and integrated structure. Exemplars of genres with very low Dimension 1 scores include official documents and academic prose (see Table 4 and Figure 4).

Table 4

Syntactic Features Underlying Dimension 1 Involved Versus Informational Production in Biber (1988) (p. 102)

Positive Features	Negative Features
Private verbs	Nouns
That deletion	Word length
Contractions	Prepositions
Present tense verbs	Type/token ratio
Second person pronouns	Attributive adjectives
Do as pro-verb	
Analytic negation	
Demonstrative pronouns	
General emphatics	
First person pronouns	
Pronoun *it*	
Be as main verb	
Causative subordination	
Discourse particles	
Indefinite pronouns	
General hedges	
Amplifiers	
Sentence relatives	
Wh questions	
Possibility modals	
Clause coordination	
Wh clauses	
Final prepositions	

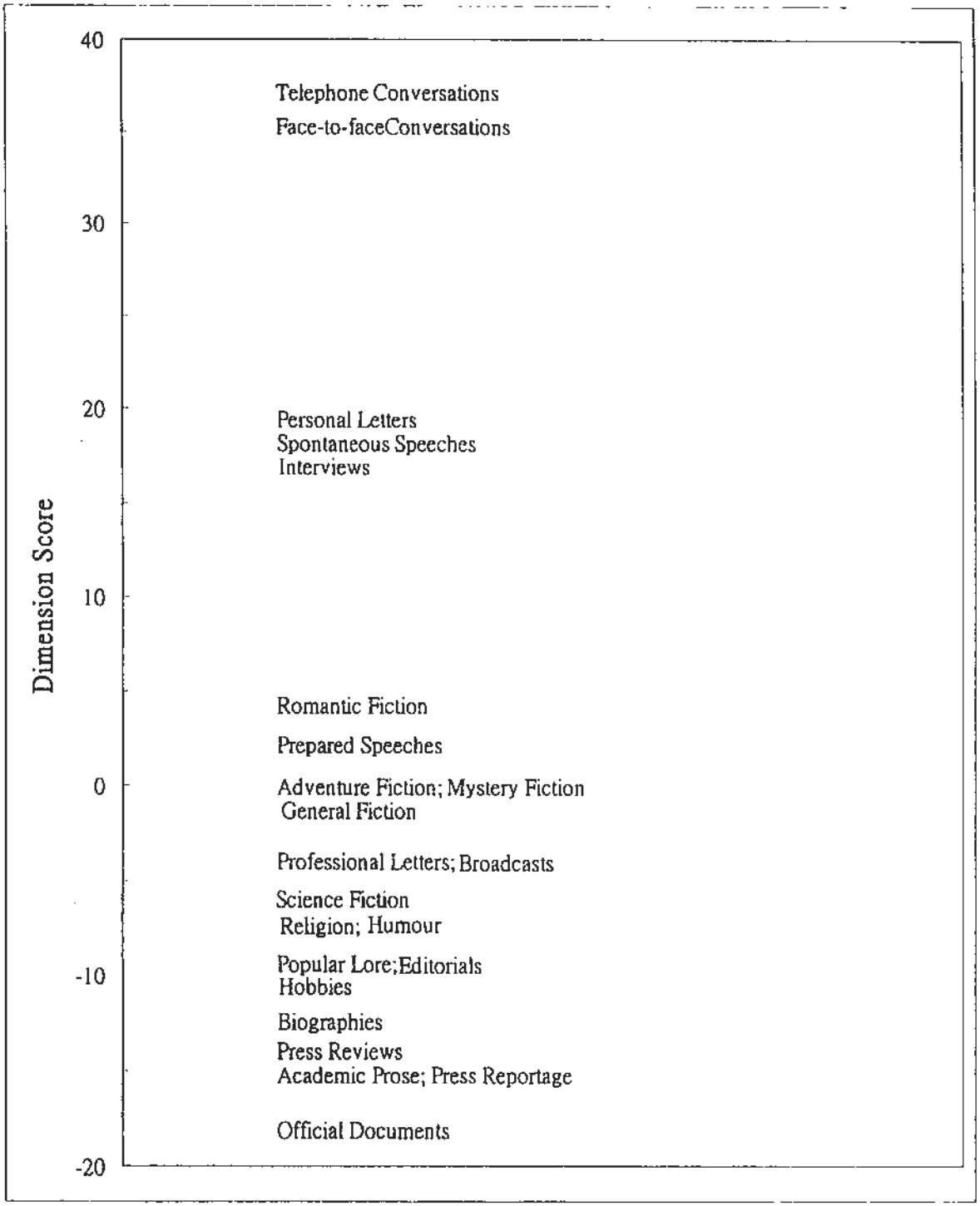

Figure 4. Distribution of the mean scores of different genres along Dimension 1 Involved versus Informational Production (Biber, 1988, p. 128)

Six positive features are loaded on Dimension 2. The positive end of Dimension 2 distinguishes narrative discourse, i.e. discourse marked by past actions, third person animate referents and reported speech, from other types of discourse. Genres with very high Dimension 2 scores are all fiction; genres with very low Dimension 2 scores include radio broadcasts and official documents (see Table 5 and Figure 5).

Table 5

Syntactic Features Underlying Dimension 2 Narrative Versus Non-narrative Concern in Biber (1988) (p. 102)

Positive Features	Negative Features
Past tense verbs	-- no negative features --
Third person personal pronouns	
Perfect aspect verbs	
Public verbs	
Synthetic negation	
Present participial clauses	

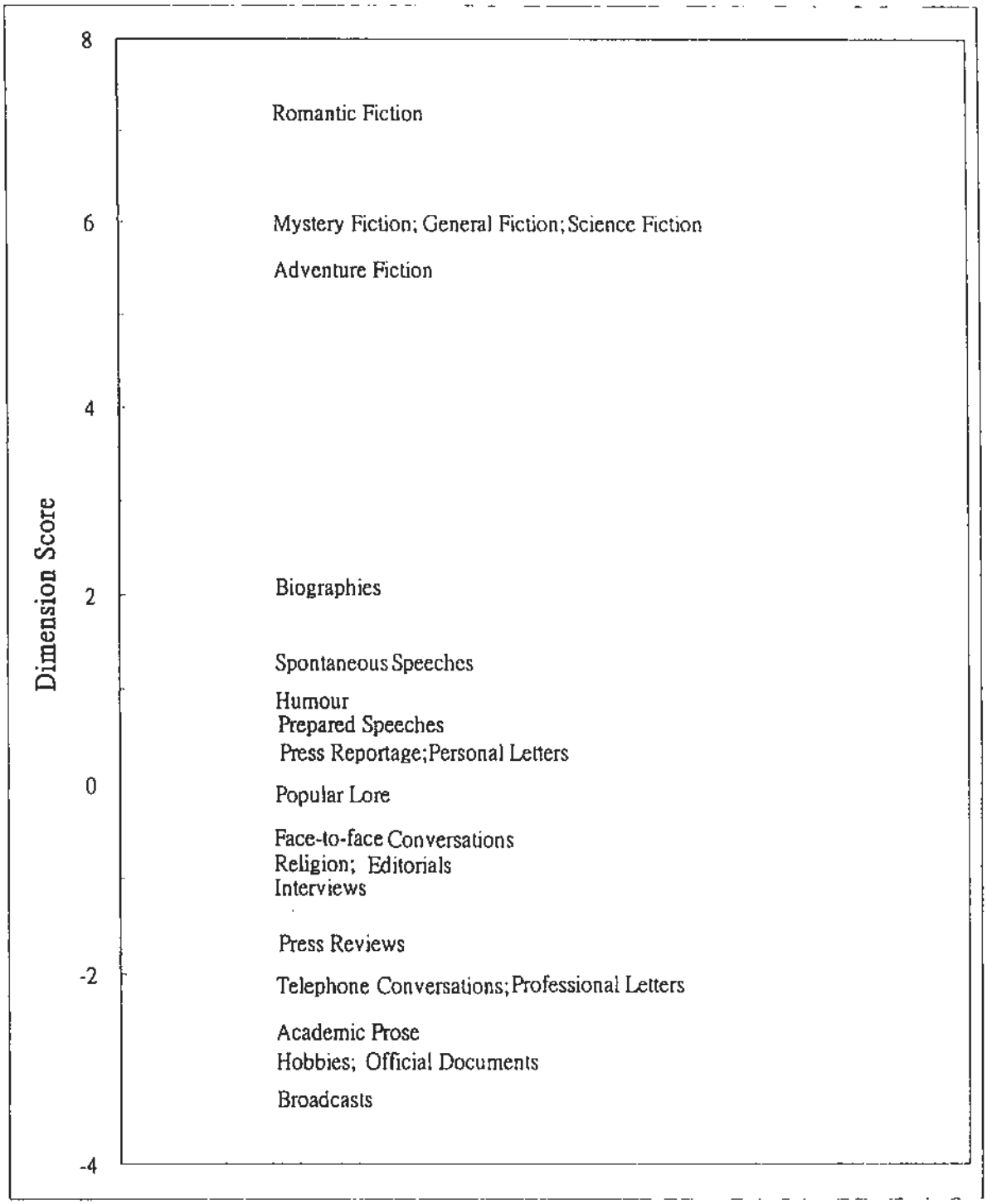

Figure 5. Distribution of the mean scores of different genres along Dimension 2 Narrative versus Non-narrative Concern (Biber, 1988, p. 136)

The five positive features and the three negative features in Dimension 3 all provide assistance for listeners and readers to identify referents in spoken and written discourse. Whereas the positive features, mainly relative clauses, offer explicit information about the referents, the negative features, e.g. time and place adverbials, give clues from the context. The positive end of Dimension 3 represents language making use of more explicit reference, as in official documents and professional letters; the negative end represents language incorporating more situation-dependent reference such as radio broadcasts and telephone conversations (see Table 6 and Figure 6).

Table 6

Syntactic Features Underlying Dimension 3 Explicit Versus Situation-dependent Reference in Biber (1988) (p. 102)

Positive Features	**Negative Features**
Wh relative clauses on object positions	Time adverbials
Pied piping constructions	Place adverbials
Wh relative clauses on subject positions	Adverbs
Phrasal coordination	
Nominalizations	

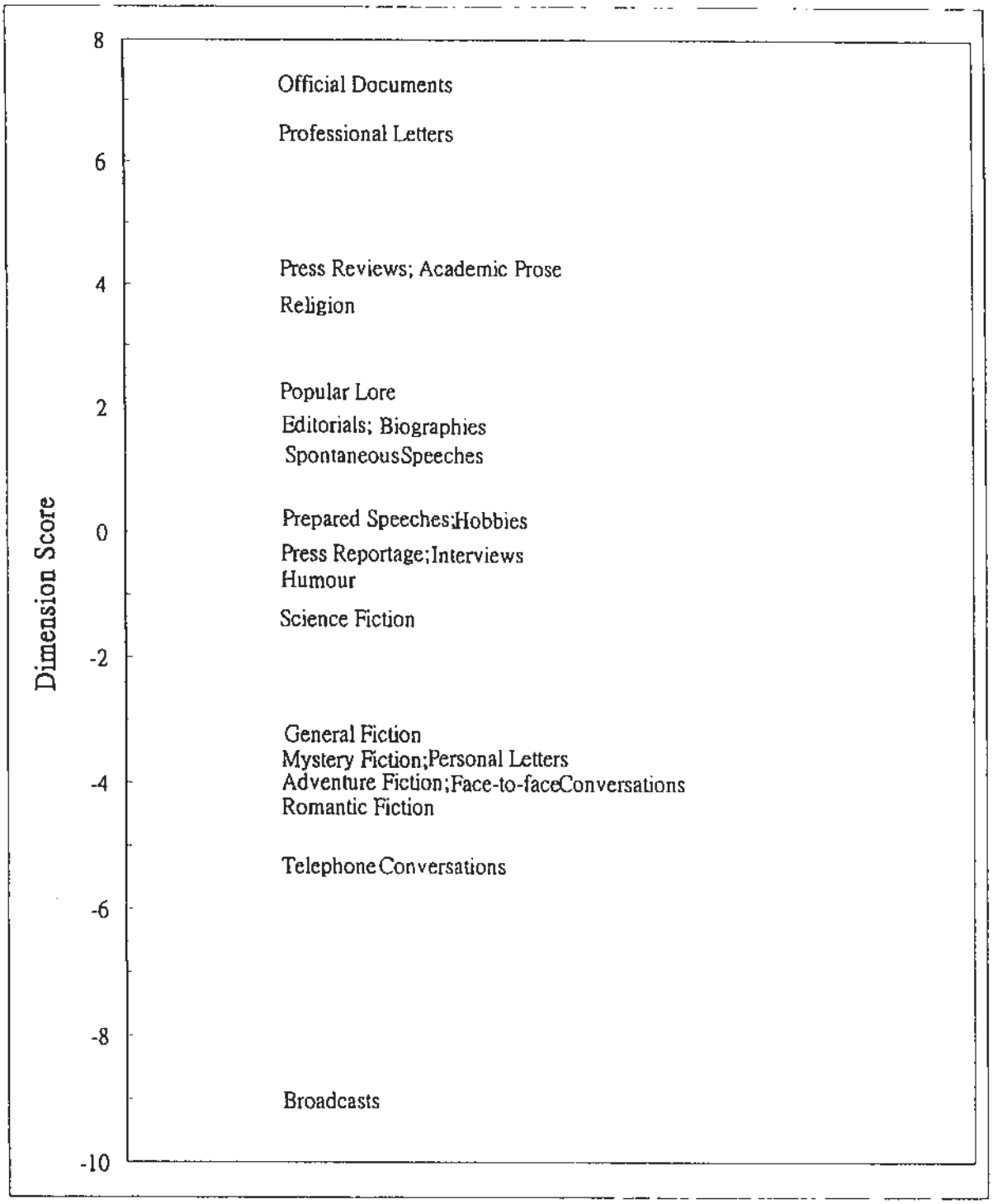

Figure 6. Distribution of the mean scores of different genres along Dimension 3 Explicit versus Situation-dependent Reference (Biber, 1988, p. 143)

Dimension 4 indicates the presence of personal opinion and the intention of persuasion. There are six positive features loaded on this dimension. Most of them, e.g. prediction modals and suasive verbs, are future-related. The positive end of Dimension 4 separates persuasive discourse from the rest of discourse. Typical examples of genres with high Dimension 4 scores are professional letters and newspaper editorials; examples of genres with low Dimension 4 scores are broadcasts and press reviews (see Table 7 and Figure 7).

Table 7

Syntactic Features Underlying Dimension 4 Overt Expression of Persuasion in Biber (1988) (p. 103)

Positive Features	**Negative Features**
Infinitives	-- *no negative features* --
Predictive modals	
Suasive verbs	
Conditional subordination	
Necessity modals	
Split auxiliaries	

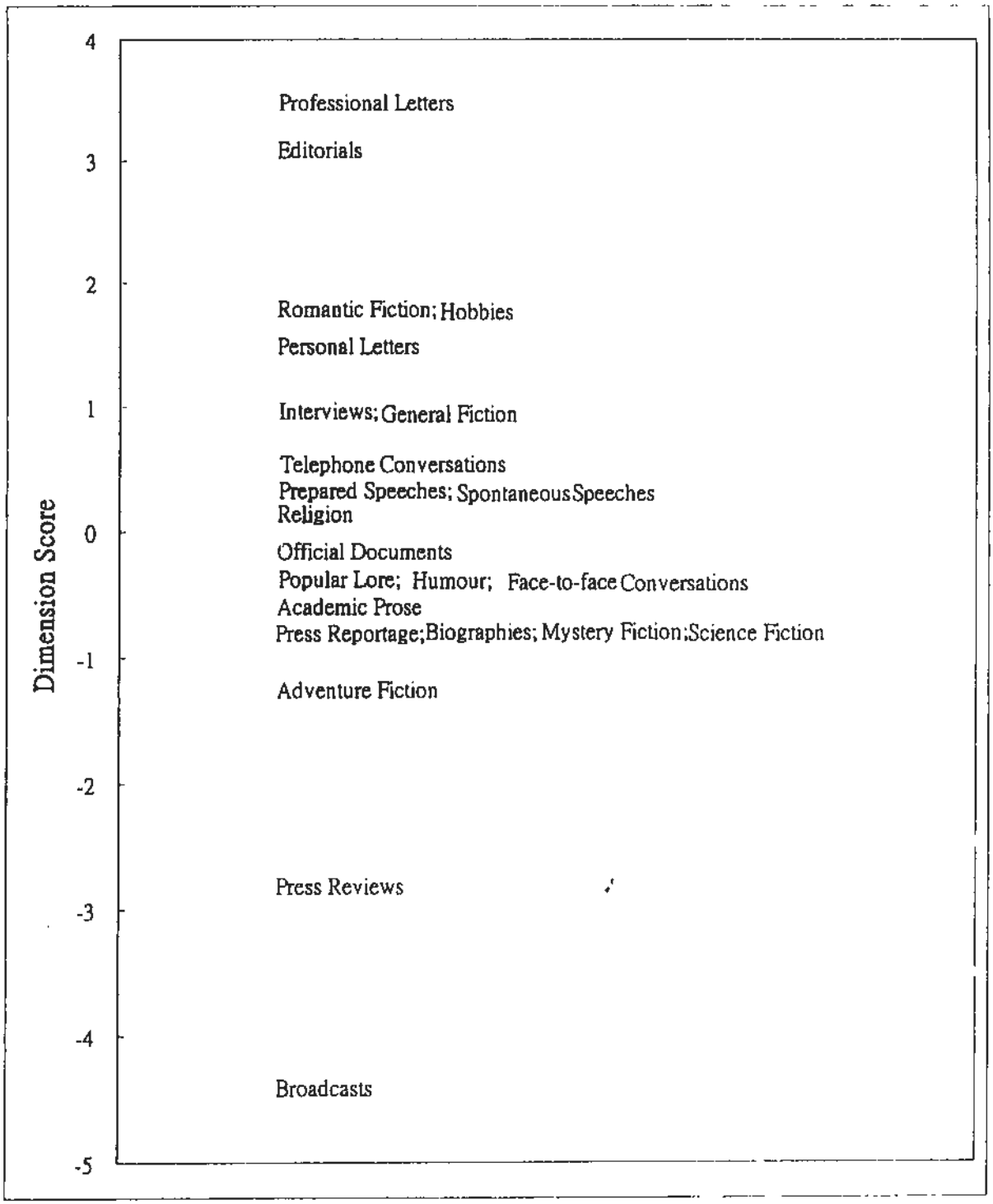

Figure 7. Distribution of the mean scores of different genres along Dimension 4 Overt Expression of Persuasion (Biber, 1988, p. 149)

Dimension 5 consists of six positive features. The positive end of Dimension 5 shows a style of abstract and technical language use characterised by passive constructions. Academic prose and official documents are two genres with very high scores in this textual dimension. In contrast, genres with very little abstract content such as telephone and face-to-face conversations tend to exhibit very low scores in Dimension 5 (see Table 8 and Figure 8).

Table 8
Syntactic Features Underlying Dimension 5 Abstract Versus Non-abstract Information in Biber (1988) (p. 103)

Positive Features	**Negative Features**
Conjuncts	*-- no negative features --*
Agentless passives	
Past participial clauses	
By-passives	
Past participial *whiz* deletions	
Other adverbial subordinators	

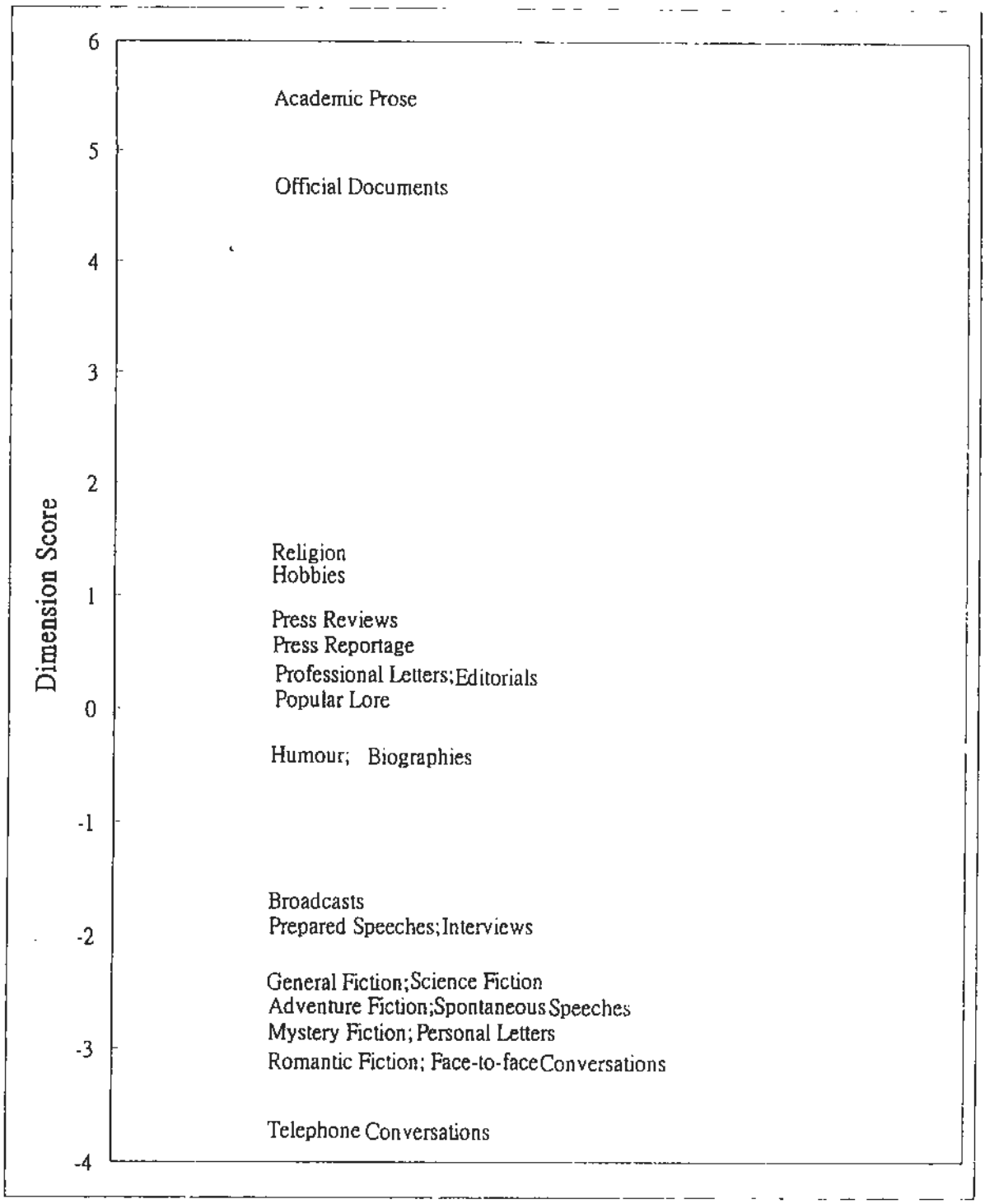

Figure 8. Distribution of the mean scores of different genres along Dimension 5 Abstract versus Non-abstract Information (Biber, 1988, p. 152)

The last dimension, Dimension 6, is the one containing the fewest syntactic features. Among the four positive features, three are *that* clauses. The positive end of Dimension 6 distinguishes informational discourse produced within limited time from other types of discourse. Genres with high Dimension 6 scores include prepared speeches and interviews, whereas those with low scores include various fiction categories (see Table 9 and Figure 9).

Table 9

Syntactic Features Underlying Dimension 6 On-line Informational Elaboration in Biber (1988) (p. 103)

Positive Features	**Negative Features**
That clauses as verb complements	*-- no negative features --*
Demonstratives	
That relative clauses on object positions	
That clauses as adjective complements	

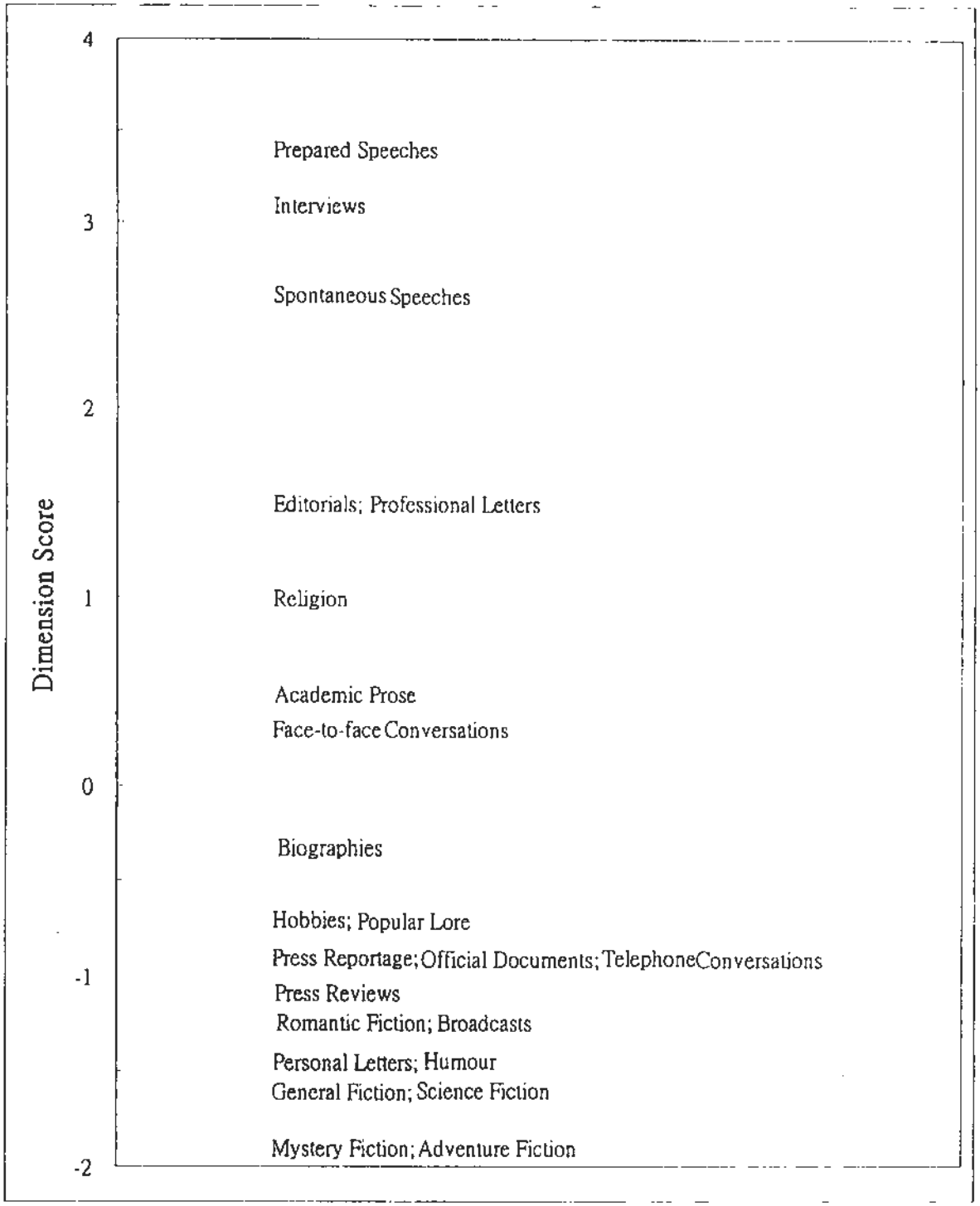

Figure 9. Distribution of the mean scores of different genres along Dimension 6 On-line Informational Elaboration (Biber, 1988, p. 155)

One major finding in Biber (1988) is that there is a high degree of overlap between spoken and written English and that there is no clear-cut distinction between the two. Nevertheless, three textual dimensions in his model can help to show the subtle differences between spoken and written English. They are Dimension 1 Involved versus Informational Production, Dimension 3 Explicit versus Situation-dependent Reference and Dimension 5 Abstract versus Non-abstract Information (Biber, 1988, pp. 160-163). Although spoken and written genres still overlap to a certain extent in these three dimensions, there is a general trend that spoken and written genres cluster around the two ends of the continua. In Biber's (1988) model, spoken genres tend to be found in the positive end of Dimension 1, the negative end of Dimension 3 and the negative end of Dimension 5; written genres tend to be found in the negative end of Dimension 1, the positive end of Dimension 3 and the positive end of Dimension 5. In other words, spoken genres usually have a more involved focus, use more situation-dependent reference and contain less abstract information; in contrast, written genres usually exhibit a more informational focus, employ more explicit reference and include more abstract information.

In Biber's (1988) opinion, a comprehensive comparison of two genres should take into consideration all the six textual dimensions but not just any one (pp. 168-169). However, given the fact that my primary interest is on the differences between speech and writing and that only Dimension 1, Dimension 3 and Dimension 5 help distinguish genres in the two modes, the analysis in the present study focuses on these three dimensions. The distribution of spoken and written genres along these three dimensions constitutes the operational definition of appropriate native-speaker

usage in the two modes.

My study is not the first to apply Biber's (1988) theoretical framework to language data other than those in his original study. Biber and his colleagues have applied the same model to some other English data. For example, Biber and Finegan (2001) investigated how different spoken and written genres in English had changed across the past three and a half centuries with reference to Biber's (1988) model. Other researchers have drawn upon Biber's (1988) model, too. For example, Collot and Belmore (1996) used it to explore the characteristics of electronic language on Bulletin Board Systems. Van Rooy (2008) applied the model in his comparison of Black South African English with Standard English in order to show that the former should be regarded as a variety of English. The fact that Biber's (1988) model has been employed in other studies shows its potential to be applied to various types of English data. In the present study, it is adopted to evaluate whether Hong Kong students' use of syntactic features is suitable for the particular production mode.

3.3 Research Design

After an introduction of the theoretical framework underlying the present study, this section provides further information about the research design. The methodological approach adopted is based on contrastive interlanguage analysis, which involves comparing a learner corpus with a control corpus of native English to examine the differences between non-native speakers' usage and native speakers' usage in terms of overuse and underuse (Granger, 2002, p. 12). To find out Hong Kong students' sensitivity to mode differences, samples of their spoken and written English are required. Two learner corpora, one containing transcripts of Hong Kong students' spoken English and one containing texts of Hong Kong students' written

English, were used in the present study. In order to perform contrastive interlanguage analysis for these two learner corpora, two reference corpora, one for the spoken data and one for the written data, were consulted. Comprising genres similar to those used in the learner corpora, these reference corpora represent native-speaker usage against which the learner usage was evaluated. Two additional sets of native-speaker language data representing general spoken and written English were also consulted. Since Hong Kong was a former British colony, British English was chosen as the control English variety.

Figure 10 is a graphic representation of my research design. The arrows with solid lines represent comparisons on syntactic features and the arrows with dotted lines represent comparisons on textual dimensions. The learner speech data, some oral presentations delivered by Hong Kong students, was compared with some prepared speeches given by native speakers in terms of the use of syntactic features. This comparison, addressing Research Question 1.1 and Research Question 1.2, was drawn in the pilot study. The rest of the comparisons were all made in the main study: The students' oral presentations were compared with the native speakers' prepared speeches and some common native-speaker written genres in terms of the textual dimension (for Research Question 1.3 and Research Question 1.4 respectively). The learner writing data, some examination essays written by Hong Kong students, was compared with some native speakers' essays in terms of the use of syntactic features (for Research Question 2.1 and Research Question 2.2). The students' essays were also compared with the native speakers' essays and some common native-speaker spoken genres in terms of the textual dimensions (for Research Question 2.3 and Research Question 2.4 respectively).

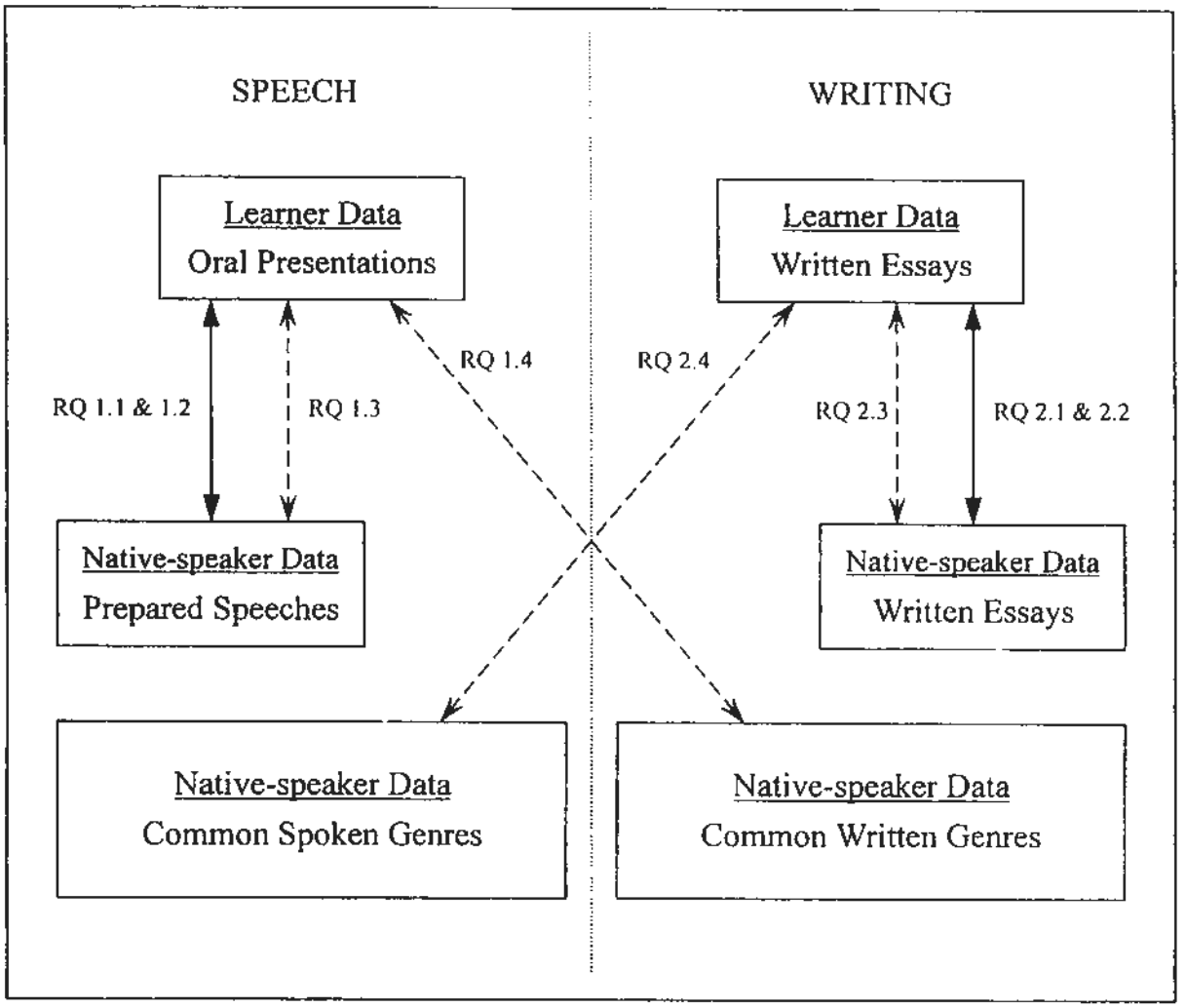

Figure 10. Research design of the present study

As discussed in Section 3.2.2, Biber's (1988) model consists of six textual dimensions, each related to different sets of syntactic features. Both Biber and Finegan (2001) and Collot and Belmore (1996), studies applying Biber's (1988) model, focus their analysis on the dimension scores. However, the present study, like van Rooy (2008), places equal emphasis on both the dimension scores and the frequencies of individual syntactic features. There are two main reasons for this decision. The first reason is that examining the frequencies of syntactic features helps identify the source of the learners' deviant usage. Each textual dimension involves a number of features. For example, Dimension 1 contains a total of 28

features. Even if the Dimension 1 score in the learner data is higher than expected, it does not necessarily mean that the students overuse every positive feature on Dimension 1. Focusing on the dimension scores alone may mask important patterns in learners' use of syntactic features. For instance, Collot and Belmore (1996) have warned their readers that the high value of Dimension 5 score in their study is caused by the high frequencies of two features with less important functional roles in the dimension (p. 25). This, in fact, can be considered a drawback in Biber's (1988) model as it seems that not every single feature has a clearly identified function closely linked to its assigned textual dimension. The second reason for studying the syntactic features is that they are of higher relevance to teachers and students. The purpose of the present study is not purely academic. More importantly, it is hoped that, by advancing current knowledge on Hong Kong students and their English, this study can help the teaching practitioners to better train their learners to become competent language users. To teachers and students, the textual dimensions are some abstract, unfamiliar concepts, but the syntactic features are not. Discussions on syntactic features thus have a higher practical value for the teaching and learning of English. Consequently, the comparisons made in the present study were based not only on the textual dimension scores, but also on the frequencies of individual syntactic features.

3.4 Data

As mentioned in Section 3.3, language data from both learners and native speakers are necessary for the implementation of contrastive interlanguage analysis. In this section, specific details of the different sets of language data employed in the present study are provided.

3.4.1 Learner Data

The Hong Kong learners under investigation are senior secondary school students who had passed the public examination, Hong Kong Certificate of Education Examination (HKCEE), and who were preparing to take or were taking the matriculation examination, Hong Kong Advanced Level Examination (HKALE). Most of them were in their late teens. Since most children in Hong Kong begin learning English at 3 years old, the learners in the present study have learned English for at least 13 years. Further information about the learner speech data and the learner writing data is given in Section 3.4.1.1 and Section 3.4.1.2.

3.4.1.1 Speech data.

The learner speech data is some self-collected oral presentations, the transcripts of which total around 10,000 tokens. It was used in both the pilot study and the main study. The following sub-sections present information about the data collection process, the data transcription process and the participants.

3.4.1.1.1 Data collection.

In the pilot study, speech samples were obtained from 64 Form 6 students in a Band 2 secondary school in New Territories in Hong Kong. The approval of the students' English teachers, including the School English Panel, was sought beforehand: Each teacher was informed of the purpose as well as the procedures of the study and was asked to sign a consent form (see Appendix B). Each student participant was also provided with information of the research and was given a consent form. If the student agreed to participate, he/she was asked to fill in a short questionnaire which included entries on sex, age and public examination results (see

Appendix C). To make sure that the students understood the content, the documents were written in both English and Chinese, i.e. their mother tongue.

The recordings were made, as suggested by the English Panel, during the school oral English examination in November 2008. The format of this examination resembled the format used in the HKALE, the public examination that the students were prepared to face in less than two years' time. The subject concerned was Advanced Supplementary Level Use of English (UE) and the part concerned was Part One in Section D Oral English. Each student was assigned a short piece of English text of about 300 words. The student was then allowed 10 minutes to read the article and jot notes on a note-card. After this 10-minute preparation, the student was asked to give a 2-minute oral presentation, with the help of the note-card but in the absence of the article. The teacher assessor, keeping a timer, would stop the student if the presentation exceeded the time limit. Altogether 16 different articles were employed as the stimuli in the examination. They were all taken from the past papers of the UE examination in the 2007 HKALE; in other words, they were real question papers from a previous public examination. Their topics are general in nature and cover issues relevant to the students' daily lives: traditional Chinese culture (Articles 1-4), air pollution in Hong Kong (Articles 5-8), food and health (Articles 9-12) and sports (Articles 13-16). The titles of these 16 articles are listed in Appendix D.

3.4.1.1.2 Data transcription.

The recordings of the students' oral presentations were transcribed into orthographic words. Since the present study focuses on lexico-grammatical features, other features like pauses and pronunciation variations were not attended to.

Incomplete words or direct repetitions of words due to hesitation were ignored. Parts that were inaudible or unclear were marked. During the transcription process, it was discovered that some of the recordings were unusable due to unsatisfactory acoustic quality. In some cases, the students' voices were too low or too soft and as a result, most parts of the presentations could not be heard; in other cases, the students' pronunciation was so inaccurate that the presentations became unintelligible. These files were thus discarded from analysis to ensure research validity and reliability. In the end, 52 transcripts of oral presentations remained to form the learner speech data. Their lengths range from 132 to 317 tokens, with an average of 192 tokens. The total length of the corpus is 9,982 tokens.

3.4.1.1.3 Participants.

These 52 participants include 28 male and 24 female Form 6 students aged from 16 to 20, with an average age of 17.7. Figure 11 and Figure 12 show the results that they obtained from the public examination, HKCEE, in the English language subject as a whole and in one sub-paper, Paper 3 Speaking, of the English language subject. Most of the students reported their 2008 HKCEE results. Three reported their results in 2007 and one did not fill in the year.

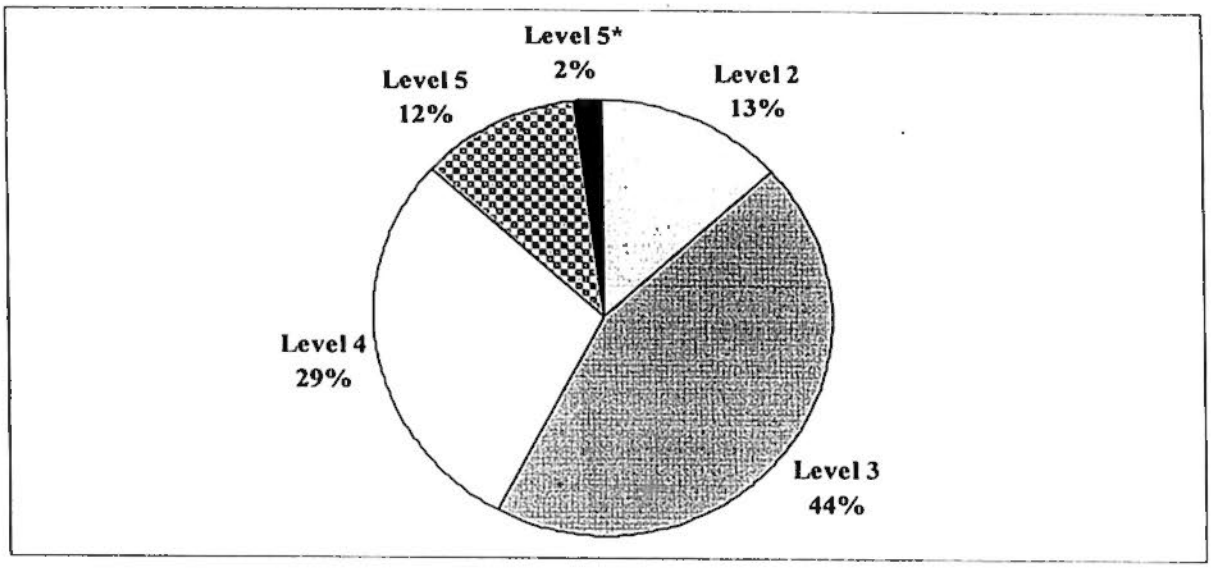

Figure 11. Participants' grades in the English language subject in HKCEE (with Level 5* as the highest grade and Level 2 as the lowest passing grade)

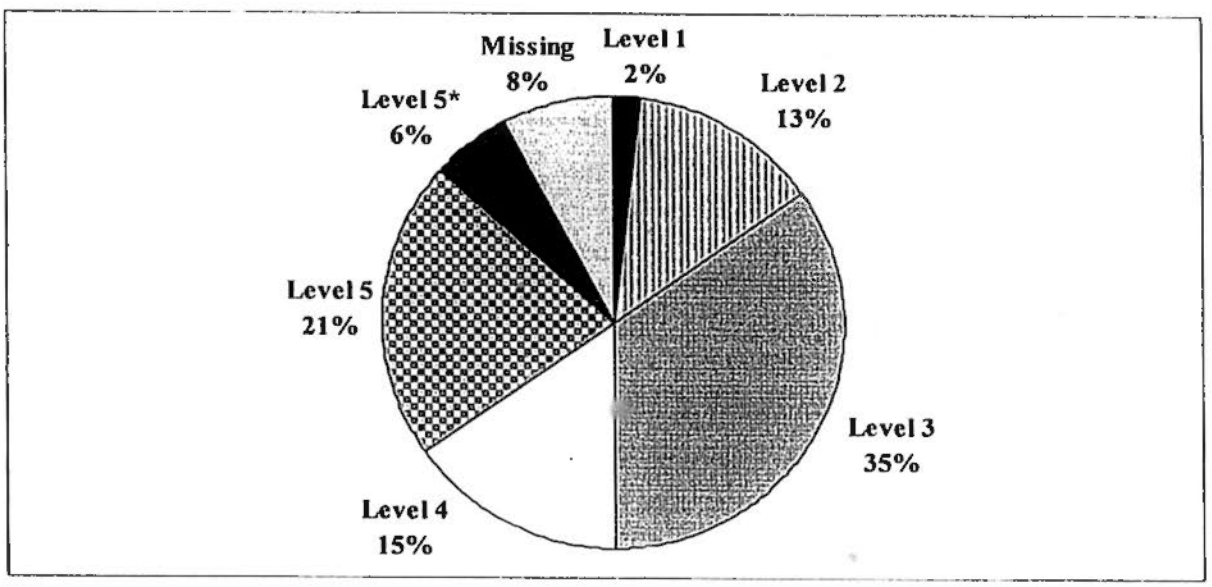

Figure 12. Participants' grades in one English sub-paper, Paper 3 Speaking, in HKCEE (with Level 5* as the highest grade and Level 2 as the lowest passing grade)

3.4.1.2 Writing data.

The learner writing data was taken from an existing corpus of public examination essays, in which there were approximately 1,500 scripts collected from the UE examination (Section B Writing) in the 1999, the 2000 and the 2001 HKALE. In this examination of English writing, students were required to produce a 500-word expository or argumentative essay, on one of the four topics provided, within 1 hour 15 minutes. Because the learner writing data was intended to be comparable with the learner speech data, the number of essays chosen had to match the number of oral presentations used. Therefore, 52 essays were sampled from the UE examination writing corpus. Nonetheless, the essays are much longer than the presentations: The average length of the former is 525 words, which is 2.5 times as long as the latter. In order to make the size of the learner writing data comparable to that of the learner speech data, the essays were not sampled in their entirety. For each essay, selection was made from the complete sentence containing the 151st word to the complete sentence containing the 350th word. In other words, about 200 words were extracted from the middle part of each essay. Among the 52 essays sampled, the shortest one contains 353 tokens and the longest one contains 745 words. By choosing the middle 200 words, it was hoped that the resultant learner writing corpus would contain mostly the main-body texts, but still some opening and closing texts. If, instead, the first 200 words had been sampled, there would have been no chance for the closing texts to be represented; similarly, if the last 200 words had been selected, the opening texts could never have been included. Thus, after careful consideration of the pros and cons of different sampling strategies, it was decided that middle-part selection should be the most suitable option for the present study.

Besides matching the amount of texts in the learner writing data with that in the learner speech data, care was taken to ensure that the student profiles in the two sets of data did not differ too much. Unfortunately, the original UE examination writing corpus contains minimal header information: Only an essay identification number and a score are present. With this limited information, there was no way to control the proportion of male to female students in the sampling process. Regarding age, the HKALE is usually taken by Hong Kong secondary school leavers, i.e. Form 7 students. Given that the participants for the learner speech corpus are Form 6 students, the ages of the two groups can be considered similar.

As for English proficiency, a score within the range from 1 to 18 is provided for each essay in the original UE examination writing corpus. Hong Kong Examinations Authority (1999, 2000, 2001) advises markers to evaluate the examination essays on a 9-point scale and classifies the essays into three proficiency levels (see Table 10). Accordingly, in the present study, essays with scores between 1 and 6 were regarded as below average. Those with scores between 7 and 12 were considered average and those with scores between 13 and 18 were treated as above average. To determine how the 52 essays should be sampled, reference was made to Table 11 and Figure 11. Table 11 summarises the performance of all the candidates taking the English examination in the 2008 HKCEE (Hong Kong Examinations and Assessment Authority, 2008). Over 50% of the students obtained Level 2 and Level 3 in the examination. Slightly more than 30% of the students obtained Level 1 and unclassified results; slightly less than 20% of the students obtained Level 4, Level 5 and Level 5*. Generally speaking, students obtaining Level 2 and Level 3 could be considered average. Those getting a higher level could be regarded as above average and those with a lower level could be seen as below average. Figure 11 gives the

proficiency profile of the participants in the learner speech data. Nearly 60% of the participants obtained Level 2 and Level 3 in the HKCEE and around 40% obtained higher levels. Therefore, to make the student proficiency in the learner writing data comparable with that in the learner speech data, about 60% of the essays sampled were of an average level, i.e. with scores between 7 and 12. About 40% of the essays sampled were of an above average level, i.e. with scores between 13 and 18.

Table 10

Nine-point Scale in Assessing Examination Essays (Hong Kong Examinations Authority, 2001, p. 15)

Mark	Proficiency
1-3	Below average
4-6	Average
7-9	Above average

Table 11

Statistics of the English Language Subject in the 2008 HKCEE (Hong Kong Examinations and Assessment Authority, 2008)

Grade	Percentage of Candidates (N = 101,656)
Level 5*	1.2%
Level 5	4.9%
Level 4	10.2%
Level 3	24.2%
Level 2	27.4%
Level 1	20.3%
Unclassified	11.8%

In the original UE examination writing corpus, 12 essay topics were involved. However, only essays of five topics were selected for the learner writing data in the present study. Other topics were not deemed appropriate because they required students to write essays in the form of newspaper articles or letters to the editor. The five essay questions are listed in Appendix E. The 52 essays sampled for the learner writing data distribute evenly across these five topics. Except for one topic, 10 essays were sampled from each topic. Altogether 11,224 tokens have been included in the learner writing data for the present study. The average length of the 52 essays is 216 tokens, the minimum being 202 tokens and the maximum 243 tokens.

3.4.2 Native-speaker Data

In Section 3.4.1, details of the learner data have been presented. In this section, information about the native-speaker speech data and the native-speaker writing data is provided. As mentioned before, the native speakers in the present study refer to native speakers of British English.

3.4.2.1 Speech data.

Two sets of native-speaker speech data have been employed in the present study. The first set involves a native-speaker spoken genre that is comparable to the oral presentations used as the learner speech data. The second set represents native speakers' spoken English in general.

3.4.2.1.1 Prepared speeches.

To draw comparison with the learner speech data, some reference data which illustrates how native speakers use English in a similar situation is needed. Two spoken genres included in Biber (1988), prepared speeches and spontaneous speeches, were considered possible reference standards to fulfil the above function. In the pilot study, statistical evidence suggests that the prepared speeches are more similar to the oral presentations in the learner speech data than the spontaneous speeches are (see Section 3.7). Consequently, the prepared speeches in Biber (1988) were adopted as the standard to which the learner speech data was compared. The genre of prepared speeches in Biber (1988) contains 14 transcripts, obtained by dividing the 5000-word transcripts from the London-Lund Corpus of Spoken English into shorter transcripts ranging from 400 to 2500 words in length (Biber, 1988, p. 67, 208). The exact number of words in this genre is not specified in Biber (1988), but another study based on the same data suggests that there are approximately 31,000 words in the genre of prepared speeches (Biber, 1995, p. 87). Specific types of speeches included in this genre are sermons, university lectures, court cases, political speeches and popular lectures (Biber, 1988, p. 69). The people who delivered these prepared speeches are 12 male British speakers: three in their thirties, one in his forties, four in their fifties, two in their sixties and two in their seventies (Greenbaum & Svartvik, 1990, pp. 44-45). From the above information, the readers can already realise that this set of native-speaker speech data, though similar to the learner speech data by being monologue produced in front of some audience, is not identical to the learner speech data in some other aspects. However, no better native-speaker reference data was identified. As a result, Biber's (1988) prepared speeches remained the native-speaker standard of reference in the speech data comparison and the

differences between the two sets of language data were taken into consideration during the interpretation of the findings. Since Biber (1988) published the descriptive statistics in his appendix, there was no need for my processing the texts, counting the syntactic features and calculating the dimension scores again. In other words, Biber's (1988) statistics for prepared speeches constitute the native-speaker speech data with which the learner speech data was compared in the pilot study (for Research Question 1.1 and Research Question 1.2) and in the main study (for Research Question 1.3).

3.4.2.1.2 Common spoken genres.

Another set of native-speaker speech data is required for the purpose of demonstrating how native speakers use syntactic features in the spoken mode in general. This second set of native-speaker speech data is Biber's (1988) statistics for six common English spoken genres. Since the comparison for Research Question 2.4 focused on the textual dimensions, only statistics of the dimension scores were adopted as this second set of native-speaker speech data. As shown in Table 2 in Section 3.2.1, the six spoken genres in Biber (1988) are face-to-face conversation, telephone conversation, interviews, broadcasts, spontaneous speeches and prepared speeches. The 141 texts in these six genres, all between 400 words and 2500 words in length, were obtained from the London-Lund Corpus of Spoken English.

3.4.2.2 Writing data.

As in the case of speech data, two sets of writing data produced by native speakers are required. Whereas the first set represents a native-speaker written genre that is similar to the essays used as the learner writing data, the second set involves

common written genres and stands for general written English produced by native speakers.

3.4.2.2.1 Written essays.

None of the 17 written genres in Biber (1988) appear to be similar to the examination essays in the learner writing data in terms of the production conditions and the nature of the discourse. Consequently, another source of data was sought. Fifty-two essays from the Louvain Corpus of Native English Essays (LOCNESS) were sampled to form the native-speaker writing data for addressing Research Question 2.1, Research Question 2.2 and Research Question 2.3. The LOCNESS is a 300,000-word corpus consisting of essays written by British and American students. In particular, it contains some argumentative essays on general topics written by British A-level students (*UCL/CECL Locness corpus*, n.d.). These essays, comparable to the Hong Kong students' written essays in terms of the type of essays and the nature of topics, were adopted as the first set of native-speaker writing data in the present study. In view of the size of the learner writing data, only fifty-two 200-word samples were taken from the LOCNESS. The original lengths of the 52 sampled essays range from 356 to 747 tokens, with an average of 532 tokens. Again the middle part, i.e. from the complete sentence containing the 151st word to the complete sentence containing the 350th word, was extracted from each essay. The resultant excerpts have an average length of 231 tokens, totaling 12,024 tokens. The shortest text contains 205 tokens and the longest one 261 tokens. The topics of these 52 essays are listed in Appendix F. The LOCNESS does not include information about the writers and the production conditions of these essays.

3.4.2.2.2 Common written genres.

The second set of native-speaker writing data represents native speakers' written language use in general. Biber's (1988) statistics for the textual dimensions of the 17 common written genres were adopted for this purpose and were used in the comparison for Research Question 1.4. These 17 genres are:

- Press reportage
- Editorials
- Press reviews
- Religion
- Skills and hobbies
- Popular lore
- Biographies
- Official documents
- Academic prose
- General fiction
- Mystery fiction
- Science fiction
- Adventure fiction
- Romantic fiction
- Humour
- Personal letters
- Professional letters

There are altogether 340 texts in the 17 genres. As mentioned in Section 3.2.1, most of these texts were language data from the LOB Corpus.

3.5 Data Processing

As pointed out in Section 3.4, among the six sets of language data used in the present study, three sets are readily available statistics published in Biber (1988) and three sets are raw texts in electronic format. These electronic texts had to be further processed before statistics could be generated. This section describes the data processing procedures of these texts and the description applies only to the following three sets of data: the learner speech data (i.e. transcripts of oral presentations), the learner writing data (i.e. written essays) and one set of the native-speaker writing

data (the set containing LOCNESS essays). The data processing procedures echo those in Biber (1988): First, the texts were tagged. Next, the 67 syntactic features were counted. Finally, relevant statistics were generated.

3.5.1 Tagging

In order to facilitate the automatic searching of the syntactic features, the electronic texts were grammatically annotated. An automatic tagger, ICE-GB Tagger 1.1 (Wallis, n.d.), was used in the present study. This software, downloaded from the Survey of English Usage website of University College London, assigns part-of-speech information to every lexical unit in a text by using the International Corpus of English (ICE) tagset, a system of word class tags developed for the ICE project. The electronic texts were first tagged by the automatic tagger and the tagged texts were then manually checked for accuracy. Inaccurate tags were corrected by the researcher. Ambiguities were resolved by consulting Nelson, Wallis and Aarts (2002) as well as the British component of the ICE corpus (ICE-GB). On average, the tagger achieved an accuracy rate of 87.4% for the learner speech data, 89.1% for the learner writing data and 87.6% for the native-speaker writing data (LOCNESS essays).

The ICE-GB Tagger 1.1 was chosen for the grammatical annotation in the present study mainly because it was developed based on Quirk et al. (1985), the same grammar reference that Biber (1988) used to describe his syntactic features. The fact that the ICE-GB Tagger 1.1 and Biber (1988) share the same reference ensures that the classification system in the ICE tagset is similar to the classification system in Biber (1988), hence greatly facilitating the subsequent data searching process. Another reason that makes the ICE tagset desirable for the present study is its user-friendliness. One major advantage of the ICE tagset over some other tagsets,

e.g. the Constituent Likelihood Automatic Word-tagging System (CLAWS) tagset, is that the ICE tags can be easily understood by people who have some knowledge of the English grammar. The tags resemble the terminology traditionally used in grammatical analysis. For example, general prepositions are tagged as "PREP(ge)" according to the ICE tagset (Nelson et al., 2002, p. 34) but they are tagged as "II" according to the CLAWS tagset (University Centre for Computer Corpus Research on Language, n.d.). As suggested by Greenbaum (1993), tags that look familiar are helpful to researchers as such tags can facilitate the checking and the searching processes (p. 13). Given the above two reasons, the ICE-GB Tagger 1.1 is a very good option of annotating programme in the present study. My data was thus grammatically annotated by using this tagger and the tagging mainly followed the practice in the ICE-GB.

One problem encountered during the tagging process was related to the English proficiency of the Hong Kong learners. The ICE-GB Tagger 1.1 was developed for and trained on English language data produced by native speakers. Since the Hong Kong students' English sometimes deviated from the native speakers' usage, the tagger could not tag the texts as satisfactorily as expected, hence resulting in the less than 90% accuracy rate. In some cases, the students' sentences were so ill-formed that it was very hard to decide which tag should be assigned. The general principle applied in the present study was that tags were determined primarily by the orthographic form of the words. When a word (e.g. *present*) was wrongly spelt (e.g. as *presence*), it was tagged according to the appearing form (i.e. *presence* in the example). But if the resultant word (e.g. *presant*) did not exist in English and it highly resembled the intended word, the word would be tagged as if it was the intended word (i.e. *present* in the example). When the same form of a word could

function as different parts of speech (e.g. *present* as a noun, a verb or an adjective), the sentence context would be consulted before a decision was made. If, unfortunately, the ambiguity could not be resolved by the above method, the word would be tagged as an indeterminate case. Nonetheless, cases like this were not common in the data.

3.5.2 Searching

After the texts were tagged and the tags were checked, automatic searching of the syntactic features could be performed. The search commands were written by the researcher by using tools of GnuWin32, a Unix-like operating system. The commands followed Biber's (1988) algorithms, as summarised in Appendix A of the thesis, as closely as possible in order to ensure that the resultant statistics were comparable to those in Biber (1988). However, some minor adjustments were necessary for a very small number of features. These adjustments are listed and explained in Appendix G. One issue that should be mentioned is that, unlike the spoken data in Biber (1988), the spoken data in the present study was not marked for tone-unit boundaries. When tone-unit boundaries were required in the identification of syntactic features, the search method had to be revised and extra checking of the search results was carried out. Overall, the minor adjustments in the search method do not seriously affect the findings of the present study. After the automatic searching process was completed, the computer-generated results were checked manually by the researcher. If necessary, the sentence context would be consulted to determine the accuracy of the search results.

Except type/token ratio (Feature 43) and mean word length (Feature 44), all the syntactic features were retrieved by the above procedures. The type/token ratio and the mean word length of each text were obtained by using WordSmith 5 (Scott, 2008). In Biber (1988), the formula for type/token ratio is "counting the number of different lexical items that occur in the first 400 words of each text, and then dividing by four" (p. 238). Nevertheless, the average length of the texts used in the present study is only about 200 words and Biber's (1988) method is not applicable. As a result, in the present study, the type/token ratio was calculated by using the total text length instead: dividing the number of different lexical items by the number of tokens and then multiplying the resultant fraction by 100. The readers should bear in mind this difference in the calculation of type/token ratio when interpreting the results of the present study.

3.5.3 Generation of Statistics

As noted earlier, the present study has adopted Biber (1988) as the theoretical framework. Therefore, the procedures for generating statistics in the present study followed exactly the procedures described in Biber (1988).

3.5.3.1 Mean frequencies of syntactic features.

By the end of the searching process, the frequency of each of the 65 syntactic features (i.e. excluding type/token ratio and mean word length) in each text was available. Nonetheless, the raw frequencies needed to be normalised before comparisons could be made across texts of different lengths. The normalised frequency for each feature was computed by dividing the raw frequency by the total number of tokens in the text and then multiplying the resultant figure by 1,000. This

normalised frequency represents the number of occurrences per 1,000 words of a text (Biber, 1988, pp. 75-76). The normalised frequencies of different texts in the same set of data were then averaged to obtain the mean normalised frequency for each syntactic feature. The corresponding standard deviation for each feature was also calculated. Regarding type/token ratio and mean word length, the figures for each text were generated directly by the WordSmith software. The averages and the standard deviations for each set of data were then calculated from these figures.

3.5.3.2 Mean dimension scores.

To compute the dimension scores, the normalised frequency for each feature in each text was first standardised: The mean normalised frequency for feature *x* in Biber's (1988) complete set of data (as listed in Biber, 1988, pp. 77-78) was subtracted from the normalised frequency for feature *x* in my text. The difference was then divided by the standard deviation for feature *x* in Biber's (1988) complete set of data (as listed in Biber, 1988, pp. 77-78). The idea of standardising frequencies is to have them represented on a scale with a mean of 0 and a standard deviation of 1, hence preventing "those features that occur very frequently from having an inordinate influence" (p. 94) on the dimension scores. Next, the score for each dimension in each text was generated by adding up the standardised frequencies of all syntactic features loaded on the positive side of the factorial structure of the dimension, minus the standardised frequencies of features loaded on the negative side, if any, of the same dimension. After this procedure, the score for each of the six dimensions in each text was obtained. Finally, the mean dimension score in each set of data was calculated by averaging the corresponding dimension scores of all the texts in the same set of data (Biber, 1988, pp. 93-95).

3.6 Data Analysis

After the three sets of data which were originally in the form of raw texts underwent the data processing procedures outlined above, their statistics on syntactic features and textual dimensions became available. As a result, the various comparisons illustrated in Figure 10 in Section 3.3 could be made. This section explains how the data in the present study was analysed in the comparisons.

3.6.1 Comparisons on Mean Frequencies of Syntactic Features

The comparisons on the mean frequencies of syntactic features refer to (a) the comparison between the learner speech data, i.e. oral presentations, and the similar native-speaker speech data, i.e. prepared speeches (for Research Question 1.1 and Research Question 1.2) and (b) the comparison between the learner writing data, i.e. written essays, and the similar native-speaker writing data, i.e. LOCNESS essays (for Research Question 2.1 and Research Question 2.2). These comparisons are aimed at highlighting how the learners' production differs from the native speakers' production. For each syntactic feature, the mean normalised frequency in the native-speaker data was subtracted from the mean normalised frequency in the learner data. The standard deviation in the native-speaker data was also subtracted from the standard deviation in the learner data. The resultant differences in mean normalised frequencies and standard deviations represent how the learners' usage deviates from the native speakers' usage. Syntactic features of the top 10 most positive and the top 10 most negative differences in mean normalised frequencies were further examined. Characteristics of the learners' production were deduced from these most heavily overused and underused features. Furthermore, these 20 syntactic features were classified according to their roles on Biber's (1988) model to

determine whether they were indicative of any confusion about mode differences.

3.6.2 Comparisons on Mean Dimension Scores

The comparisons on the mean dimension scores refer to (a) the comparison between the learner speech data, i.e. oral presentations, and the similar native-speaker speech data, i.e. prepared speeches (for Research Question 1.3); (b) the comparison between the learner speech data and common native-speaker written genres (for Research Question 1.4); (c) the comparison between the learner writing data, i.e. written essays, and the similar native-speaker writing data, i.e. LOCNESS essays (for Research Question 2.3) and (d) the comparison between the learner writing data and common native-speaker spoken genres (for Research Question 2.4). The first and the third comparisons focus on how the learners' production deviates from the native speakers' production. The second and the fourth comparisons are intended to find out whether there is any similarity between the learners' production and common native-speaker genres of the opposite mode. In all four comparisons, only scores for Dimension 1, Dimension 3 and Dimension 5, i.e. the three dimensions that Biber (1988) found to distinguish between speech and writing, were considered. To facilitate the comparisons, the dimension scores of different sets of data were plotted on graphs.

3.7 Pilot Study

The above sections in this chapter have described the methodology of the present study. So far, most details about the research implementation have been unfolded. What remains to be explained in this section is the pilot study. At the early stage of this research project, during which the ideas of research design were formed,

a pilot study was conducted to experiment with the research method. The pilot study started from November 2008 and lasted until August 2009. It involved (a) the collection and the transcription of the learner speech data, i.e. oral presentations (as described in Section 3.4.1.1); (b) the selection of a suitable tagger, the automatic tagging of the data and the manual editing process (as described in Section 3.5.1); (c) the devising of computer search commands for the 67 syntactic features, the automatic search for the features and the manual checking of the results (as described in Section 3.5.2); (d) the generation of statistics according to Biber's (1988) description (as described in Section 3.5.3) and (e) the comparison of the learner speech data with a comparable native-speaker spoken genre for addressing Research Question 1.1 and Research Question 1.2 (as described in Section 3.6.1).

During the pilot study, various problems arose in the implementation of the research plan. For example, the tagging of the learner speech data was not as straightforward as expected. As pointed out in Section 3.5.1, the learners' English was less than perfect. Decisions about how to tag the erroneous forms had to be made. What was particularly problematic was the writing of the computer search commands and the checking of the search results of the 67 syntactic features. Although Biber (1988) included a lot of information about the features and their search algorithms, it was not until the time when I had to replicate the search in the pilot study that I realised the description in Biber (1988) was not detailed enough. For example, the criterion for identifying gerunds (Feature 15), "participle forms serving nominal functions" (Biber, 1988, p. 227), is less than clear. Biber (1988) merely added that he counted participle forms separately as gerunds, participial adjectives and participial clauses. Unfortunately, distinguishing among these different participial functions is not an easy task. Quirk et al. (1985) give 14

examples of *painting* to illustrate the complexity involved in identifying the functions of *-ing* forms which range from purely nominal to purely verbal (pp. 1290-1292). Gerunds, representing a "mixture of nominal and verbal characteristics in the *-ing* forms" (p. 1291), are somewhere in the middle of the continuum. Since there were no absolute guidelines for gerunds, the search results of this feature had to be double-checked by my supervisor to ensure the accuracy of my decision. On the whole, most of the problems encountered in the searching process of the pilot study were solved by my consulting Quirk et al. and my supervisor. The research method proved to be feasible. The specific practice in the pilot was recorded so that the same data processing and analysis procedures could be repeated in a consistent manner in the main study.

One issue worth mentioning in this section is the choice of a similar native-speaker genre for the comparison with the learner speech data in the pilot study. There are two possible options for the native-speaker genre: spontaneous speeches and prepared speeches, both of which are data in Biber (1988). However, the oral presentations used as the learner speech data in the present study do not totally fit in with the two genre categories. In a description of the London-Lund Corpus, from which Biber's (1988) spoken data was obtained, Greenbaum and Svartvik (1990) characterise the genre of spontaneous speeches as "nearest to conversation in being relatively unplanned" (p. 12) and the genre of prepared speeches as "closest to written English but retain[ing] some spontaneity in not being read from a script and therefore allowing for improvisation" (p. 12). The learner speech data could not be considered "spontaneous" as the students did spend time to plan for their presentations. Nevertheless, the students were only allowed 10 minutes for preparation, during which they also had to read a short article. It is doubtful

whether the oral presentations are as "prepared" as sermons, university lectures, court cases, political speeches and popular lectures, the prepared speeches used in Biber (1988).

To find out which is a better standard of reference for the learner speech data, distance indices have been calculated in the pilot study. One of the major proximity measures in statistics, especially in the field of cluster analysis, is the calculation of "distance" between two sets of data and the most common distance measure is the Euclidean distance (Khattree & Naik, 2000, pp. 356-359; Lorr, 1983, p. 33). The distance indices used in the pilot study are also based on the Euclidean distance. First of all, for each syntactic feature, the mean normalised frequency in the spontaneous speeches was subtracted from the mean normalised frequency in the oral presentations to obtain a difference in their mean normalised frequencies. A difference in their standard deviations was also obtained in a similar manner. Next, each of the differences in mean normalised frequencies and standard deviations was squared. Afterwards, the squared mean frequency differences of all syntactic features were added up. The squared standard deviation differences of all syntactic features were also added up separately. Finally, the square root values of these two sums became the distance indices for the spontaneous speech data. The distance indices for the prepared speech data were calculated by following the same procedures. According to Cramer (2003), squaring the differences can ensure that "this [distance] measure is not affected by the sign or direction of the difference" (p. 47). The distance index of the mean frequency differences (79.6) and that of the standard deviation differences (36.1) for the prepared speeches are smaller than the corresponding distance indices (mean frequency differences = 85.3; standard deviation differences = 42.5) for the spontaneous speeches (see Appendix H for

details). In other words, the students' oral presentation data is less "distant" from the prepared speech data in terms of the attributes of mean and standard deviation. This suggests that the learner speech data is more similar to the prepared speeches than to the spontaneous speeches. Consequently, the prepared speech data in Biber (1988) has been adopted as the comparable native-speaker spoken genre in the comparison with the learner speech data.

Overall, the pilot study was a success. The objective of testing the research method was achieved. Practical problems associated with the implementation of the research plan were solved. The results of the pilot study appeared to be satisfactory. Thus the main study could be conducted by using the same procedures as in the pilot study. For the integrity of the research findings and the ease of comprehension, the results of the pilot study are reported together with the results of the main study in Chapter 4 of this thesis.

3.8 Summary

To evaluate if Hong Kong students produce English language suitable for the mode concerned, samples of their spoken and written English were compared and contrasted with native speakers' spoken and written English with reference to Biber's (1988) multi-dimensional model. The language data of the present study includes:

1. Learner speech data: Oral presentations (9,982 tokens)
2. Learner writing data: Examination essays (11,224 tokens)
3. Native-speaker speech data:
    a. Prepared speeches in Biber (1988)
    b. 6 spoken genres in Biber (1988)

4. Native-speaker writing data:

    a. Argumentative essays from the LOCNESS (12,024 tokens)

    b. 17 written genres in Biber (1988)

Among the above six sets of data, 3a, 3b and 4b were adopted in the form of descriptive statistics from Biber (1988). The other data sets, 1, 2 and 4a, were processed to generate the required statistics: The transcripts and the texts were tagged for part-of-speech information. Frequencies of 67 syntactic features were counted and scores for six textual dimensions were calculated. Comparisons were made to see (a) if Hong Kong students' English usage deviated from native speakers' usage in a similar genre in the same mode and (b) if Hong Kong students' English usage showed any similarity to native speakers' usage in the opposite mode.

CHAPTER 4 RESULTS AND DISCUSSION: SPEECH DATA OF HONG KONG LEARNERS

4.1 Introduction

Chapter 3 has provided a detailed description of the methodology of the present study: the theoretical framework, the research design, the different sets of English language data and the data processing and analysis procedures. All information required for the interpretation of the research results has been elaborated. In this chapter, findings from the first half of the learner language data, i.e. English speech data produced by Hong Kong students, are presented. As mentioned in Chapter 3 Section 3.6, data in the present study was analysed on two levels: the use of individual syntactic features and the overall textual dimensions. Chapter 4 is organised in accordance with these two levels of analysis. The first part of this chapter (Section 4.2) reports findings from the comparison on the mean frequencies of the 67 syntactic features; the second part of this chapter (Section 4.3) examines findings from the comparison on the mean scores of the three relevant textual dimensions.

In Section 4.2, Hong Kong students' oral presentations are compared with similar native-speaker oral production, prepared speeches, in terms of the use of 67 syntactic features. In Section 4.2.1, the mean normalised frequencies of the syntactic features in Hong Kong students' oral presentations are presented. Differences between the learners' presentations and the native speakers' prepared speeches are also listed. In Section 4.2.2, key differences between Hong Kong students' and native speakers' use of syntactic features are explored. With reference to Research

Question 1.1, characteristics of the students' usage are identified by considering the common functions of the most heavily overused and underused features. In response to Research Question 1.2, these features of most deviant usage are classified according to Biber's (1988) model so as to decide whether they are indicative of any confusion about mode differences.

In Section 4.3, Hong Kong students' oral presentations are compared and contrasted with native-speaker production in terms of the three textual dimensions that Biber (1988) considers most relevant to the variation across speech and writing. In Section 4.3.1, the mean scores of all six textual dimensions of Hong Kong students' presentations are provided. In Section 4.3.2, Hong Kong students' oral presentations are contrasted with comparable native-speaker oral production, prepared speeches, in order to show how the learners' spoken English deviates from the native speakers' spoken English on the three relevant textual dimensions. Whereas this comparison on differences is intended to answer Research Question 1.3, another comparison on similarities is made to address Research Question 1.4. In this second comparison, Hong Kong students' oral presentations are compared with some common native-speaker written genres to determine whether any similarity exists on the level of textual dimensions. It is hoped that the analysis in Section 4.2 and Section 4.3 can help to judge whether Hong Kong students' spoken English exhibit characteristics typically found in English writing.

4.2 Comparison on Mean Frequencies of Syntactic Features (Pilot Study)

This section investigates the differences between Hong Kong students' oral presentations and comparable native-speaker oral production, prepared speeches, on the level of syntactic features. The results reported here represent the preliminary findings of the pilot study. They are incorporated into Chapter 4 for two reasons: the integrity of the findings on the learner speech data and the maintenance of an organisation that is parallel in the two Results and Discussion chapters (Chapter 4 and Chapter 5). The discussion on the pilot study results has been substantially revised and expanded with the addition of qualitative analysis. In the rest of this section, statistical information on the use of 67 syntactic features in the learner speech data is provided. The statistics were generated by following the procedures described in Chapter 3 Section 3.5.3.1. The learner speech data is then compared with the prepared speech genre in Biber (1988). Most noticeable differences between the two groups of speakers' usage are explored.

4.2.1 Results

Reported in this section are (a) descriptive statistics, mean and standard deviation, on the frequency of occurrence of the 67 syntactic features in Hong Kong students' oral presentations, and (b) differences between the descriptive statistics in Hong Kong students' oral presentations and the descriptive statistics in native speakers' prepared speeches.

4.2.1.1 Use of 67 syntactic features in Hong Kong learners' oral presentations.

Table 12 presents the descriptive statistics of the learner speech data, i.e. Hong Kong students' oral presentations. The first column classifies the syntactic features into some general categories according to Biber (1988). The second column lists the 67 syntactic features adopted from Biber's (1988) model. The third column shows the mean normalised frequency for each feature, i.e. an average frequency of the feature per 1,000 words of the oral discourse. The last column reports the standard deviation for each feature. It can be observed in Table 12 that more than half of the syntactic features exhibit relatively low frequencies. Some categories, e.g. place and time adverbials (Features 4-5), questions (Feature 13) and passives (Features 17-18), have consistently low frequencies. In the category of subordination features (Features 21-38) and the category of lexical classes (Features 45-51), there is a general tendency for low frequencies, despite the existence of few exceptional cases: *that* verb complements (Feature 21), infinitives (Feature 24), causative adverbial subordinators (Feature 35) and demonstratives (Feature 51). In contrast, in the categories of tense and aspect markers (Features 1-3), nominal forms (Features 14-16) and prepositional phrases, adjectives and adverbs (Features 39-42), a trend of high frequencies is observed. Again individual members in the categories may deviate from the general trend. Overall, there is no category with consistently high frequencies.

Table 12

Mean Normalised Frequencies (M) and Standard Deviations (SD) of the 67 Syntactic Features in Hong Kong Students' Speech Data

Category	Syntactic Feature	M	SD
Tense and aspect markers	1. Past Tense	15.7	10.8
	2. Perfect Aspect	2.7	3.6
	3. Present Tense	102.1	22.2
Place and time adverbials	4. Place Adverbials	0.9	2.7
	5. Time Adverbials	2.8	3.6
Pronouns and pro-verbs	6. First Person Pronouns	39.2	25.8
	7. Second Person Pronouns	8.4	10.5
	8. Third Person Personal Pronouns	29.5	22.2
	9. Pronoun *it*	16.9	15.4
	10. Demonstrative Pronouns	4.9	5.5
	11. Indefinite Pronouns	3.9	5.6
	12. Pro-verb *do*	0.4	1.3
Questions	13. Direct *wh*-questions	0.5	1.7
Nominal forms	14. Nominalizations	18.7	16.7
	15. Gerunds	2.8	5.0
	16. Total Other Nouns	189.5	32.4
Passives	17. Agentless Passives	3.5	4.4
	18. *By*-passives	0.1	0.7
Stative forms	19. *Be* as Main Verb	14.8	9.8
	20. Existential *there*	5.2	5.7
Subordination features	21. *That* Verb Complements	9.7	9.2
	22. *That* Adjective Complements	0.1	0.8
	23. *Wh* Clauses	1.0	3.0
	24. Infinitives	20.5	12.0
	25. Present Participial Clauses	0.1	0.7
	26. Past Participial Clauses	0.2	1.3
	27. Past Participial *whiz* Deletion Relatives	0.5	1.5
	28. Present Participial *whiz* Deletion Relatives	0.6	1.8
	29. *That* Relative Clauses on Subject Position	0.7	1.7
	30. *That* Relative Clauses on Object Position	0.0	0.0
	31. *Wh* Relatives on Subject Position	1.8	3.8
	32. *Wh* Relatives on Object Position	0.1	0.8
	33. Pied-piping Relatives Clauses	0.1	0.6
	34. Sentence Relatives	0.1	0.7
	35. Causative Adverbial Subordinators	7.5	6.7

Table 12 (contd)

Category	Syntactic Feature	M	SD
Subordination features (contd)	36. Concessive Adverbial Subordinators	0.7	1.8
	37. Conditional Adverbial Subordinators	2.3	3.6
	38. Other Adverbial Subordinators	1.2	2.7
Prepositional phrases, adjectives, and adverbs	39. Total Prepositional Phrases	58.5	18.4
	40. Attributive Adjectives	42.7	27.2
	41. Predicative Adjectives	3.6	5.4
	42. Total Adverbs	40.7	17.5
Lexical specificity	43. Type/Token Ratio*	52.7	4.7
	44. Mean Word Length#	4.4	0.2
Lexical classes	45. Conjuncts	3.3	4.8
	46. Downtoners	0.9	2.0
	47. Hedges	0.4	1.3
	48. Amplifiers	4.2	5.6
	49. Emphatics	1.2	2.8
	50. Discourse Particles	0.0	0.0
	51. Demonstratives	10.9	9.5
Modals	52. Possibility Modals	13.7	12.5
	53. Necessity Modals	4.5	5.9
	54. Predictive Modals	4.6	9.4
Specialized verb classes	55. Public Verbs	4.8	5.9
	56. Private Verbs	14.2	7.9
	57. Suasive Verbs	0.7	2.5
	58. *Seem* and *appear*	0.5	1.4
Reduced forms and dispreferred structures	59. Contractions	10.2	10.0
	60. Subordinator *that* deletion	5.3	5.9
	61. Stranded Prepositions	0.1	0.5
	62. Split Infinitives	0.1	0.7
	63. Split Auxiliaries	1.9	3.5
Coordination	64. Phrasal Coordination	6.0	5.8
	65. Independent Clause Coordination	9.5	8.8
Negation	66. Synthetic Negation	0.6	2.0
	67. Analytic Negation	10.7	8.2

* The type/token ratio, calculated by dividing the number of different words by the number of tokens and then multiplying the resultant fraction by 100, refers to the average number of different words per 100 tokens.

The mean word length refers to the average number of orthographic letters in a word.

It is worth noticing that a few large standard deviation figures appear in Table 12. Present tense (Feature 3), first person pronouns (Feature 6), third person personal pronouns (Feature 8) and attributive adjectives (Feature 40) all exhibit standard deviations larger than 20. The feature of total other nouns (Feature 16) has a standard deviation of 32.4, which is the largest among the 67 syntactic features. A large standard deviation is often regarded as a sign of a high degree of variation in a set of data. However, standard deviation should not be interpreted in isolation, as it is a function of frequency. The syntactic features with large standard deviations mentioned above also have very high frequencies in the learner speech data. In other words, the unusually large standard deviations may be partially affected by the frequent occurrence of these features, as well as the large individual differences among the Hong Kong learners.

4.2.1.2 Comparison between Hong Kong learners' oral presentations and native speakers' prepared speeches.

The previous section has reported the mean frequencies of the 67 syntactic features in Hong Kong students' oral presentations. This section compares these frequencies in the learner speech data with the frequencies in native speakers' prepared speeches. The descriptive statistics of prepared speeches can be found in Appendix III of Biber (1988), but they have been reproduced in Appendix I in this thesis for the readers' easy reference. Table 13 illustrates the differences in Hong Kong students' and native speakers' use of syntactic features. Included in the table are the differences in mean normalised frequency and the differences in standard deviation. These differences were calculated by subtracting the data in native speakers' prepared speeches from the data in Hong Kong students' oral presentations.

For example, the mean frequency difference for past tense (-32.6) was obtained by subtracting the mean normalised frequency of past tense in Appendix I (48.3) from the mean normalised frequency of past tense in Table 12 (15.7).

A negative mean frequency difference means that the Hong Kong students used a smaller number of the syntactic feature concerned in their oral presentations than the native speakers did in prepared speeches. In contrast, a positive mean frequency difference suggests that the learners used a larger number of the syntactic feature than the native speakers did. As shown in Table 13, over half of the differences in mean frequency are negative: There are 37 negative mean frequency differences but only 26 positive differences. Some categories of syntactic features exhibit consistent positive mean frequency differences. These categories include questions (Feature 13), lexical specificity (Features 43-44) and coordination (Features 64-65). In other words, Hong Kong students employed more of these features in oral presentations than native speakers did in prepared speeches. Some other categories like place and time adverbials (Features 4-5), passives (Features 17-18), prepositional phrases, adjectives and adverbs (Features 39-42), and specialized verb classes (Features 55-58) show consistent negative mean frequency differences, indicating the learners' comparative underuse of these features.

Table 13

Differences in Mean Normalised Frequency (M Diff.) and Differences in Standard Deviation (SD Diff.) between Hong Kong Students' Speech Data and Native Speakers' Prepared Speech Data in Biber (1988)

Category	Syntactic Feature	M Diff.	SD Diff.
Tense and aspect markers	1. Past Tense	-32.6	-11.7
	2. Perfect Aspect	-8.6	-5.5
	3. Present Tense	+31.6	+1.1
Place and time adverbials	4. Place Adverbials	-1.0	+0.6
	5. Time Adverbials	-4.3	-0.8
Pronouns and pro-verbs	6. First Person Pronouns	-2.6	+4.4
	7. Second Person Pronouns	+3.3	+5.6
	8. Third Person Personal Pronouns	-7.6	+6.9
	9. Pronoun *it*	+8.0	+11.0
	10. Demonstrative Pronouns	-2.0	+3.4
	11. Indefinite Pronouns	+2.4	+4.4
	12. Pro-verb *do*	-2.0	-1.0
Questions	13. Direct *wh*-questions	+0.2	+1.1
Nominal forms	14. Nominalizations	-1.9	+5.2
	15. Gerunds	-2.3	+3.1
	16. Total Other Nouns	+0.4	+10.8
Passives	17. Agentless Passives	-6.1	+0.5
	18. *By*-passives	-0.1	+0.3
Stative forms	19. *Be* as Main Verb	-15.7	+4.0
	20. Existential *there*	+2.1	+4.3
Subordination features	21. *That* Verb Complements	+2.7	+4.7
	22. *That* Adjective Complements	-0.5	+0.1
	23. *Wh* Clauses	+0.8	+2.6
	24. Infinitives	+4.3	+5.4
	25. Present Participial Clauses	-0.1	+0.1
	26. Past Participial Clauses	+0.2	+1.3
	27. Past Participial *whiz* Deletion Relatives	-0.4	+0.1
	28. Present Participial *whiz* Deletion Relatives	-0.8	+0.6
	29. *That* Relative Clauses on Subject Position	+0.4	+1.0
	30. *That* Relative Clauses on Object Position	-1.6	-1.4
	31. *Wh* Relatives on Subject Position	-0.6	+1.5
	32. *Wh* Relatives on Object Position	-2.4	-1.0
	33. Pied-piping Relatives Clauses	-1.0	-1.0
	34. Sentence Relatives	0.0	+0.4
	35. Causative Adverbial Subordinators	+5.9	+5.0

Table 13 (contd)

Category	Syntactic Feature	M	SD
Subordination features (contd)	36. Concessive Adverbial Subordinators	+0.6	+1.4
	37. Conditional Adverbial Subordinators	-0.1	+2.5
	38. Other Adverbial Subordinators	+0.4	+1.3
Prepositional phrases, adjectives, and adverbs	39. Total Prepositional Phrases	-54.1	+5.9
	40. Attributive Adjectives	-6.2	+13.6
	41. Predicative Adjectives	0.0	+3.8
	42. Total Adverbs	-21.5	+7.7
Lexical specificity	43. Type/Token Ratio*	+3.7	+1.3
	44. Mean Word Length#	0.0	0.0
Lexical classes	45. Conjuncts	+2.8	+4.1
	46. Downtoners	-0.6	+1.1
	47. Hedges	+0.2	+0.7
	48. Amplifiers	+1.1	+3.5
	49. Emphatics	-3.6	-0.4
	50. Discourse Particles	-2.4	-2.4
	51. Demonstratives	-2.0	+5.3
Modals	52. Possibility Modals	+8.1	+9.4
	53. Necessity Modals	+1.9	+3.0
	54. Predictive Modals	-0.4	+6.4
Specialized verb classes	55. Public Verbs	-3.1	+1.1
	56. Private Verbs	-3.4	+3.0
	57. Suasive Verbs	-2.6	-1.0
	58. *Seem* and *appear*	0.0	+0.7
Reduced forms and dispreferred structures	59. Contractions	-3.1	-0.4
	60. Subordinator *that* deletion	+3.4	+4.3
	61. Stranded Prepositions	-3.6	-1.4
	62. Split Infinitives	+0.1	+0.7
	63. Split Auxiliaries	-3.4	+0.6
Coordination	64. Phrasal Coordination	+4.9	+4.6
	65. Independent Clause Coordination	+0.9	+5.4
Negation	66. Synthetic Negation	-1.2	+0.6
	67. Analytic Negation	+2.3	+3.9

* The type/token ratio, calculated by dividing the number of different words by the number of tokens and then multiplying the resultant fraction by 100, refers to the average number of different words per 100 tokens.

The mean word length refers to the average number of orthographic letters in a word.

Whereas positive figures occupy only slightly more than one-third of the column of mean frequency differences, they dominate in the column of standard deviation differences. Among the 67 syntactic features, 54 exhibit positive standard deviation differences, so in 80% of the cases, there seems to be a higher degree of variation among the Hong Kong students' usage than among the native speakers' usage. As suggested in Section 4.2.1.1, features with very high frequencies are more likely to have large standard deviations. Therefore, when a large standard deviation is accompanied by a low frequency, the wide variation in this set of data is more noticeable. In Table 13, there are 25 features with positive standard deviation differences and negative mean frequency differences:

- 4. place adverbials
- 6. first person pronouns
- 8. third person personal pronouns
- 10. demonstrative pronouns
- 14. nominalizations
- 15. gerunds
- 17. agentless passives
- 18. *by*-passives
- 19. *be* as main verb
- 22. *that* adjective complements
- 25. present participial clauses
- 27. past participial *whiz* deletion relatives
- 28. present participial *whiz* deletion relatives
- 31. *wh* relatives on subject position
- 37. conditional adverbial subordinators
- 39. total prepositional phrases
- 40. attributive adjectives
- 42. total adverbs
- 46. downtoners
- 51. demonstratives
- 54. predictive modals
- 55. public verbs
- 56. private verbs
- 63. split auxiliaries
- 66. synthetic negation

The smaller mean frequencies and the larger standard deviations of these features clearly demonstrate that there is wider variation in the learner speech data than in the comparable native-speaker oral production.

The greater variation among Hong Kong students' use of syntactic features may be attributed to their heterogeneous English proficiency. As illustrated in Chapter 3 Section 3.4.1.1.3, the students' general English proficiency differs considerably from one to another. Although nearly half of the participants obtained Level 3 in the English language subject in the public examination, there were also more than one-tenth of the students obtaining Level 2, i.e. the minimum passing grade, and more than one-tenth of the students obtaining Level 5 or 5*, i.e. the top most level (see Figure 11 in Chapter 3 Section 3.4.1.1.3). The students' oral English proficiency varies even more widely. While some of them were among the best student speakers in Hong Kong, some of them could not even pass the speaking examination (see Figure 12 in Chapter 3 Section 3.4.1.1.3). Consequently, the big difference in the Hong Kong students' English proficiency may be one reason behind the large variability in their use of syntactic features.

4.2.2 Discussion

After a report on the statistics concerning the differences between Hong Kong students' and native speakers' use of syntactic features in similar oral production tasks, this section examines these differences more carefully in order to provide answers for Research Question 1.1 and Research Question 1.2. However, not all 67 syntactic features are considered. The following discussion only focuses on the 20 syntactic features that show the biggest differences between Hong Kong students and native speakers. These features, the top 10 most overused and the top 10 most

underused syntactic features in Hong Kong students' oral presentations, are reproduced in Table 14 and Table 15 respectively.

Table 14

Differences in Mean Normalised Frequency for the Top 10 Most Overused Syntactic Features in Hong Kong Students' Speech Data, When Compared with Native Speakers' Prepared Speech Data in Biber (1988)

Syntactic Feature		Difference in Mean Frequency (per 1000 words)
3.	Present Tense	+31.6
52.	Possibility Modals	+8.1
9.	Pronoun *it*	+8.0
35.	Causative Adverbial Subordinators	+5.9
64.	Phrasal Coordination	+4.9
24.	Infinitives	+4.3
43.	Type/Token Ratio*	+3.7
60.	Subordinator *that* deletion	+3.4
7.	Second Person Pronouns	+3.3
45.	Conjuncts	+2.8

* Regarding the type/token ratio, the number indicated in the table refers to the difference between the average type/token ratios of the two data sets concerned.

Table 15

Differences in Mean Normalised Frequency for the Top 10 Most Underused Syntactic Features in Hong Kong Students' Speech Data, When Compared with Native Speakers' Prepared Speech Data in Biber (1988)

Syntactic Feature		Difference in Mean Frequency (per 1000 words)
39.	Total Prepositional Phrases	-54.1
1.	Past Tense	-32.6
42.	Total Adverbs	-21.5
19.	*Be* as Main Verb	-15.7
2.	Perfect Aspect	-8.6
8.	Third Person Personal Pronouns	-7.6
40.	Attributive Adjectives	-6.2
17.	Agentless Passives	-6.1
5.	Time Adverbials	-4.3
49.	Emphatics	-3.6

A word of warning should be included here for a feature in Table 14. Type/token ratio (Feature 43) is the seventh most overused feature, which means that Hong Kong students' oral presentations have a larger type/token ratio than native speakers' prepared speeches. Nevertheless, the type/token ratios in the two sets of data were not obtained by the same method (see Chapter 3 Section 3.5.2). The type/token ratio for the prepared speeches in Biber (1988) was based on the first 400 tokens of a text, but none of the transcripts in the learner speech data reached a length of 400 tokens. The type/token ratio for the learners' oral presentations was computed by using the total length of each transcript instead. Since type/token ratio is a measure sensitive to the length of a text (Read, 2000, p. 201), the type/token ratio of the learner speech data and the type/token ratio of the native-speaker speech

data are not directly comparable. Readers should bear this in mind when reading Table 14.

This discussion section is organised according to the research questions. Section 4.2.2.1 addresses Research Question 1.1: How are Hong Kong students' oral presentations different from comparable native-speaker oral production (prepared speeches), in terms of the use of syntactic features? The analysis is based on the top 10 most overused and the top 10 most underused syntactic features in Table 14 and Table 15. The features are considered with reference to their shared textual and communicative functions. To better understand characteristics of Hong Kong students' spoken English, qualitative analysis of the learner speech data has also been conducted. Section 4.2.2.2 addresses Research Question 1.2: Do the differences in the use of syntactic features observed suggest that Hong Kong students' oral presentations exhibit written-like characteristics? The 20 syntactic features listed in Table 14 and Table 15 are grouped according to the textual dimensions that they belong to in Biber's (1988) model. Through this categorisation, it can be decided whether the major differences between Hong Kong students' and native speakers' use of syntactic features hint at the written-like nature of Hong Kong students' oral English.

4.2.2.1 Characteristics of Hong Kong learners' oral presentations.

In response to Research Question 1.1, the discussion in this section is centred around some general characteristics observed in Hong Kong students' oral presentations when they are contrasted with native speakers' prepared speeches in Biber (1988). These characteristics are identified by careful consideration of the

relationships among the syntactic features in Table 14 and Table 15. In other words, these 20 cases of most deviated usage are not discussed one by one; rather, they are grouped according to the common textual and communicative functions of the syntactic features. Findings from the qualitative analysis of the learner speech data are also cited to support the discussion of the different characteristics. References to citations of the learner speech data are made by using the name of the computer file (e.g. STUAA, in which *STU* represents files from the learner speech data and *AA* represents a student identification code) and the line numbers in the tagged file (e.g. 10-20). The four general patterns of Hong Kong students' use of syntactic features to be examined in this section are (a) the heavy use of present tense constructions (Section 4.2.2.1.1), (b) the preference for a tentative tone (Section 4.2.2.1.2), (c) the reliance on noun phrases with simple structure (Section 4.2.2.1.3), and (d) the frequent use of clausal relationship markers (Section 4.2.2.1.4).

4.2.2.1.1 Popularity of present tense constructions.

The first characteristic observed in Hong Kong students' oral presentations is that Hong Kong students seem to strongly favour the use of present tense. As shown in Table 14, present tense (Feature 3) is the top most overused syntactic feature. In contrast, other tense and aspect markers investigated in the present study have been seriously underused. In Table 15, past tense (Feature 1) is the second most underused feature and the perfect aspect (Feature 2) is the fifth most underused feature. Hong Kong students' preference for present tense and their avoidance of past tense become even more evident when the frequencies in the learner speech data are directly compared with those in the native-speaker speech data (see Table 16). In the learner speech data, there are 102.1 present tense verbs per 1,000 words (SD = 22.2), which

is about 50% more than the number of present tense verbs in the native-speaker speech data (M= 70.5, SD = 21.1), but there are only 15.7 past tense verbs per 1,000 words (SD = 10.8), which is just one-third of the number of past tense verbs in the native-speaker data (M= 48.3, SD = 22.5). In the light of the above information, it seems that Hong Kong students always adopt the relatively unmarked present perspective in spoken English.

Table 16

Comparison on Mean Normalised Frequencies of Past Tense, Perfect Aspect and Present Tense in Hong Kong Students' Speech Data and Native Speakers' Prepared Speech Data in Biber (1988)

Speech Data	Past Tense	Perfect Aspect	Present Tense
HK Learners	15.7	2.7	102.1
Native speakers	48.3	11.3	70.5

In fact, the underuse of past tense forms in the learner speech data can be viewed as a direct consequence of the non-narrative focus of the elicitation task. Biber (1988) regards past tense as "the primary surface marker of narrative" (p. 223). In both Biber (1986) and Biber (1988), the co-occurrence of past tense, the perfect aspect and third person personal pronouns (i.e. different forms of *he*, *she* and *they*, excluding *it*) has been found in narrative discourse such as various kinds of fiction (Biber, 1986, pp. 393, 396, 400; Biber, 1988, pp. 102, 108-109, 136). In the present study, these three features all appear in Table 15: In addition to past tense and the perfect aspect, third person personal pronouns (Feature 8) is also among the top 10 most underused syntactic features in Hong Kong students' oral presentations. Since

these three features mark narrative discourse, their underuse suggests that there is relatively little narrative content in the learner speech data. This low level of narrative focus is not surprising given the nature of the elicitation task. The students were asked to give speeches based on some articles. The verbatim instruction that the participants received was "Read and make notes on the following passage. You will use it as the basis for your presentation. You may add your own opinions or ideas." The students' primary task was to summarise the articles and their secondary task was to express their views (for further information about the elicitation task, see Chapter 3 Section 3.4.1.1.1). It is thus understandable that little narrative is involved in the learner speech data. Moreover, narrative is not prevalent in the 16 articles used as stimuli for the oral elicitation. The article content focuses on issues related to the modern Hong Kong society (see Chapter 3 Section 3.4.1.1.1 and Appendix D). The dominant tense in these articles is present tense. Past tense is only used for recounting past events reflecting the social phenomena or for reporting people's opinions. When summarising these articles, the students naturally retained the use of present tense most of the time.

Besides the nature of the elicitation task of the learner speech data, the oral discourse used as the native-speaker speech data also helps account for the underuse of past tense observed in Hong Kong students' oral presentations. Underuse is a relative concept which can be influenced by the standard of reference adopted in the comparison. If a lot of past tense verbs are present in the native-speaker genre adopted as the standard of reference, the learners' underuse will be exaggerated. In the present study, the native-speaker speech data, i.e. the prepared speeches in Biber (1988), contains some sermons and court cases (see Chapter 3 Section 3.4.2.1.1). It is not difficult to imagine that considerable narrative exists in these two types of

speech. As a result, the underuse of past tense constructions in Hong Kong students' presentations can be largely explained by the differential nature of the oral production tasks in the learner data and the native-speaker data.

However, during the data transcription process, the researcher did notice another contributing factor to the underuse of past tense: Occasionally the student presenters used present tense forms in situations where past tense forms were required. A qualitative analysis of the learner speech data reveals that 91 cases of such errors exist. These cases, representing 9% of the total number of present tense verbs (raw frequency = 1021), should not be overlooked, especially when the total number of past tense verbs is only 155 (raw frequency). In slightly more than half of the 91 erroneous cases, the students did not use past tense forms when recounting some past events:

> in 2006 there is a campaign that called Lights Out Hong Kong which is ask the residents and companies switch their light off for three minutes to support the pollution problems <#> and but few of them participate in the campaign (STUFB:19-58)

> last year there are school children from some Hong Kong school take part in a letter-writing campaign urging the Chief Executive Donald Tsang <#> this campaign is start by a school called Beacon Hill School in Kowloon Tong <#> this action is start by primary school age between five to eleven (STUGA:33-81)

In both excerpts, the time adverbials, *in 2006* and *last year*, clearly indicate that the events happened in the past, but the students still used the present tense forms. In the remaining erroneous cases, the student presenters simply omitted the past tense inflections required in indirect speech:

but the writer said it is being practised less in Hong Kong because teenagers do not follow any particular faith (STUDB:24-42)

some biologist said that it is the only way we can make sure that enough food is satisfies for our future population (STUKB:83-104)

Therefore, Hong Kong students' inappropriate use of tense can be another factor leading to their overuse of present tense and underuse of past tense in oral presentations.

One may readily attribute Hong Kong students' tense errors to mother tongue interference because there is no tense inflection in the Chinese verb system. Nevertheless, Riddle (1986) has noticed that past tense is problematic not only to English language learners whose mother tongue has no tense inflection, but also to learners whose mother tongue has such inflection (p. 267). Moreover, in the present study, there are far fewer cases of past tense omission in the learner writing data than in the learner speech data (see Chapter 5 Section 5.2.2.1.1). It seems that, instead of negative transfer, the production conditions of the learner speech data may better explain Hong Kong students' tense errors in the present study.

Two factors in the production conditions may be relevant to the higher density of tense errors in the learner speech data than in the learner writing data. The first one is the medium of the production. The students produced the essays, i.e. the learner writing data, by writing them on paper. The relatively permanent nature of writing allowed the students more opportunities to revise and edit their essays, hence reducing the number of grammatical mistakes. In contrast, the students' oral presentations, i.e. the learner speech data, were transient in nature. When monitoring the accuracy of their utterances, the students had to resort to their limited working memory, which meant that extensive revision and editing was not possible. The

second factor affecting the number of tense errors is time. In the present study, the students were given more time to write the essays. The time limit for the production of the 500-word essay was 1 hour 15 minutes, but the time limit for the preparation of the 2-minute presentation was only 10 minutes (see Chapter 3 Section 3.4.1.1.1 and Section 3.4.1.2). Time is necessary not only for editing, but also for planning. Krashen (1982) comments that without adequate time, second language learners cannot apply grammar rules to produce appropriate language output (p. 89). Ochs (1979) has observed that in relatively unplanned discourse, present tense is often used to refer to past events because inadequate planning forces language users to depend more on early acquired morphosyntactic structures (pp. 68, 70). Færch and Kasper (1983) also suggest that second language learners may choose to omit grammatical morphemes with no contribution to the sentential meaning so as to maintain fluency in speech production (p. 42). For example, in the four excerpts of oral presentations cited in this section, the omitted past tense morphemes have nothing but grammatical functions. All in all, the popularity of present tense constructions and the rarity of past tense forms in Hong Kong students' oral presentations are affected by the task requirement, the limited planning and editing opportunities, and the learners' use of communication strategy.

4.2.2.1.2 Inclination to adopt a tentative tone.

The second characteristic that is noticeable in Hong Kong students' oral presentations is that the learners tend to use a relatively tentative speech style. Table 14 and Table 15 show that possibility modals (Feature 52) and emphatics (Feature 49) are ranked the second most overused syntactic features and the tenth most underused feature respectively. Possibility modals, as the name suggests, indicate a degree of

possibility in utterances. Emphatics such as *for sure* and *really* act as signals of certainty (Biber, 1988, p. 241). The overuse of possibility modals and the underuse of emphatics in the learner speech data hint that Hong Kong students prefer a tentative tone in spoken English.

However, modal verbs are well known for their multiple functions. To prove that the learners' overuse of possibility modals is associated with some sense of uncertainty, a decision has been made to closely examine these modals in the learner speech data. Table 17 summarises the distribution of the four possibility modals, *can*, *could*, *may* and *might*, counted in the present study. The vast majority (82%) of the possibility modals identified in the learner speech data are instances of *can* and only 13% are instances of *may*. According to Quirk et al. (1985), the four modals can express both intrinsic modality, i.e. permission, and extrinsic modality, i.e. possibility (including ability) (p. 221). Another classification is made to distinguish between these two kinds of modality. Among the 141 possibility modals (raw frequency), only two have been categorised as expressing permission:

> fifteen pupils took letters to Government House <#> they hope that they <u>can</u> meet a member from of Chief Executive Office but they <u>can't</u> (STUGC:113-135)

Excluding 12 indeterminate cases (the meaning of which are unclear due to incomplete utterances), all the remaining possibility modals (90%) are indeed expressing possibility. It seems highly likely that the relatively frequent occurrence of these modals causes Hong Kong students' oral presentations to sound more tentative than native speakers' prepared speeches.

Table 17

Percentages of Different Possibility Modals (Raw Frequency = 141) Found in Hong Kong Students' Speech Data

Possibility Modal	Percentage*
can	82%
could	4%
may	13%
might	0%

*The four percentages do not add up to 100% simply because they are rounded off to the nearest integer.

One may wonder why Hong Kong students tend to use so many possibility modals in oral presentations. Three major patterns of the learners' usage have been identified. Firstly, in about one-fourth of the cases, a sense of possibility is already present in the articles used as stimuli for the oral elicitation task. In order to capture the original meaning faithfully in their summaries, the students either repeated the whole phrases, which might contain possibility modals, from the articles or paraphrased the parts concerned by using possibility modals. This is one situation in which the student presenters used possibility modals in the learner speech data.

Secondly, sometimes the original articles do not contain any indication of possibility but the students inserted possibility modals into their paraphrases. This kind of cases accounts for one-fifth of the modals expressing possibility in the learner speech data. The following excerpt is an example:

> Alex Lai Hoi-wing a champion of a horse racing even describe it as people can get bad can badly injured <#> he may be a lucky one but Sunny Ho is not (STUNB:178-208)

The original article reads like this:

> 'You have to accept that it is a dangerous sport. Despite safety regulations, people do get badly hurt.' Alex has been fortunate and has never been seriously injured, but Hong Kong jockey Sunny Ho was not so lucky.

The writer of the article adopts a much stronger and more certain tone by using the emphatic construction *do get* and the present perfect construction *has been*. In comparison, the version that the student produced sounds much weaker and less certain. Cases like this cannot be easily explained. There are many possibilities behind the students' usage. For example, the use of possibility modals may reflect the students' comprehension problem, i.e. their inability to accurately interpret the writers' attitudes in the articles. The possibility modals may also mark the students' uncertainty towards the article content, as they were not allowed to refer to the articles during the presentation. Further studies are needed to fully understand the incentives for the learners' usage.

Finally, one-third of the possibility modals in the learner speech data appear in utterances in which the students gave their own comments or suggestions:

> I think the government can do something to encourage people to join this course the Chinese opera course <#> for example they can promote this great culture to the school and also to the society and they can provide for the course and also and provide some department for the people to join the course (STUBC:126-182)

By using the possibility modal *can*, instead of the necessity modals like *should* and *must*, the student diminished the government's responsibility in sustaining the traditional Chinese culture. Furthermore, the student greatly lowered the degree of persuasiveness in the speech. Hong Kong students' inclination to use possibility

modals seems to suggest that they do not have much confidence in voicing their opinions and they do not want to commit themselves.

Equally important to the above analysis of the learner language data is a consideration of the nature of the native-speaker speech data used for comparison in the present study. As mentioned in Chapter 3 Section 3.4.2.1.1, the prepared speeches that Biber (1988) has analysed consist of sermons, university lectures, court cases, political speeches and popular lectures. One striking similarity among these different kinds of spoken discourse is that they were all produced by relatively authoritative figures in some specific settings. According to Greenbaum and Svartvik (1990), the speakers who delivered these speeches include ministers of religion, academics, judges, a counsel and a politician (pp. 44-45). The prepared speeches were produced by these authoritative figures for professional purposes. Tentativeness is generally not encouraged in these situations. A judge cannot be uncertain in a court because he/she is the one responsible for decision-making in a trial. Likewise, a minister cannot sound unsure when preaching a sermon in church. It can thus be assumed that possibility modals, which signal uncertainty, are usually avoided in sermons, lectures, court cases and political speeches. Indeed, the mean normalised frequency of possibility modals in prepared speeches is the second lowest among the six spoken genres in Biber (1988) (the lowest being radio broadcast). Given that the overuse of possibility modals in the learner speech data is relative to the use in the native-speaker data, the production conditions which inhibit the use of possibility modals in native speakers' prepared speeches can help explain why there is an overuse in Hong Kong students' oral presentations.

4.2.2.1.3 Dependence on simple noun phrase structure.

The third characteristic observed in Hong Kong students' oral presentations concerns the structure of noun phrases. There are a few syntactic features closely related to noun phrases in Table 14 and Table 15. Total prepositional phrases (Feature 39) and attributive adjectives (Feature 40) are the most underused and the seventh most underused features in Table 15. According to Quirk et al. (1985), prepositional phrases mainly function as postmodifiers in noun phrases, as adverbials and as complementation of verbs and adjectives (p. 657). Attributive adjectives premodify nouns. Therefore, the common function of these two features is to add extra information to noun phrases. Chafe (1982) and Chafe (1985) regard these two features as devices for information integration and idea unit expansion because the features allow more information to be packed into an idea unit, hence increasing its size and complexity (Chafe, 1982, pp. 39-43; Chafe, 1985, pp. 108-109). The underuse of prepositional phrases and attributive adjectives hints that Hong Kong students do not use as many noun phrases with extensive modification as the native speakers do. In Table 14, the pronoun *it* (Feature 9) is the third most overused feature. *It* is a pronoun that could substitute for a range of items, from concrete objects to abstract concepts, from individual words to long phrases (Biber, 1988, pp. 225-226; Quirk et al., 1985, pp. 347-348). The overuse of *it* reinforces the belief that Hong Kong students prefer simple nominal structure to complex one in oral presentations.

As mentioned in the last paragraph, attributive adjectives and prepositional phrases are devices for increasing information density in noun phrases, which are major carriers of meaning in sentences. The underuse of these two features seems to imply that the noun phrases in the learner speech data are not heavily packed with information. This implication is quite puzzling, given the informational nature of the

elicitation task: The students were supposed to present not only the article content, but also their own views within two minutes. The oral presentations simply cannot be non-informational in nature. If one takes a closer look at Table 14, one will notice two other syntactic features commonly used for structuring information: phrasal coordination (Feature 64), the fifth most overused feature, and infinitives (Feature 24), the sixth most overused feature. Chafe (1982) and Chafe (1985) classify both of the features into devices for integrating and elaborating information (Chafe, 1982, pp. 39-44; Chafe, 1985, pp. 108-110). The overuse of these two features suggests that the learner speech data does contain an informational focus.

What is becoming evident in the above discussion is that Hong Kong students appear to prefer some devices over others in structuring information. For example, the mean frequency of phrasal coordination in the learner speech data is 6.0 per 1,000 words (SD = 5.8), which is more than five times the corresponding mean frequency in the native-speaker speech data (M = 1.1, SD = 1.2). An examination of various instances of phrasal coordination in the learner speech data reveals that most of them are related to nouns. Table 18 shows the distribution of different conjoined structures in Hong Kong students' oral presentations. In line with Biber (1988), phrasal coordination in the present study refers to two nouns, two adjectives, two verbs or two adverbs joined together by the coordinator *and* (see Appendix A). As shown in Table 18, the most common type of phrasal coordination is the coordination of nouns. In fact, one third of the conjoined noun phrases do appear in the articles that the students read before the presentations. However, half of these represent very common combinations such as *history and culture, fruit and vegetables,* and *health and happiness*. Only another half of the conjoined noun phrases consist of relatively idiosyncratic phrases like *thickness and flexibility* and

Confucianism and Buddhism. On the whole, borrowing from the source articles does not appear to be the main reason for the relative prevalence of conjoined noun phrases in the learner speech data. When the dominance of conjoined noun phrases in Table 18 and the underuse of prepositional phrases and attribute adjectives in Table 15 are taken into consideration, it can be deduced that even though Hong Kong students do not tend to produce heavily modified noun phrases, they can expand information by joining simple noun phrases together. It seems that the learners favour syntactically simpler constructions over more complex ones for encoding information.

Table 18
Percentages of Different Kinds of Phrasal Coordination (Raw Frequency = 61) Found in Hong Kong Students' Speech Data

Phrasal Coordination	Percentage*
noun *and* noun	77%
adjective *and* adjective	10%
verb *and* verb	7%
adverb *and* adverb	7%

*The four percentages do not add up to 100% simply because they are rounded off to the nearest integer.

Hong Kong students' dependence on simple noun phrases can be traced to three factors. First, when compared with native speakers, Hong Kong students have limited English proficiency. It is not surprising that the noun phrases that they use are, in general, syntactically simpler than the noun phrases that the native speakers use. Second, the learners in the present study were only allowed 10 minutes to

prepare for the oral presentations. Although the preparation time for the prepared speeches is not specified in the London-Lund Corpus, it is almost certain that the preparation time for sermons, university lectures, court cases and political speeches is much longer than 10 minutes. The native speakers had an opportunity to organise information in their speeches in a more integrated manner, which was not available to the Hong Kong students in the present study. Third, Hong Kong students' underuse of prepositional phrases may be a result of negative transfer from the students' mother tongue. In the present study, the mean frequency of prepositional phrases in the learner speech data (M = 58.5, SD = 18.4) is just half of the mean frequency of prepositional phrases in the native-speaker speech data (M = 112.6, SD = 12.5). The feature is not only the most underused feature in the learner speech data (see Table 15), but also the second most underused feature in the learner writing data (see Chapter 5 Section 5.2.2). It seems that Hong Kong students generally avoid using prepositional phrases. According to P.-C. Yip and Rimmington (1997), modifiers appear before noun heads in Chinese noun phrases (p. 1). Even when prepositional phrases are used as modifiers in Chinese, they are placed before the nouns that they modify (p. 26). The habit of not using post-modifiers in nouns phrases in Chinese may have contributed to Hong Kong students' avoidance of prepositional phrases in English.

4.2.2.1.4 Explicit marking of clausal relationships.

The final characteristic of Hong Kong students' oral presentations discussed in Section 4.2.2.1 is the learners' heavy reliance on explicit organisational markers. In Table 14, causative adverbial subordinators (Feature 35) and conjuncts (Feature 45) are the fourth most overused and the tenth most overused syntactic features

respectively. The mean normalised frequency of causative adverbial subordinators in the learner speech data (M = 7.5, SD = 6.7) is nearly five times the mean normalised frequency of causative adverbial subordinators in the native-speaker speech data (M = 1.6, SD = 1.7). Fung and Carter's (2007) study on discourse markers has also found that Hong Kong students overuse *because* in oral discussion (p. 429). The mean normalised frequency of conjuncts in the learner speech data (M = 3.3, SD = 4.8) is six times more than the mean normalised frequency of conjuncts in the native-speaker speech data (M = 0.5, SD = 0.7). Following Biber (1988), the counting of causative adverbial subordinators in the present study is only limited to the subordinator *because*. Like other subordinating conjunctions, *because* combines two clauses into a complex sentence (Quirk et al., 1985, p. 44). The feature of conjuncts in Biber (1988) involves more than 40 specific items, including both single-word conjuncts and multi-word conjuncts (see Appendix A for details). Conjuncts enable writers to indicate their views on how two sentences or two paragraphs should be connected (Quirk et al., 1985, pp. 631-632). Since the concepts of sentences and paragraphs are not applicable to spoken language, the two features, causative adverbial subordinators and conjuncts, can be considered very similar in their functions of organising information and expressing clausal relations. The overuse of these two features suggests that, when compared with native speakers, Hong Kong learners tend to overtly mark relationships between clauses.

More in-depth analysis has been conducted to promote understanding of Hong Kong students' use of causative adverbial subordinators and conjuncts. There are a total of 75 instances of *because* (raw frequency) in the learner speech data. According to Quirk et al. (1985), this subordinator indicates a reason relationship of a subordinate clause to a matrix clause (p. 1103). Nonetheless, some instances of

because in the learner speech data do not fulfil this role:

> I think every time I say Cantonese opera everyone will feel quite sleepy because it is different with the with our pop song (STUBB:50-72)

> Hong Kong soccer team dedicated in Homeless World Cup is a example <#> those members are homeless because they feel always feel that they are lonely and they have negative feeling of themselves (STUPA:96-127)

Reading the above excerpts, one cannot help feeling puzzled. It is really hard to imagine why the fact that Cantonese opera is different from pop music can cause people to feel sleepy. It is equally difficult to understand why negative feelings like loneliness can make people homeless. Despite the existence of the subordinator *because*, there is no causal relationship between the connected clauses. Accounting for nearly 20% of the total number of causative adverbial subordinators in the learner speech data, the use of *because* to connect irrelevant clauses is one factor contributing to the overuse of causative adverbial subordinators in Hong Kong students' oral presentations.

The remaining instances of *because* in the learner speech data, except for a few unclassified cases due to incomplete utterances, all denote causal relationships. One characteristic noticed in these instances is that some students used *because* for marking an extended or indirect relationship:

> nowadays many people choice vegetable for choose vegetable but give up meat for their lunch or for their daily life because someone some people think that the bird flu is come from the animals <#> and when the human when us eat the animals we may be will get those bird flu or others thing that effect our health (STUJA:20-78)

If the propositions in the above example are labelled in this way:

1. many people choose vegetables and give up meat
2. bird flu comes from animals
3. people may get bird flu when they eat meat from animals

the logic behind what the student said can be expressed as 2 → 3 → 1. In other words, it is because *bird flu comes from animals* that *people may get bird flu when eating meat from animals*. It is because *people may get bird flu when eating meat from animals* that *many people choose vegetables and give up meat*. The causal relation between Proposition 2 and Proposition 1, the two clauses connected by *because*, is not direct. In this example, the student presenter made the logic complete by adding Proposition 3. Nevertheless, in some cases, the students did not supply the missing link to the listeners:

> and nowadays more and more people like the Hong Kong learning western instrument <#> why they learn western instrument <#> the writer think that because the teachers are expensive for learning the Chinese instrument (STUAC:88-124)

The fact that *teachers are expensive for learning the Chinese instrument* cannot be the direct reason for *why they learn western instrument*. The two propositions in this example are:

1. more and more Hong Kong people learn western instruments
2. learning Chinese instruments from teachers is expensive

In order to understand the student' utterance, the listeners had to supply the missing link by drawing on their own encyclopaedic knowledge of the world:

3. people don't choose to learn Chinese instruments
4. people choose to learn other instruments instead

Now the student's logic, 2 → 3 → 4 → 1, could be easily understood. As in the previous example (i.e. the example from STUJA), the causal relation between Proposition 2 and Proposition 1 is an indirect one, but the student in this example considered the missing link common sense and expected the listeners to supply it by themselves. Altogether more than 20% of all instances of *because* are used for marking indirect causal relationships in the learner speech data.

In the present study, the learners' oral presentations are based on the content of some written articles. There is a possibility that the overuse of *because* is influenced by these articles. However, after a comparison of the learner speech data with the source articles, there are only five cases of *because* in the learner speech data (about 7% of the total raw frequency) that can be attributed to borrowing, owing to the presence of causal subordinators, *as* and *because*, in the corresponding parts of the articles. In other words, direct copying of *because* from the source articles is not common. Moreover, while it is true that as many as 50% of the total instances of *because* were employed by the student presenters to provide explanations for the article content, not all of these explanations originated in the articles. Sometimes the students used the subordinator to add their own reasons. Consequently, influence from the source articles is not a main cause for the overuse of *because* in the present study.

An analysis of the conjuncts used in the learner speech data shows that the articles do not have much influence on Hong Kong students' overuse of conjuncts either. None of the conjuncts were borrowed directly from the articles: None of them appear in the parts of the articles that the students were referring to during the presentations. There are only six cases, four instances of *however* and two instances of *for example*, in which the same conjuncts appear elsewhere in the articles. Based

on the discussion in these two paragraphs, the articles which the students read before the presentations have little effect on the overuse of causative adverbial subordinators and conjuncts found in the learner speech data.

Table 19 lists the different conjuncts used by the student presenters in the present study. Half of the 35 conjuncts (raw frequency) are instances of *for example*. The second and the third most common conjuncts are *however* and *moreover* respectively. Some of the conjuncts in the table (e.g. *however* and *nonetheless*) share the same functions (e.g. concession), so an attempt has been made to classify all the conjuncts according to seven semantic relations: listing, summative, appositive, resultive, inferential, contrastive and transitional (Quirk et al., 1985, pp. 634-636). The results are presented in Table 20. The category of appositive contains only one conjunct, *for example*. The category of contrastive includes *however*, *notwithstanding*, *nonetheless*, and *on the other hand* (one instance). The category of listing consists of *moreover* and *on the other hand* (one instance). The category of resultive comprises *therefore* and *as a result*. As evident from the table, Hong Kong students used conjuncts most frequently to mark appositive relations and contrastive relations.

Table 19

Percentages of Different Conjuncts (Raw Frequency = 35) Found in Hong Kong Students' Speech Data

Conjunct	Percentage*
For example	49%
However	17%
Moreover	11%
Therefore	9%
On the other hand	6%
Notwithstanding	3%
As a result	3%
Nonetheless	3%

*The eight percentages do not add up to 100% simply because they are rounded off to the nearest integer.

Table 20

Classification of Different Conjuncts (Raw Frequency = 35) Found in Hong Kong Students' Speech Data

Category of Conjunct	Percentage
Appositive	49%
Contrastive	26%
Listing	14%
Resultive	11%

Among the total number of conjuncts found in the learner speech data, nearly 30% represent problematic usage. In some cases, there is simply no connection between the two clauses and the use of conjuncts does not make any sense at all:

> as we know Italy is the birthplace of the western opera and however Hong Kong people are not that interested in Cantonese opera (STUBA:41-62)

In some other cases, the use of conjuncts is unsatisfactory because the relationships between the clauses do not totally match the choice of conjuncts:

> this article gives us a example in 1970s in LA USA <#> the governor of LA is trying to spend around ten thousand per person to solve the air pollution problems <#> that's a lot of money <#> we notwithstanding Hong Kong officers said that they can do nothing about the air pollution problem nearby in Guangdong because you know we are not going to build the Great Wall to block all the airs (STUED:96-171)

> I think that there are both sides about GM food <#> on the one hand firstly it can got more quickly with less water <#> on the other hand GM food can have rich in vitamins and tastes better <#> I suppose Hongkongers should not tamper with the nature (STUKB:135-183)

In addition to the formality of *notwithstanding* and its rarity in speech (Quirk et al., 1985, p. 636), another problem of the first excerpt is the unbalanced contrast between the American case and the Hong Kong case. Whereas the student emphasised the monetary issue in the American case, he/she focused on the attitudinal issue in the Hong Kong case. In the second excerpt, the conjunct *on the other hand*, together with *on the one hand*, appears to list two advantages of GM food. Nonetheless, the introductory phrase *both sides* seems to suggest that the student was going to talk about both advantages and disadvantages. The last comment, *I suppose Hongkongers should not tamper with the nature*, implies that the student's conclusion was drawn after the discussion of a disadvantage. It is ambiguous what kind of semantic relation *on the other hand* indicates in the second example.

In the discussion in this section, it has been demonstrated that a considerable number of clausal relationship markers, 20% of causative adverbial subordinators and 30% of conjuncts, were misused in the oral presentations produced by Hong Kong students. The misuse of these markers constitutes one reason why causative adverbial subordinators and conjuncts were overused in the learner speech data. The examples quoted in this section show that Hong Kong students sometimes use causative adverbial subordinators and conjuncts even when these devices are not necessary or not appropriate in the utterances. It seems that the students employ these clausal relationship markers as a communication strategy "to compensate for breakdowns in communication due to ... insufficient competence" (Canale, 1983, p. 10). When they encounter difficulties in expressing themselves clearly in spoken English, they turn to these clausal relationship markers for help, hoping that these magic words would facilitate the listeners' understanding.

Tannen (1985) remarks that the lexicalization of relationships between propositions is not necessary in speech due to the availability of nonverbal and paralinguistic communication channels. In her opinion, this lexicalization strategy has to be written-based (pp. 130-131). Hong Kong students' reliance on overt markers of clausal relationships suggests that they may be transferring strategies from written discourse to spoken discourse. Altenberg (1984) and Altenberg (1986) have compared causal links and contrastive links in spontaneous conversations in the London-Lund Corpus and informative prose in the LOB Corpus. According to Altenberg (1984), *because* is not particularly linked to written language: Although *because* is the most common causal link in informative prose, it occurs far more frequently in conversations (pp. 39, 41). However, five out of the eight different conjuncts identified in the learner speech data, *however, therefore, on the other hand,*

as a result and *nonetheless*, are more common in informative prose than in spontaneous conversations (Altenberg, 1984, p. 40; Altenberg, 1986, p. 16). It is quite possible that Hong Kong students' use of conjuncts in spoken English is affected by their use of conjuncts in written English.

After a discussion on the last characteristic, explicit marking of clausal relationships, a summary is provided here to finish Section 4.2.2.1. In this section, four characteristics of Hong Kong students' use of syntactic features in oral presentations have been examined in order to answer Research Question 1.1. Hong Kong students' oral presentations have been found to differ from comparable native-speaker oral production in terms of the use of tense, the degree of tentativeness, the structure of noun phrases and the linkage between clauses. First, Hong Kong students usually employ more present tense verbs than the native speakers do. The underuse of past tense verbs and perfective verbs, together with the underuse of third person personal pronouns, indicates the non-narrative nature of the learner speech data. The omission of past tense inflections observed hints that the limited planning and editing opportunities in the elicitation task lower the students' grammatical accuracy. Second, Hong Kong students have been found to use more possibility modals, but fewer emphatics, showing the learners' preference for a tentative tone. The students tend to add possibility modals when paraphrasing and when expressing their opinions. Nevertheless, it seems that the choice of reference standard has intensified the degree of uncertainty in the learners' oral presentations because the native-speaker data selected for comparison represents some speeches delivered by relatively authoritative figures in professional contexts. Third, when compared with native speakers, Hong Kong students rely more on noun phrases with

simple structure. They use fewer prepositional phrases and attributive adjectives, both of which are common modifiers of noun phrases, but they use more pronoun *it*, which can substitute for noun phrases, as well as phrasal coordination and infinitives, other devices for structuring information. It is suggested that the learners' use of simpler noun phrases may be affected by their English proficiency, the limited planning time of the oral elicitation task and mother tongue interference. Finally, Hong Kong students use more conjuncts and the subordinator *because*, suggesting that they are eager to mark clausal relationship explicitly. Qualitative analysis of the learner speech data reveals that the students mainly use appositive and contrastive conjuncts. There is hardly any evidence that the students' overuse of conjuncts and *because* is a result of direct copying from the source articles. In fact, quite a number of these clausal relationship markers involve problematic usage. It appears that Hong Kong students employ these devices to compensate for their inability to express their meanings clearly and this communication strategy may be transferred from written English to spoken English.

4.2.2.2 Relationship between the deviated features and Biber's (1988) model.

In Section 4.2.2.1, the top 10 most overused and the top 10 most underused syntactic features in Hong Kong students' oral presentations have been discussed in relation to the features' common textual and communicative functions. In this section, these 20 cases of most heavily deviated usage are further explored. To address Research Question 1.2 (Do the differences in the use of syntactic features observed suggest that Hong Kong students' oral presentations exhibit written-like characteristics?), the overused and the underused features are classified according to

their roles in Biber's (1988) multi-dimensional model. By this method, it can be determined whether these cases of overuse and underuse hint at any written-like usage in Hong Kong students' spoken English.

Table 21 and Table 22 contain all the information necessary for the analysis in this section. In the left-hand column, the top 10 most overused and the top 10 most underused features are reproduced from Table 14 and Table 15. In the middle column, each syntactic feature is categorised into the textual dimension that it belongs to in Biber (1988). The numbers, 2, 4 and 6, are put in brackets because Dimension 2, Dimension 4 and Dimension 6 are not directly relevant to the distinctions between spoken and written English. In the right-hand column, each feature classified into Dimension 1, Dimension 3 or Dimension 5 is further distinguished between the two sides of the dimension. The two ends of the dimension are shown and the underlined one is the end that the feature is loaded on. Features other than those in Dimension 1, Dimension 3 and Dimension 5 are not considered and a mark N/A is put in this column.

In Table 21, nine out of 10 most overused syntactic features belong to Dimension 1, Dimension 3 and Dimension 5, strongly indicating their relevance to the variation across speech and writing. Type/token ratio, the seventh most overused feature, is a feature on Dimension 1 in Biber's (1988) model. A higher type/token ratio usually means that the discourse is more informational. However, as mentioned at the beginning of Section 4.2.2, the type/token ratio in the learner speech data and the type/token ratio in the native-speaker speech data are not comparable owing to the different text lengths of the two sets of data. The higher type/token ratio may not truly represent the more informational nature of the learner speech data. Concerning Dimension 3, phrasal coordination, the fifth most overused feature, is a feature on

the explicit side. The overuse of this feature implies that Hong Kong students employ more explicit reference than native speakers do. Regarding Dimension 5, the tenth most overused feature, conjuncts, is loaded on the abstract side. According to Biber's (1988) model, the presence of conjuncts is a contributing factor to the level of abstractness in a piece of discourse. The overuse of this feature thus hints at the possibility of Hong Kong students' oral presentations being more abstract than native speakers' prepared speeches. In Biber's (1988) opinion, the most typical writing is characterised by an informational focus, the use of explicit reference and abstract content (p. 163). Therefore, Table 21 has shown, through the overuse of phrasal coordination and conjuncts, that the learners' presentations may be more written-like than the native speakers' prepared speeches in certain aspects.

Table 21

The Top 10 Most Overused Syntactic Features in Hong Kong Students' Speech Data, Classified According to Biber's (1988) Three Dimensions (1, 3 & 5)

Syntactic Feature		Dimension	Which Side the Feature Lies in the Dimension
3.	Present Tense	1	involved vs informational
52.	Possibility Modals	1	involved vs informational
9.	Pronoun *it*	1	involved vs informational
35.	Causative Adverbial Subordinators	1	involved vs informational
64.	Phrasal Coordination	3	explicit vs situation-dependent
24.	Infinitives	(4)	N/A
43.	Type/Token Ratio	1	involved vs informational
60.	Subordinator *that* deletion	1	involved vs informational
7.	Second Person Pronouns	1	involved vs informational
45.	Conjuncts	5	abstract vs non-abstract

Table 22

The Top 10 Most Underused Syntactic Features in Hong Kong Students' Speech Data, Classified Aceording to Biber's (1988) Three Dimensions (1, 3 & 5)

Syntactic Feature		Dimension	Which Side the Feature Lies in the Dimension
39.	Total Prepositional Phrases	1	involved vs informational
1.	Past Tense	(2)	N/A
42.	Total Adverbs	3	explicit vs situation-dependent
19.	*Be* as Main Verb	1	involved vs informational
2.	Perfect Aspect	(2)	N/A
8.	Third Person Personal Pronouns	(2)	N/A
40.	Attributive Adjectives	1	involved vs informational
17.	Agentless Passives	5	abstract vs non-abstract
5.	Time Adverbials	3	explicit vs situation-dependent
49.	Emphatics	1	involved vs informational

In contrast to Table 21, Table 22 should be interpreted in an opposite way. For example, the fourth most underused syntactic feature, *be* as main verb, is a feature of Dimension 1 and its presence adds to the level of involvement in a piece of discourse. But since Table 22 lists features that have been underused by Hong Kong students, the fact that *be* as main verb appears in Table 22 means that there are fewer such verbs in the learner speech data. The smaller number of *be* as main verb signals the less involved and hence the more informational nature of the learner speech data because the involved focus and the informational focus are conceptualised as two ends of a continuum in Biber (1988). On the same principle, the underuse of emphatics, another feature loaded on the involved side of Dimension 1, also hints at the more informational nature of the learner speech data. As for Dimension 3, the third most underused feature, total adverbs, and the ninth most underused feature, time adverbials, both lie in the situation-dependent side. The underuse of these two

features implies that Hong Kong students use less situation-dependent reference, which in turn means that they employ more explicit reference in oral presentations. Overall, Table 22 has demonstrated that, seven out of the top 10 most underused syntactic features are related to Dimension 1, Dimension 3 and Dimension 5. Among these seven cases of underuse, four suggest that Hong Kong students' oral presentations should be more written-like than native speakers' prepared speeches, as the four underuse cases show that the learners' presentations have less involved concern and less situation-dependent reference in some manner.

In summary, Table 21 and Table 22 have revealed that 16 out of the 20 cases of most deviated usage are relevant to the three textual dimensions distinguishing between spoken and written English. In particular, six of these 16 cases suggest that Hong Kong students' oral presentations may have more informational focus, explicit reference and abstract content, all of which are typical characteristics of English writing. These six cases are:

Overuse

- Phrasal Coordination
- Conjuncts

Underuse

- Total Adverbs
- *Be* as Main Verb
- Time Adverbials
- Emphatics

Consequently, Section 4.2.2.2 has answered Research Question 1.2: The differences between Hong Kong students' and native speakers' use of syntactic features suggest that Hong Kong students' oral presentations exhibit some written-like characteristics.

Section 4.2 has compared Hong Kong students' oral presentations with similar native-speaker oral production, prepared speeches, with reference to the use of syntactic features. Statistics of all 67 syntactic features have been reported and the 20 features with the biggest frequency differences between the two sets of data have been discussed. Hong Kong students' oral presentations show some interesting deviation patterns from native speakers' prepared speeches in terms of the temporal perspective, the tentative tone, the noun phrase structure and the clausal linkage. These characteristics may be affected by many factors including the production conditions of the oral presentations, transfer from the students' mother tongue, the choice of the native-speaker reference, the students' English proficiency and their use of communication strategies. It has been observed that a large proportion of the top most deviant usage in the learners' presentations is related to mode differences in English. In particular, the overuse and the underuse of six features point to the possibility that Hong Kong students' oral presentations are more written-like than native speakers' prepared speeches.

Nevertheless, as suggested in Chapter 2 Section 2.2, the relationship between spoken and written English is very complicated. One cannot conclude that Hong Kong students' spoken English is written-like simply based on a few syntactic features. For one thing, there are only six cases of most deviated usage indicative of the written-like nature of the learners' presentations, but there are nine cases pointing to an opposite possibility: Hong Kong students' oral presentations are more spoken-like than native speakers' prepared speeches. Therefore, to arrive at a more convincing conclusion, one should also take into account the multi-faceted nature of mode differences. The analysis in this section has tackled the issue of Hong Kong students' sensitivity to spoken English from a microscopic perspective, i.e. by

studying individual syntactic features. In the next section, the same issue will be investigated from a macroscopic perspective and the more abstract concepts of textual dimensions will be considered. It is hoped that these two different levels of analysis can offer the readers a more comprehensive view about Hong Kong students' sensitivity to mode differences.

4.3 Comparison on Mean Dimension Scores

After a thorough investigation of the use of syntactic features in Hong Kong students' oral presentations in Section 4.2, Section 4.3 analyses the learner speech data from another perspective. Hong Kong students' oral presentations are compared and contrasted with native-speaker production on the level of textual dimensions. Dimension scores for the learner speech data have been calculated as specified in Chapter 3 Section 3.5.3.2 and the results are reported in Section 4.3.1. A discussion then follows in Section 4.3.2: The first part of the discussion contrasts Hong Kong students' oral presentations with native speakers' prepared speeches so as to find out how they differ from one another; the second part compares the learners' presentations with common native-speaker written genres to discern similarity. Providing answers for Research Question 1.3 and Research Question 1.4, the analysis on textual dimensions in Section 4.3 should further determine whether Hong Kong students' spoken English exhibits characteristics typically found in English writing.

4.3.1 Results: Dimension Scores of Hong Kong Learners' Oral Presentations

This section offers statistical information on the dimension scores of Hong Kong students' oral presentations. Included in Table 23 are the mean scores and the standard deviations of the six textual dimensions of the learner speech data. Even though the discussion in Section 4.3.2 will only focus on three dimensions, statistics of all six dimensions are reported in the table so that the readers can get a complete picture of the learner data investigated in the present study. The mean score for Dimension 1 Involved versus Informational Production is 10.0. The mean score for Dimension 3 Explicit versus Situation-dependent Reference is 2.1. The mean score for Dimension 5 Abstract versus Non-abstract Information is -0.2. The standard deviation figure of Dimension 1 is the highest in the table. The Dimension 1 scores of the 52 presentations vary widely, from -18.8 to 38.2, resulting in a very large range of 57.1, which is about three times the range of the Dimension 3 scores (19.9) and the range of the Dimension 5 scores (18.5). In other words, there is a greater amount of variation among the student presenters in deciding how involved or how informational their presentations should be.

Table 23

Mean Dimension Scores and Standard Deviations of Hong Kong Students' Speech Data

Dimension	Mean Score	Standard Deviation
Dimension 1	10.0	11.5
Dimension 2	-3.7	2.3
Dimension 3	2.1	4.0
Dimension 4	-0.4	4.3
Dimension 5	-0.2	5.1
Dimension 6	1.4	4.0

4.3.2 Discussion

In the previous section, dimension scores of the learner speech data have been reported. For the readers' easy reference, the mean dimension scores and the standard deviations of the native-speaker speech data are reproduced from Biber (1988) and included in Appendix J. In this section, Hong Kong students' oral presentations are characterised with reference to Biber's (1988) multi-dimensional model. The results in Section 4.3.1 are discussed and emphasis is placed on Dimension 1, Dimension 3 and Dimension 5, the three textual dimensions that can distinguish between spoken and written English. Three figures have been constructed to give a graphical representation of the mean dimension scores of the learner speech data and the mean scores of other native-speaker genres in Biber (1988). It is hoped that through Figure 13, Figure 14 and Figure 15, which illustrate how different genres are situated along Dimension 1, Dimension 3 and Dimension 5, the nature of Hong Kong students' oral presentations can be revealed.

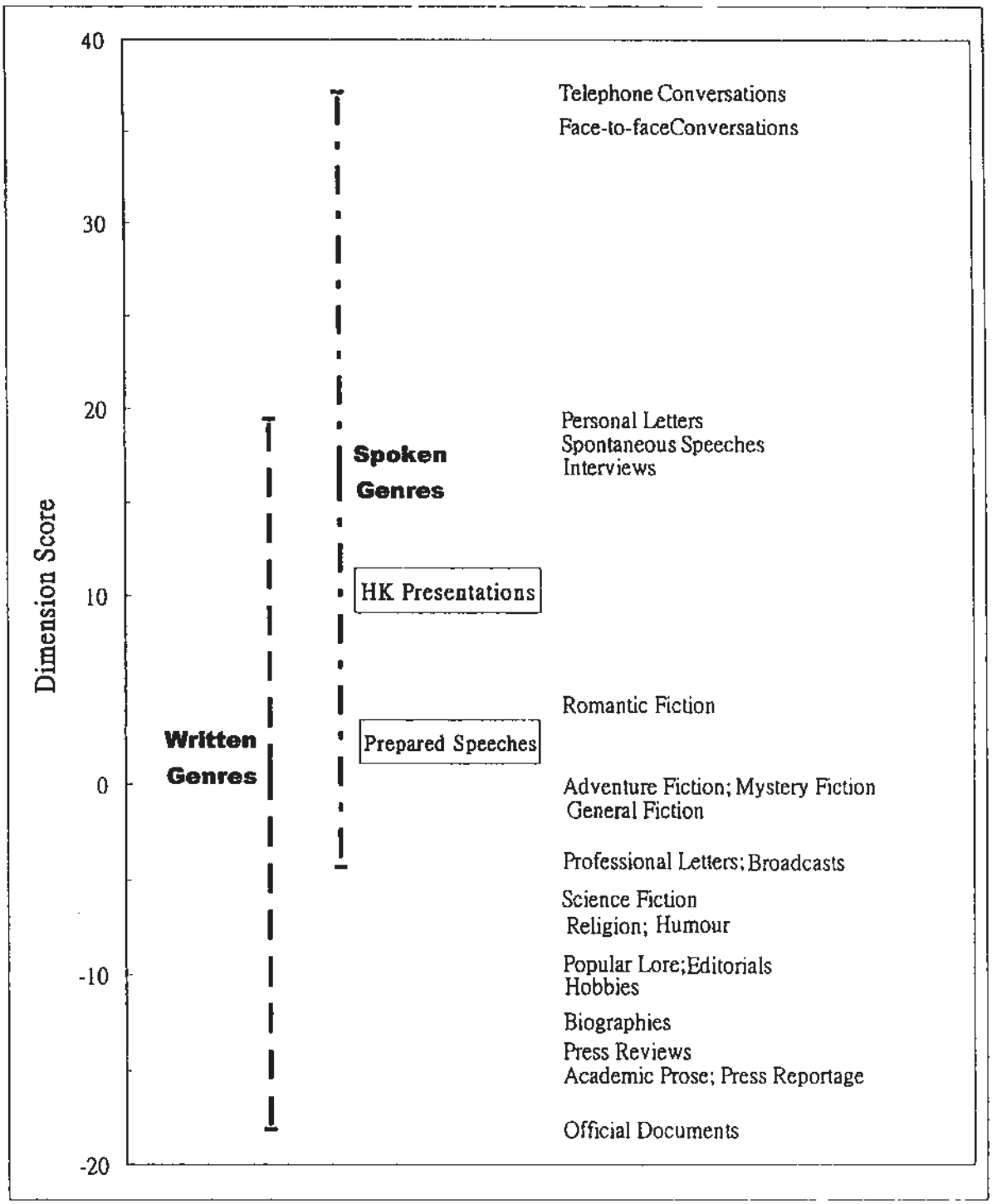

Figure 13. Distribution of the mean scores of the learner speech data and Biber's (1988) different genres (p. 128) along Dimension 1 Involved versus Informational Production

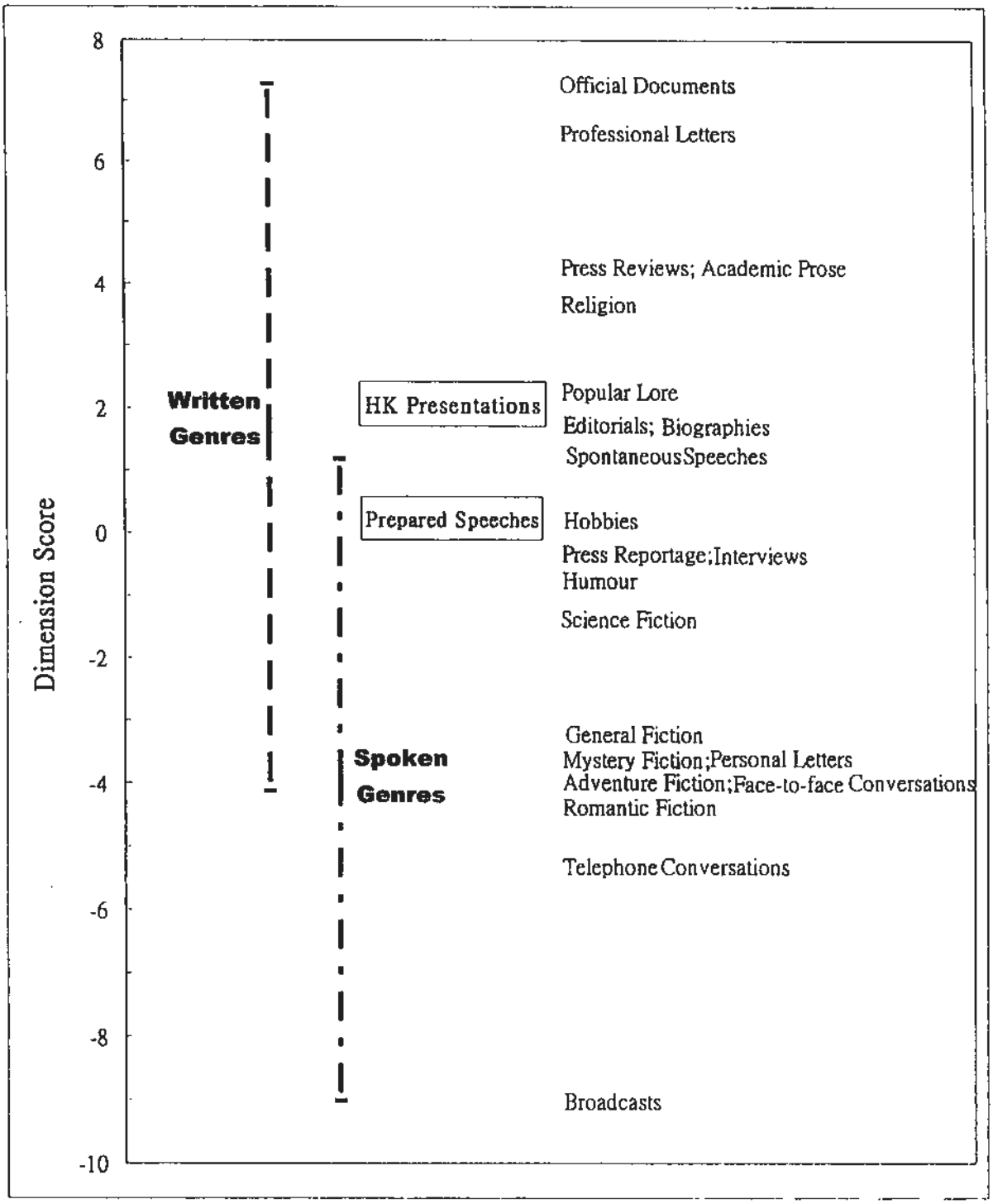

Figure 14. Distribution of the mean scores of the learner speech data and Biber's (1988) different genres (p. 143) along Dimension 3 Explicit versus Situation-dependent Reference

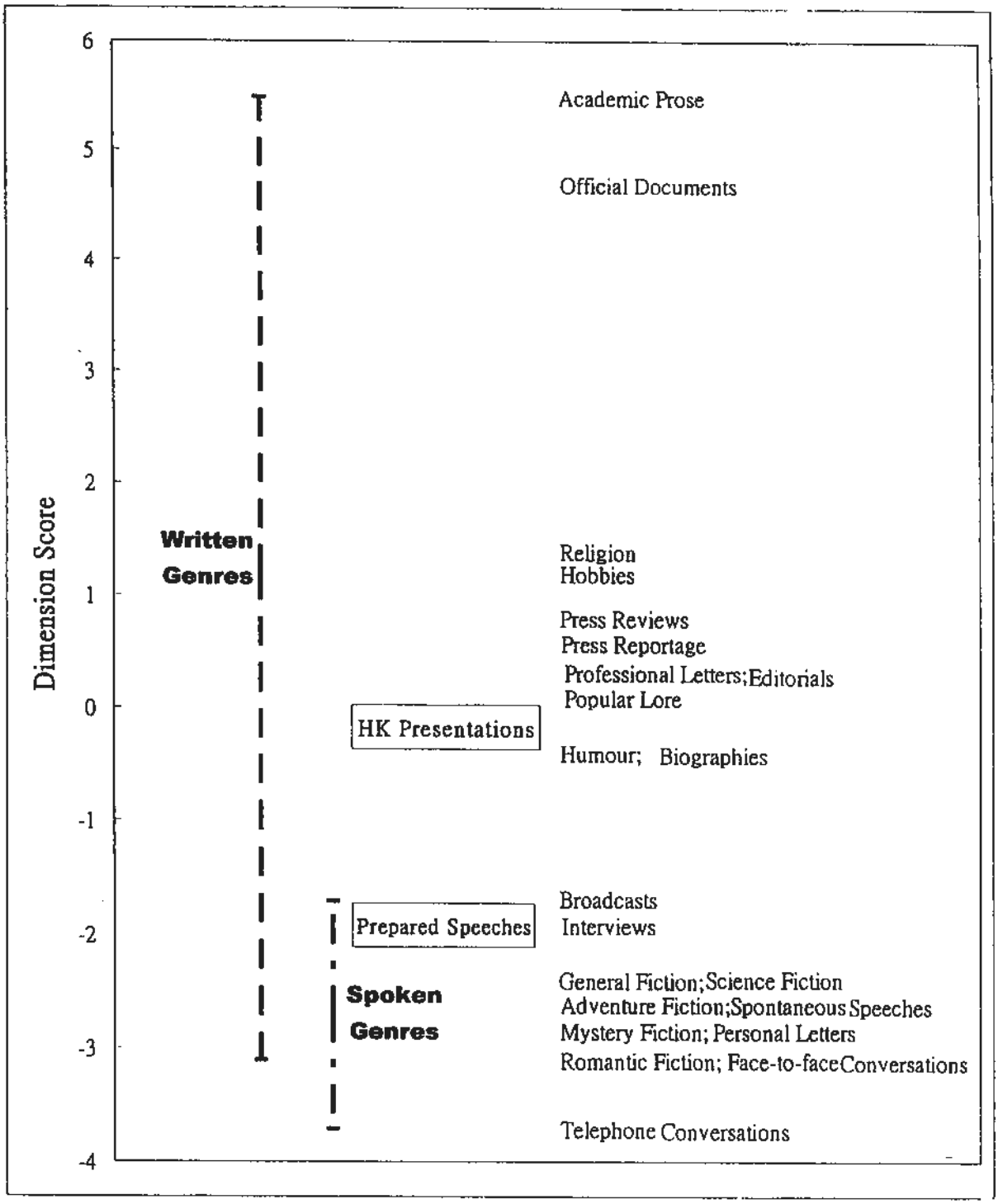

Figure 15. Distribution of the mean scores of the learner speech data and Biber's (1988) different genres (p. 152) along Dimension 5 Abstract versus Non-abstract Information

In these three figures, the vertical axis represents the scale of dimension scores. Biber's (1988) model is reproduced and different genres are plotted vertically on the right side of each graph according to their mean dimension scores. To prevent labels from overlapping each other, genres with very similar mean scores are plotted right next to each other. The right side of the graph shows labels of 22 genres studied by Biber (1988). The genre of prepared speeches is plotted separately in the middle part of the graph so as to highlight its comparison with the learner speech data (labelled as "HK Presentations"). On the left side of the graph, two vertical dotted lines are drawn to represent the whole range of scores for spoken genres and the whole range of scores for written genres in Biber's (1988) model.

In the rest of this section, Figure 13, Figure 14 and Figure 15 are examined with reference to Research Question 1.3 and Research Question 1.4. In Section 4.3.2.1, the positions of Hong Kong students' oral presentations and native speakers' prepared speeches in the three figures are compared in order to determine how the two sets of language data differ from each other. In Section 4.3.2.2, native-speaker genres located around the learner speech data are identified so that it can be decided whether there is any similarity between Hong Kong students' spoken English and native speakers' real-life written language use. The discussion in Section 4.3.2 as a whole will promote understanding of Hong Kong students' oral English and their sensitivity to mode differences.

4.3.2.1 Comparison between Hong Kong learners' oral presentations and native speakers' prepared speeches.

The discussion in this section addresses Research Question 1.3: How are Hong Kong students' oral presentations different from comparable native-speaker oral production (prepared speeches), in terms of the three textual dimensions identified in Biber's (1988) model? Figure 13, Figure 14 and Figure 15 are scrutinised, with special attention paid to the left and the middle parts of the figures. On Dimension 1 Involved versus Informational Production, Hong Kong students' oral presentations have a larger mean score than native speakers' prepared speeches (see Figure 13). Since a larger Dimension 1 score represents a higher level of involvement in Biber's (1988) model (p. 135), Hong Kong students' presentations can be considered more involved than the comparable native-speaker spoken production. In Figure 13, the dotted line of the written genres shows that common native-speaker written genres tend to have low scores along this textual dimension. In view of this, it can be decided that Hong Kong students' oral presentations are not more written-like than native speakers' prepared speeches on Dimension 1.

On Dimension 3 Explicit versus Situation-dependent Reference, the learner speech data has a higher mean score than the comparable native-speaker spoken production (see Figure 14). According to Biber (1988), the higher the Dimension 3 score, the greater the dependence on explicit reference. In contrast, the lower the Dimension 3 score, the greater the reliance on situation-dependent reference (p. 142). The higher Dimension 3 score of the learner speech data thus indicates that there is more explicit reference in Hong Kong students' oral presentations than in native speakers' prepared speeches. There is a general tendency in Figure 14 that native-speaker written genres tend to exhibit higher Dimension 3 scores than

native-speaker spoken genres. Therefore, the higher Dimension 3 score of the learner speech data also suggests that Hong Kong students' oral presentations are more written-like than native speakers' prepared speeches. Moreover, it is worth pointing out that, in the figure, whereas the label for prepared speeches is located in an area in which scores of native-speaker spoken genres and scores of native-speaker written genres overlap, the label for Hong Kong students' presentations falls outside the range of scores of native-speaker spoken genres into an area that is exclusive for native-speaker written genres. Given all the above observations, Hong Kong students' oral presentations are more written-like than the native speakers' prepared speeches along Dimension 3.

On Dimension 5 Abstract versus Non-abstract Information, Hong Kong students' presentations are found to have a larger mean score than the native-speakers' prepared speeches (see Figure 15). In Biber's (1988) framework, a higher mean score on this textual dimension refers to a higher level of abstractness (p. 151). As a result, the larger Dimension 5 score of the learner speech data indicates that Hong Kong students' oral presentations are more abstract than native speakers' prepared speeches. It also shows that Hong Kong students' oral presentations are more written-like, because it is usually the written genres that exhibit higher scores on this dimension. Another observation that is very similar to the one made in the last paragraph is that falling into a range that is exclusive for native-speaker written genres, the score of Hong Kong students' presentations is totally out of the range of scores for native-speaker spoken genres. In comparison, the score of prepared speeches lies in an overlapping area of the native-speaker spoken genres and written genres. Therefore, Hong Kong students' oral presentations are more written-like than native speakers' prepared speeches along Dimension 5.

It has been revealed in Section 4.3.2.1 that Hong Kong students (a) focus more on the involved communicative function, (b) employ more explicit reference and (c) include more abstract information in their oral presentations than native speakers do in prepared speeches. According to Biber (1988), typical written discourse usually relies more heavily on explicit reference and contains more abstract content (p. 163). The above findings clearly demonstrate that Hong Kong students' oral presentations differ from comparable native-speaker spoken production by behaving more written-like on Dimension 3 Explicit versus Situation-dependent Reference and Dimension 5 Abstract versus Non-abstract Information. At the same time, Hong Kong students' spoken English in oral presentations appears to violate native speakers' expectation of conventional spoken English on these two textual dimensions.

4.3.2.2 Comparison between Hong Kong learners' oral presentations and common native-speaker written genres.

Whereas the previous section emphasises the differences between the learner speech data and the comparable native-speaker oral production, this section is aimed at discovering potential similarities between the learner speech data and common native-speaker written genres. Figure 13, Figure 14 and Figure 15 continue to be the centre of discussion, but the subsequent discussion focuses on the right side of the figures, instead of the left and the middle parts. It is hoped that this section, together with Section 4.3.2.1, can offer evidence on the written-like nature of Hong Kong students' spoken English.

According to Figure 13, Hong Kong students' oral presentations do not bear resemblance to any native-speaker spoken genre or written genre along Dimension 1 Involved versus Informational Production. Lying in an overlapping area of the native-speaker spoken genres and the native-speaker written genres, the mean score of Hong Kong students' oral presentations (10.0) is halfway in-between the mean score of interviews (17.1) and the mean score of romantic fiction (4.3). It seems that the learners' presentations are rather unique in terms of the relative focus on involved and informational communicative functions.

In Figure 14, the label for Hong Kong students' oral presentations is right next to the labels of popular lore, editorials, biographies and spontaneous speeches. The proximity of these labels suggests that these genres are very similar on Dimension 3 Explicit versus Situation-dependent Reference. The mean score of the learner speech data (2.1) is close to the mean score of popular lore (2.3), the mean score of editorials (1.9) and the mean score of biographies (1.7). The fact that these scores are so similar indicates that Hong Kong students use as much explicit reference in their oral presentations as native speakers do in popular lore, editorials and biographies, which is a clear signal of the written-like nature of Hong Kong students' presentations. The only native-speaker spoken genre that can be considered similar to the learner speech data on Dimension 3 is spontaneous speeches. With a mean score of 1.2, the genre of spontaneous speeches represents the native-speaker spoken genre that employs the most explicit reference. Since the score of the learner speech data even exceeds that of spontaneous speeches, Hong Kong students' oral presentations can be regarded as a genre containing more explicit reference than all native-speaker spoken genres. In summary, Figure 14 shows that Hong Kong students rely much on explicit reference in oral presentations, to the extent that their

language use is very similar to native speakers' language use in written genres like popular lore, editorials and biographies.

As shown in Figure 15, Hong Kong students' oral presentations behave similarly to some written genres on Dimension 5 Abstract versus Non-abstract Information. The mean score of the learner speech data (-0.2) is close to the mean score of popular lore (0.1), the mean score of humour (-0.4) and the mean score of biographies (-0.5). In other words, Hong Kong students' presentations contain as much abstract content as these native-speaker written genres do. As mentioned in Section 4.3.2.1, the score of the learner speech data lies outside the range of scores for spoken genres. As a result, there is no native-speaker spoken genre which is comparable to Hong Kong students' oral presentations on Dimension 5. Overall, the learner speech data exhibits similarity to native-speaker written genres, but not native-speaker spoken genres, in terms of the level of abstractness.

Is there any similarity between Hong Kong students' oral presentations and common native-speaker written genres, in terms of the three textual dimensions identified in Biber's (1988) model? The discussion in this section has provided an affirmative answer to Research Question 1.4. Evidence from Figure 14 and Figure 15 has shown that the learners' use of explicit/situation-dependent reference in oral presentations resembles native speakers' use of reference in popular lore, editorials and biographies and that the amount of abstract/non-abstract content in the learners' presentations resembles the amount of abstract/non-abstract content in popular lore, humour and biographies produced by native speakers. Therefore, Hong Kong students' oral presentations are similar to native-speaker written language use on Dimension 3 Explicit versus Situation-dependent Reference and Dimension 5 Abstract versus Non-abstract Information.

4.3.2.3 Written-like nature of Hong Kong learners' oral presentations.

Section 4.3.2.1 and Section 4.3.2.2 have demonstrated clearly that Hong Kong students' oral presentations are written-like on Dimension 3 Explicit versus Situation-dependent Reference and Dimension 5 Abstract versus Non-abstract Information. What is particularly surprising is that the mean scores of the learner speech data on these two dimensions fall completely outside the range of scores of native-speaker spoken genres. The learner speech data investigated in the present study was elicited immediately after the students read some written articles. The content of the students' presentations was also based on these articles (see Chapter 3 Section 3.4.1.1.1). This special nature of the elicitation task may pose threats to research validity. It is possible to argue that the written-like nature of the learner speech data is mainly caused by the articles used as stimuli in the oral elicitation task. Consequently, it is necessary to dispel doubts about the internal validity before the end of the discussion section. In this last section of the discussion on textual dimensions, the composition of the scores for Dimension 3 and that for Dimension 5 are analysed so that the readers can evaluate the extent to which the elicitation procedure affects the findings of the present study.

Table 24 and Table 25 present the standardised frequencies of syntactic features on Dimension 3 and those on Dimension 5 respectively. In both tables, the first column indicates whether the syntactic feature is a positive one or a negative one. On Dimension 3, positive features contribute to the end of explicit reference of the continuum and negative features contribute to the end of situation-dependent reference. On Dimension 5, all features are positive ones which add to the level of abstractness of discourse. The second column lists all the syntactic features loaded

on the specific dimension. The third column gives the standardised frequencies of these features in Hong Kong students' oral presentations in the present study. Information on how these standardised frequencies were generated can be found in Chapter 3 Section 3.5.3.2. The fourth column shows the standardised frequencies of the same features in native speakers' prepared speeches in Biber (1988). The frequencies in this column were calculated by using the mean normalised frequencies of prepared speeches published in Appendix III of Biber (1988). The last column compares the standardised frequencies in Hong Kong students' oral presentations with those in native speakers' prepared speeches by subtracting the latter from the former.

In both tables, the mean dimension score of a particular set of data can be obtained by having the sum of standardised frequencies of all positive features minus the sum of standardised frequencies of all negative features. For example, the Dimension 5 score of the learner speech data can be calculated in this way:

Dimension 5 score = 1.32 + (-0.92) +0.37 + (-0.54) + (-0.65) + 0.20 = -0.22

However, the Dimension 3 score of the learner speech data as calculated from Table 24 (2.15) and the Dimension 3 score of the native-speaker speech data as calculated from the same table (0.36) differ slightly from the Dimension 3 score of the learner speech data reported in Section 4.3.1 (2.1) and the Dimension 3 score of the native-speaker speech data reported in Biber (1988) (0.3). The differences exist only because all figures in Table 24 have been rounded off to two decimal places for the clarity of data presentation.

Table 24

Comparison Between Standardised Frequencies of Syntactic Features on Dimension 3 in Hong Kong Students' Speech Data ($SFreq_{HK}$) and Standardised Frequencies of Syntactic Features on Dimension 3 in Native Speakers' Prepared Speech Data ($SFreq_{NS}$)

	Syntactic Features on Dimension 3	$SFreq_{HK}$	$SFreq_{NS}$	$SFreq_{HK}$ - $SFreq_{NS}$
Positive Features	*Wh* relative clauses on object positions	-0.76	0.65	-1.41
	Pied piping constructions	-0.56	0.36	-0.92
	Wh relative clauses on subject positions	-0.14	0.15	-0.29
	Phrasal coordination	0.95	-0.85	+1.8
	Nominalizations	-0.08	0.05	-0.13
Negative Features	Time adverbials	-0.68	0.54	-1.22
	Place adverbials	-0.64	-0.35	-0.29
	Adverbs	-1.42	-0.19	-1.23

Table 25

Comparison Between Standardised Frequencies of Syntactic Features on Dimension 5 in Hong Kong Students' Speech Data ($SFreq_{HK}$) and Standardised Frequencies of Syntactic Features on Dimension 5 in Native Speakers' Prepared Speech Data ($SFreq_{NS}$)

	Syntactic Features on Dimension 5	$SFreq_{HK}$	$SFreq_{NS}$	$SFreq_{HK}$ - $SFreq_{NS}$
Positive Features	Conjuncts	1.32	-0.44	+1.76
	Agentless passives	-0.92	0.00	-0.92
	Past participial clauses	0.37	-0.25	+0.62
	By-passives	-0.54	-0.46	-0.08
	Past participial *whiz* deletions	-0.65	-0.52	-0.13
	Other adverbial subordinators	0.20	-0.18	+0.38

On Dimension 3, the score of Hong Kong students' oral presentations (2.1) is larger than the score of native speakers' prepared speeches (0.3). So what makes the Dimension 3 score of the learner speech data larger than the Dimension 3 score of the native-speaker speech data? Since the dimension score equals the sum of standardised frequencies of all positive features minus the sum of standardised frequencies of all negative features, positive features with larger standardised frequencies and negative features with smaller standardised frequencies can both cause the Dimension 3 score of the learner speech data to become larger. In other words, positive features with positive differences in the last column of Table 24 and negative features with negative differences in the same column all contribute to the larger Dimension 3 score of Hong Kong students' oral presentations. These features, arranged in descending order of the magnitude of the standardised frequency difference, are phrasal coordination (Feature 64), total adverbs (Feature 42), time adverbials (Feature 5) and place adverbials (Feature 4). Therefore, factors that contribute to the written-like nature of Hong Kong students' oral presentations on Dimension 3, in descending order of influence, are the overuse of phrasal coordination, the underuse of total adverbs, the underuse of time adverbials and the underuse of place adverbials.

On Dimension 5, the score of Hong Kong students' oral presentations (-0.2) is also larger than the score of native speakers' prepared speeches (-1.9). In a similar vein to the explanation in the previous paragraph, syntactic features with larger standardised frequencies can lead to the larger Dimension 5 score of the learner speech data. (There is no negative feature on Dimension 5.) Features with positive standardised frequency differences in the last column of Table 25, in descending order of the magnitude of the difference, are conjuncts (Feature 45), past participial

clauses (Feature 26) and other adverbial subordinators (Feature 38). Thus, factors that contribute to the written-like nature of Hong Kong students' oral presentations on Dimension 5, in descending order of influence, are the overuse of conjuncts, the overuse of past participial clauses and the overuse of other adverbial subordinators.

As suggested in the discussion in Section 4.2.2.1, the overuse of phrasal coordination and the overuse of conjuncts are not caused by the students' borrowing phrases from the source articles to the oral presentations. Total adverbs, time adverbials and place adverbials are underused in the learner speech data and hence, direct copying from the written articles is not a problem at all. What remains is the overuse of past participial clauses and the overuse of other adverbial subordinators. An examination of the data shows that the past participial clauses (raw frequency = 2) in the learner speech data do not appear in the source articles. Two out of the total number of other adverbial subordinators (raw frequency = 13) can be found in the articles, but they were used in a different way in the students' presentations. On the whole, the investigation of different syntactic features in this chapter cannot yield any evidence of the students' borrowing constructions extensively from the written articles. Even if the research method of the present study exaggerates the written-like nature of the learner speech data, such influence cannot be the dominant factor behind the findings of the present study. For one thing, regarding the features that contribute to the written-like nature of the learner speech data on Dimension 3 and Dimension 5, the students' usage is not heavily influenced by the articles acting as the stimuli for the elicitation task. Moreover, the choice of the native-speaker reference, the genre of prepared speeches, represents one of the most written-like native-speaker spoken genres on Dimension 3 and Dimension 5. All the above pieces of evidence prove that the written-like nature of the learner speech data is not an

artefact of the research methodology of the present study; rather, it reflects Hong Kong students' language habits.

Section 4.3 has examined the textual dimensions of Hong Kong students' oral presentations. It has been found that, in comparison with comparable native-speaker oral production, the language that Hong Kong students use shifts towards a direction of typical written language use on Dimension 3 and Dimension 5. It has also been noticed that, in terms of the use of explicit/situation-dependent reference and the inclusion of abstract/non-abstract information, Hong Kong students' oral presentations are highly similar to some real-life native-speaker written genres. Especially interesting is the observation that on both Dimension 3 and Dimension 5, the mean scores of the learner speech data completely fall outside the range of scores for native-speaker spoken genres. In other words, although Hong Kong students deliver the presentations through the oral channel, some features that they use are more appropriate for writing than for speech. The findings in Section 4.3 have constituted strong evidence that Hong Kong students do not possess adequate sensitivity to mode differences in English. They experience difficulty in adjusting the proportion of different syntactic features to suit oral communication.

4.4 Summary

This chapter describes and explains findings on the speech data of Hong Kong learners. It examines Hong Kong students' oral presentations from two perspectives: the use of individual syntactic features and the overall textual dimensions. Main findings are summarised as follows:

Use of syntactic features

- When compared with native speakers, Hong Kong students overuse the following features most in their oral presentations:
  - ☐ Present Tense
  - ☐ Possibility Modals
  - ☐ Pronoun *it*
  - ☐ Causative Adverbial Subordinators
  - ☐ Phrasal Coordination
  - ☐ Infinitives
  - ☐ Type/Token Ratio
  - ☐ Subordinator *that* Deletion
  - ☐ Second Person Pronouns
  - ☐ Conjuncts
- When compared with native speakers, Hong Kong students underuse the following features most in their oral presentations:
  - ☐ Total Prepositional Phrases
  - ☐ Past Tense
  - ☐ Total Adverbs
  - ☐ *Be* as Main Verb
  - ☐ Perfect Aspect
  - ☐ Third Person Personal Pronouns
  - ☐ Attributive Adjectives
  - ☐ Agentless Passives
  - ☐ Time Adverbials
  - ☐ Emphatics
- When these 20 cases of most deviated usage are all taken into consideration, four general patterns emerge:
  - ☐ Hong Kong students rely more heavily on present tense constructions, but avoid other temporal perspectives.
  - ☐ Hong Kong students adopt a more tentative style by employing more possibility modals and fewer emphatics.

- Hong Kong students prefer syntactically simpler features like phrasal coordination for structuring information and they do not often use noun phrases heavily modified by attributive adjectives and prepositional phrases.
- Hong Kong students tend to mark clausal relationships explicitly by using conjuncts and the subordinator *because*.

- Among the 20 cases of most deviated usage, 16 features are related to the differences between spoken and written English. A total of six cases of overuse and underuse imply that Hong Kong students' oral presentations are more written-like than native speakers' prepared speeches.

Textual Dimensions

- On Dimension 1 Involved versus Informational Production:
  - Hong Kong students' oral presentations are more involved than native speakers' prepared speeches.
  - Hong Kong students' presentations are not particularly similar to any native-speaker spoken or written genres.
- On Dimension 3 Explicit versus Situation-dependent Reference:
  - Hong Kong students' oral presentations contain more explicit reference than native speakers' prepared speeches do.
  - Hong Kong students' presentations clearly fall outside the range for spoken genres; they lie solely in the range for written genres.
  - Hong Kong students' presentations are very similar to native speakers' popular lore, editorials and biographies.

- On Dimension 5 Abstract versus Non-abstract Information:
  - Hong Kong students' oral presentations are more abstract than native speakers' prepared speeches.
  - Hong Kong students' presentations clearly fall outside the range for spoken genres; they lie solely in the range for written genres.
  - Hong Kong students' presentations are very similar to native speakers' popular lore, humour and biographies.
- Hong Kong students' oral presentations are more written-like than native speakers' prepared speeches on Dimension 3 and Dimension 5. Hong Kong students' use of language is also comparable to native speakers' written language use on these two textual dimensions.

In response to Research Question 1, this chapter has proved that Hong Kong students' spoken English exhibits characteristics typically found in English writing.

CHAPTER 5 RESULTS AND DISCUSSION: WRITING DATA OF HONG KONG LEARNERS

5.1 Introduction

After a thorough analysis of Hong Kong students' speech data in Chapter 4, this chapter examines another half of the learner language data in the present study: Hong Kong students' writing data. Organised in a way that is identical to the previous chapter, this chapter begins with a comparison on the mean frequencies of the 67 syntactic features (Section 5.2) and ends with a comparison on the mean scores of the three textual dimensions (Section 5.3). For both comparisons, the statistical results are first presented and then discussed with reference to the corresponding research questions.

Section 5.2 compares Hong Kong students' written essays with similar native-speaker written production, written essays, in terms of the use of 67 syntactic features. In Section 5.2.1, the mean normalised frequencies in Hong Kong students' essays and the mean normalised frequencies in native speakers' essays, as well as the differences between these two sets of statistics, are provided. In Section 5.2.2, major differences between Hong Kong students' and native speakers' use of syntactic features are discussed. Patterns are identified to show characteristics of Hong Kong students' usage and answer Research Question 2.1. The major differences in the use of syntactic features are also considered in relation to Biber's (1988) model to determine whether they reflect a lack of sensitivity to mode differences. This analysis is intended to address Research Question 2.2.

Section 5.3 compares and contrasts Hong Kong students' written essays with native-speaker production in terms of the three textual dimensions that have been found to distinguish between speech and writing in Biber (1988). In Section 5.3.1, the mean scores of all six textual dimensions of Hong Kong students' essays and the mean scores of all six textual dimensions of native speakers' essays are presented. In Section 5.3.2, Hong Kong students' written essays are contrasted with native speakers' written essays with reference to the three relevant textual dimensions. This comparison is aimed at responding to Research Question 2.3 and showing how the learners' essays deviate from the native speakers' essays. In addition, Hong Kong students' essays are compared with some common native-speaker spoken genres to judge if the former bears any resemblance to the latter in terms of the three textual dimensions. Through this analysis, the last research question, Research Question 2.4, can be answered.

5.2 Comparison on Mean Frequencies of Syntactic Features

This section examines how Hong Kong students' written essays deviate from comparable native-speaker written production, written essays, on the level of syntactic features. Statistics on the use of 67 syntactic features in both sets of data are provided. The statistical findings are then considered in relation to Research Question 2.1 and Research Question 2.2 to see if there is any trace of spoken-like usage in Hong Kong students' written English. Information on how the mean frequencies of syntactic features were generated can be found in Chapter 3 Section 3.5.3.1. Unless otherwise specified, the frequencies reported in this section are normalised to a text length of 1,000 words.

5.2.1 Results

The results presented in this section include (a) descriptive statistics, mean and standard deviation, on the frequency of occurrence of the 67 syntactic features in Hong Kong students' written essays, (b) descriptive statistics, mean and standard deviation, on the frequency of occurrence of the 67 syntactic features in native speakers' written essays, and (c) differences between the descriptive statistics in (a) and (b) above.

5.2.1.1 Use of 67 syntactic features in Hong Kong learners' written essays.

Table 26 lists the mean normalised frequency and the standard deviation of each of the 67 syntactic features in the learner writing data, i.e. the public examination essays written by Hong Kong students. Some categories of the features exhibit consistently low frequencies. These include place and time adverbials (Features 4-5), questions (Feature 13) and passives (Features 17-18). In some other categories, a general trend of low frequencies could be observed, with individual features deviating from the norm. For example, most subordination features (Features 21-38) exhibit extremely low frequencies, but *that* verb complements (Feature 21), infinitives (Feature 24) and conditional adverbial subordinators (Feature 37) show moderate to high frequencies. The categories of lexical classes (Features 45-51), specialised verb classes (Features 55-58), and reduced forms and dispreferred structures (Features 59-63) all exhibit a general low-frequency trend, with individual features showing slightly higher frequencies. None of the categories in Table 26 have consistently high frequencies, but some are dominated by high-frequency features. Categories with a general high-frequency trend are tense and aspect markers

(Features 1-3), nominal forms (Features 14-16), prepositional phrases, adjectives and adverbs (Features 39-42), and modals (Features 52-54).

There are large standard deviations among some of the features in Table 26. The standard deviation figures for present tense (Feature 3), third person personal pronouns (Feature 8), total prepositional phrases (Feature 39), attributive adjectives (Feature 40) and total adverbs (Feature 42) are all above 20. The standard deviation of total other nouns (Feature 16) is even greater than 40. These large standard deviations indicate that there are big individual differences in Hong Kong students' use of these syntactic features. However, it should be noted that these features with very large standard deviations are also features with very high frequencies. Even though high frequencies do not necessarily guarantee large standard deviations, the computation of standard deviations do depend upon frequencies. Therefore, the above-mentioned large standard deviation figures in the learner writing data reflect not only the wide variation among Hong Kong students' usage, but also the frequent occurrence of the syntactic features concerned.

Table 26

Mean Normalised Frequencies (M) and Standard Deviations (SD) of the 67 Syntactic Features in Hong Kong Students' Writing Data

Category	Syntactic Feature	M	SD
Tense and aspect markers	1. Past Tense	12.3	12.6
	2. Perfect Aspect	2.1	3.5
	3. Present Tense	98.8	20.6
Place and time adverbials	4. Place Adverbials	1.3	2.8
	5. Time Adverbials	2.0	2.7
Pronouns and pro-verbs	6. First Person Pronouns	14.2	14.7
	7. Second Person Pronouns	12.9	19.2
	8. Third Person Personal Pronouns	33.6	25.7
	9. Pronoun *it*	14.4	10.3
	10. Demonstrative Pronouns	3.5	4.6
	11. Indefinite Pronouns	3.1	4.7
	12. Pro-verb *do*	0.5	1.5
Questions	13. Direct *wh*-questions	0.3	1.1
Nominal forms	14. Nominalizations	14.2	13.2
	15. Gerunds	4.2	6.5
	16. Total Other Nouns	193.9	41.2
Passives	17. Agentless Passives	3.4	5.2
	18. *By*-passives	0.7	1.8
Stative forms	19. *Be* as Main Verb	13.1	9.3
	20. Existential *there*	4.1	4.1
Subordination features	21. *That* Verb Complements	6.5	6.0
	22. *That* Adjective Complements	0.4	1.2
	23. *Wh* Clauses	1.3	2.8
	24. Infinitives	20.6	10.0
	25. Present Participial Clauses	0.3	1.1
	26. Past Participial Clauses	0.0	0.0
	27. Past Participial *whiz* Deletion Relatives	0.9	2.3
	28. Present Participial *whiz* Deletion Relatives	1.0	2.1
	29. *That* Relative Clauses on Subject Position	0.1	0.7
	30. *That* Relative Clauses on Object Position	0.1	0.6
	31. *Wh* Relatives on Subject Position	2.4	4.1
	32. *Wh* Relatives on Object Position	0.2	0.8
	33. Pied-piping Relatives Clauses	0.0	0.0
	34. Sentence Relatives	0.1	0.6
	35. Causative Adverbial Subordinators	2.2	3.6

Table 26 (contd)

Category	Syntactic Feature	M	SD
Subordination features (contd)	36. Concessive Adverbial Subordinators	0.5	1.5
	37. Conditional Adverbial Subordinators	6.0	5.7
	38. Other Adverbial Subordinators	1.8	4.1
Prepositional phrases, adjectives, and adverbs	39. Total Prepositional Phrases	76.2	21.5
	40. Attributive Adjectives	51.0	22.4
	41. Predicative Adjectives	5.4	7.0
	42. Total Adverbs	59.5	22.3
Lexical specificity	43. Type/Token Ratio*	54.0	4.6
	44. Mean Word Length#	4.7	0.3
Lexical classes	45. Conjuncts	10.4	7.2
	46. Downtoners	1.9	3.0
	47. Hedges	0.3	1.1
	48. Amplifiers	3.8	5.6
	49. Emphatics	1.8	2.8
	50. Discourse Particles	0.1	0.6
	51. Demonstratives	5.7	5.5
Modals	52. Possibility Modals	20.3	13.5
	53. Necessity Modals	3.7	5.1
	54. Predictive Modals	10.1	9.4
Specialized verb classes	55. Public Verbs	0.8	1.9
	56. Private Verbs	6.9	5.6
	57. Suasive Verbs	0.4	1.2
	58. *Seem* and *appear*	0.3	1.2
Reduced forms and dispreferred structures	59. Contractions	3.9	4.7
	60. Subordinator *that* deletion	1.4	2.9
	61. Stranded Prepositions	0.5	1.5
	62. Split Infinitives	0.0	0.0
	63. Split Auxiliaries	5.5	5.9
Coordination	64. Phrasal Coordination	6.7	7.7
	65. Independent Clause Coordination	2.0	3.3
Negation	66. Synthetic Negation	1.7	3.3
	67. Analytic Negation	16.7	8.5

* The type/token ratio, calculated by dividing the number of different words by the number of tokens and then multiplying the resultant fraction by 100, refers to the average number of different words per 100 tokens.

The mean word length refers to the average number of orthographic letters in a word.

5.2.1.2 Use of 67 syntactic features in native speakers' written essays.

Table 27 shows the mean normalised frequencies and the standard deviations of the 67 syntactic features in the native-speaker writing data, i.e. the LOCNESS essays. The categories of place and time adverbials (Features 4-5), questions (Feature 13) and specialised verb classes (Features 55-58) exhibit consistently low frequencies. In the categories of subordination features (Features 21-38), lexical classes (Features 45-51), and reduced forms and dispreferred structures (Features 59-63), most features have very low frequencies, with the exception of one or two features. Among the subordination features, only the feature of infinitives (Feature 24) shows a fairly high frequency. Among the lexical classes, conjuncts (Feature 45) and demonstratives (Feature 51) are the ones with moderate frequencies. Among the reduced forms and dispreferred structures, split auxiliaries (Feature 63) is the exceptional case. In contrast with the above categories with infrequent occurrence, the categories of tense and aspect markers (Features 1-3), nominal forms (Features 14-16), and prepositional phrases, adjectives and adverbs (Features 39-42) reflect a general tendency of high frequencies, with individual features showing relatively low frequencies. Overall, the distribution of high-frequency features and low-frequency features in the native-speaker writing data is rather similar to the distribution in the learner writing data.

Table 27

Mean Normalised Frequencies (M) and Standard Deviations (SD) of the 67 Syntactic Features in Native Speakers' Writing Data (LOCNESS Essays)

Category	Syntactic Feature	M	SD
Tense and aspect markers	1. Past Tense	23.7	18.8
	2. Perfect Aspect	10.0	8.6
	3. Present Tense	79.6	25.4
Place and time adverbials	4. Place Adverbials	1.8	2.5
	5. Time Adverbials	4.0	4.0
Pronouns and pro-verbs	6. First Person Pronouns	8.5	12.0
	7. Second Person Pronouns	2.1	5.4
	8. Third Person Personal Pronouns	17.9	15 0
	9. Pronoun *it*	14.8	9.5
	10. Demonstrative Pronouns	6.6	5.4
	11. Indefinite Pronouns	2.1	4.0
	12. Pro-verb *do*	0.9	2.4
Questions	13. Direct *wh*-questions	0.1	0.6
Nominal forms	14. Nominalizations	18.4	12.0
	15. Gerunds	0.9	2.3
	16. Total Other Nouns	210.7	30.3
Passives	17. Agentless Passives	10.0	6.5
	18. *By*-passives	1.1	2.0
Stative forms	19. *Be* as Main Verb	12.6	8.0
	20. Existential *there*	3.4	4.2
Subordination features	21. *That* Verb Complements	3.8	4.7
	22. *That* Adjective Complements	0.2	0.9
	23. *Wh* Clauses	0.9	1.9
	24. Infinitives	19.5	9.5
	25. Present Participial Clauses	0.6	1.5
	26. Past Participial Clauses	0.0	0.0
	27. Past Participial *whiz* Deletion Relatives	1.2	2.3
	28. Present Participial *whiz* Deletion Relatives	0.8	2.1
	29. *That* Relative Clauses on Subject Position	0.7	1.8
	30. *That* Relative Clauses on Object Position	0.3	1.2
	31. *Wh* Relatives on Subject Position	3.8	5.3
	32. *Wh* Relatives on Object Position	0.4	1.6
	33. Pied-piping Relatives Clauses	0.2	1.0
	34. Sentence Relatives	0.5	1.4
	35. Causative Adverbial Subordinators	1.8	3.5

Table 27 (contd)

Category	Syntactic Feature	M	SD
Subordination features (contd)	36. Concessive Adverbial Subordinators	0.9	2.0
	37. Conditional Adverbial Subordinators	4.0	4.7
	38. Other Adverbial Subordinators	1.6	2.8
Prepositional phrases, adjectives, and adverbs	39. Total Prepositional Phrases	89.7	19.4
	40. Attributive Adjectives	48.9	19.6
	41. Predicative Adjectives	5.7	6.3
	42. Total Adverbs	53.2	13.9
Lexical specificity	43. Type/Token Ratio*	58.5	4.0
	44. Mean Word Length#	4.6	0.3
Lexical classes	45. Conjuncts	6.7	6.3
	46. Downtoners	3.0	3.8
	47. Hedges	0.6	1.5
	48. Amplifiers	2.9	3.7
	49. Emphatics	2.2	3.5
	50. Discourse Particles	0.1	0.6
	51. Demonstratives	6.6	5.6
Modals	52. Possibility Modals	10.2	7.6
	53. Necessity Modals	3.7	4.0
	54. Predictive Modals	11.8	12.0
Specialized verb classes	55. Public Verbs	1.3	2.5
	56. Private Verbs	4.0	4.5
	57. Suasive Verbs	0.4	1.3
	58. *Seem* and *appear*	0.6	1.7
Reduced forms and dispreferred structures	59. Contractions	1.1	3.2
	60. Subordinator *that* deletion	1.2	2.5
	61. Stranded Prepositions	0.5	1.6
	62. Split Infinitives	0.3	1.4
	63. Split Auxiliaries	8.2	6.3
Coordination	64. Phrasal Coordination	6.3	5.8
	65. Independent Clause Coordination	1.4	2.8
Negation	66. Synthetic Negation	1.9	2.9
	67. Analytic Negation	7.6	6.2

* The type/token ratio, calculated by dividing the number of different words by the number of tokens and then multiplying the resultant fraction by 100, refers to the average number of different words per 100 tokens.

The mean word length refers to the average number of orthographic letters in a word.

In Table 27, the largest standard deviation, 30.3, is associated with total other nouns (Feature 16). Other than that, only one standard deviation exceeds 20 (Feature 3 present tense) and three standard deviations approach 20 (Feature 1 past tense, Feature 39 total prepositional phrases, and Feature 40 attributive adjectives). As in the case of the learner writing data, these features with large standard deviations are at the same time features with high frequencies.

5.2.1.3 Comparison between Hong Kong learners' written essays and native speakers' written essays.

Table 28 provides the differences between the descriptive statistics of Hong Kong students' essays and the descriptive statistics of native speakers' essays. The differences were obtained by subtracting the data in Table 27 from the corresponding data in Table 26. For example, the mean frequency difference for past tense (-11.4) was generated by subtracting the mean normalised frequency of past tense in Table 27 (23.7) from the mean normalised frequency of past tense in Table 26 (12.3). A negative mean frequency difference indicates that the Hong Kong students used fewer of the syntactic feature concerned than the native speakers did, whereas a positive mean frequency difference means that the learners used more of the syntactic feature than the native speakers did. Among the 67 linguistic features, 30 show positive differences in mean frequency and 34 show negative differences. In other words, about half of the features were used more frequently by the Hong Kong students and the other half were used more by the native speakers. The categories of questions (Feature 13), stative forms (Features 19-20) and coordination (Features 64-65) show consistent positive mean frequency differences. In contrast, the categories of place and time adverbials (Features 4-5) and passives (Features 17-18)

show consistent negative mean frequency differences. Other categories of syntactic features exhibit mixed results of positive and negative differences.

Regarding differences in standard deviation, 35 of the features show positive figures, hinting that the Hong Kong students' usage is more varied than the native speakers' usage; 28 of the features show negative differences, which implies that the learners' usage is less varied than the native speakers'. In the categories of questions (Feature 13), nominal forms (Features 14-15), prepositional phrases, adjectives and adverbs (Features 39-42), coordination (Features 64-65) and negation (Features 66-67), the standard deviation differences are all positive. But in the categories of tense and aspect markers (Features 1-3) and passives (Features 17-18), all the standard deviation differences are negative. In the remaining categories, both positive and negative differences exist.

As mentioned in Section 5.2.1.1, the value of standard deviation is partially affected by the size of frequency. If, in the comparison with native-speaker data, a larger standard deviation in the learner data is a signal of the wider variation in the students' usage, a larger standard deviation together with a lower mean frequency is an even stronger signal of such variation. In Table 28, there are eight instances of this type (i.e. cases with a positive standard deviation difference and a negative mean frequency difference): place adverbials (Feature 4), pronoun *it* (Feature 9), nominalizations (Feature 14), total other nouns (Feature 16), total prepositional phrases (Feature 39), predicative adjectives (Feature 41), type/token ratio (Feature 43) and synthetic negation (Feature 66). Without doubt, there is more variation concerning the use of these syntactic features within the group of Hong Kong students than within the group of native speakers in the present study.

Table 28

Differences in Mean Normalised Frequency (M Diff.) and Differences in Standard Deviation (SD Diff.) between Hong Kong Students' Writing Data and Native Speakers' Writing Data (LOCNESS Essays)

Category	Syntactic Feature	M Diff.	SD Diff.
Tense and aspect markers	1. Past Tense	-11.4	-6.3
	2. Perfect Aspect	-7.9	-5.1
	3. Present Tense	+19.3	-4.7
Place and time adverbials	4. Place Adverbials	-0.4	+0.3
	5. Time Adverbials	-2.1	-1.3
Pronouns and pro-verbs	6. First Person Pronouns	+5.7	+2.7
	7. Second Person Pronouns	+10.8	+13.8
	8. Third Person Personal Pronouns	+15.8	+10.7
	9. Pronoun *it*	-0.4	+0.8
	10. Demonstrative Pronouns	-3.0	-0.8
	11. Indefinite Pronouns	+1.0	+0.7
	12. Pro-verb *do*	-0.4	-1.0
Questions	13. Direct *wh*-questions	+0.2	+0.5
Nominal forms	14. Nominalizations	-4.3	+1.3
	15. Gerunds	+3.2	+4.1
	16. Total Other Nouns	-16.8	+10.9
Passives	17. Agentless Passives	-6.6	-1.3
	18. *By*-passives	-0.4	-0.2
Stative forms	19. *Be* as Main Verb	+0.5	+1.2
	20. Existential *there*	+0.7	-0.1
Subordination features	21. *That* Verb Complements	+2.7	+1.3
	22. *That* Adjective Complements	+0.2	+0.4
	23. *Wh* Clauses	+0.4	+0.8
	24. Infinitives	+1.1	+0.4
	25. Present Participial Clauses	-0.3	-0.4
	26. Past Participial Clauses	0.0	0.0
	27. Past Participial *whiz* Deletion Relatives	-0.3	0.0
	28. Present Participial *whiz* Deletion Relatives	+0.2	+0.1
	29. *That* Relative Clauses on Subject Position	-0.6	-1.1
	30. *That* Relative Clauses on Object Position	-0.3	-0.5
	31. *Wh* Relatives on Subject Position	-1.4	-1.2
	32. *Wh* Relatives on Object Position	-0.3	-0.7
	33. Pied-piping Relatives Clauses	-0.2	-1.0
	34. Sentence Relatives	-0.4	-0.8
	35. Causative Adverbial Subordinators	+0.5	+0.1

Table 28 (contd)

Category	Syntactic Feature	M	SD
Subordination features (contd)	36. Concessive Adverbial Subordinators	-0.4	-0.5
	37. Conditional Adverbial Subordinators	+2.0	+1.0
	38. Other Adverbial Subordinators	+0.1	+1.3
Prepositional phrases, adjectives, and adverbs	39. Total Prepositional Phrases	-13.5	+2.1
	40. Attributive Adjectives	+2.1	+2.8
	41. Predicative Adjectives	-0.3	+0.7
	42. Total Adverbs	+6.3	+8.4
Lexical specificity	43. Type/Token Ratio*	-4.5	+0.6
	44. Mean Word Length#	+0.1	0.0
Lexical classes	45. Conjuncts	+3.8	+0.9
	46. Downtoners	-1.1	-0.8
	47. Hedges	-0.3	-0.4
	48. Amplifiers	+0.9	+1.9
	49. Emphatics	-0.4	-0.7
	50. Discourse Particles	0.0	0.0
	51. Demonstratives	-0.9	-0.1
Modals	52. Possibility Modals	+10.1	+5.9
	53. Necessity Modals	+0.1	+1.1
	54. Predictive Modals	-1.7	-2.6
Specialized verb classes	55. Public Verbs	-0.5	-0.6
	56. Private Verbs	+2.9	+1.1
	57. Suasive Verbs	-0.1	-0.1
	58. *Seem* and *appear*	-0.2	-0.5
Reduced forms and dispreferred structures	59. Contractions	+2.8	+1.5
	60. Subordinator *that* deletion	+0.2	+0.3
	61. Stranded Prepositions	0.0	-0.1
	62. Split Infinitives	-0.3	-1.4
	63. Split Auxiliaries	-2.7	-0.4
Coordination	64. Phrasal Coordination	+0.4	+2.0
	65. Independent Clause Coordination	+0.6	+0.5
Negation	66. Synthetic Negation	-0.3	+0.4
	67. Analytic Negation	+9.0	+2.3

* The type/token ratio, calculated by dividing the number of different words by the number of tokens and then multiplying the resultant fraction by 100, refers to the average number of different words per 100 tokens.

The mean word length refers to the average number of orthographic letters in a word.

5.2.2 Discussion

Section 5.2.1 has presented all the statistical information concerning the use of syntactic features in both Hong Kong students' and native speakers' written essays. In this section, these results are explored in greater detail. However, given the vast amount of information present in Table 28, the discussion in this section is bound to be selective. Emphasis will be placed on the 20 syntactic features that show the most seriously deviated usage when the learner writing data is compared with the native-speaker data. Table 29 and Table 30 highlight these 20 features: Table 29 lists the top 10 most overused features in Hong Kong students' essays and Table 30 shows the top 10 most underused features.

Table 29
Differences in Mean Normalised Frequency for the Top 10 Most Overused Syntactic Features in Hong Kong Students' Writing Data, When Compared with Native Speakers' Writing Data (LOCNESS Essays)

Syntactic Feature		Difference in Mean Frequency (per 1000 words)
3.	Present Tense	+19.3
8.	Third Person Personal Pronouns	+15.8
7.	Second Person Pronouns	+10.8
52.	Possibility Modals	+10.1
67.	Analytic Negation	+9.0
42.	Total Adverbs	+6.3
6.	First Person Pronouns	+5.7
45.	Conjuncts	+3.8
15.	Gerunds	+3.2
56.	Private Verbs	+2.9

Table 30

Differences in Mean Normalised Frequency for the Top 10 Most Underused Syntactic Features in Hong Kong Students' Writing Data, When Compared with Native Speakers' Writing Data (LOCNESS Essays)

	Syntactic Feature	Difference in Mean Frequency (per 1000 words)
16.	Total Other Nouns	-16.8
39.	Total Prepositional Phrases	-13.5
1.	Past Tense	-11.4
2.	Perfect Aspect	-7.9
17.	Agentless Passives	-6.6
43.	Type/Token Ratio*	-4.5
14.	Nominalizations	-4.3
10.	Demonstrative Pronouns	-3.0
63.	Split Auxiliaries	-2.7
5.	Time Adverbials	-2.1

* Regarding the type/token ratio, the number indicated in the table refers to the difference between the average type/token ratios of the two data sets concerned.

The whole discussion section is divided into two parts. In the first part, Research Question 2.1 (How are Hong Kong students' written essays different from comparable native-speaker written production, in terms of the use of syntactic features?) is addressed. Possible relationships between the top 10 most overused and the top 10 most underused syntactic features are identified and characteristics of Hong Kong students' usage are discussed. Results from the qualitative analysis of the learner writing data and the native-speaker writing data are also cited to promote the readers' understanding of Hong Kong students' written English. In the second part, Research Question 2.2 (Do the differences in the use of syntactic features observed suggest that Hong Kong students' written essays exhibit spoken-like

characteristics?) is considered. The 20 syntactic features in Table 29 and Table 30 are classified according to their roles in Biber's (1988) model to decide whether they are indicative of any spoken-like quality in Hong Kong students' writing.

5.2.2.1 Characteristics of Hong Kong learners' written essays.

As the first part of the discussion section, this section examines the data of the present study in relation to Research Question 2.1. Table 29 and Table 30 have already illustrated how Hong Kong students' use of syntactic features differs from native speakers' use in written essays. Instead of discussing these 20 cases of most deviated usage one by one, this section groups related features together by considering their common textual and communicative functions. It is hoped that through this approach, underlying causes of the students' characteristic usage can be revealed. Whenever necessary, excerpts from the data are cited to support discussion. References to these citations are made by referring to the name of the computer file (e.g. EXMAA and BRWAA, in which *EXM* represents files from the learner writing data and *BRW* represents files from the native-speaker writing data) and the line numbers in the tagged file (e.g. 10-20). The three general patterns of Hong Kong students' use of syntactic features to be discussed in this section are (a) the heavy use of present tense constructions (Section 5.2.2.1.1), (b) the more frequent substitution of third person personal pronouns for noun phrases (Section 5.2.2.1.2), and (c) the adoption of a more interactive tone by using second person pronouns (Section 5.2.2.1.3).

5.2.2.1.1 Dominance of present time frame.

The first characteristic observed in Hong Kong students' written essays is that Hong Kong students tend to use a lot of present time reference but avoid adopting other temporal perspectives. In Table 29 and Table 30, present tense (Feature 3) is the most overused feature whereas past tense (Feature 1) and the perfect aspect (Feature 2) are the third most underused and the fourth most underused features. Table 31 reproduces the frequency information of these three features from Table 26 and Table 27. The number of past tense verbs in the learner writing data is just half of the number in the native-speaker writing data. The number of perfective verbs in the learner data is only one-fifth of the number in the native-speaker data. It is evident from the table that Hong Kong students rely much less on past tense and the perfect aspect when constructing sentences. They seem to prefer using present tense most of the time. In fact, the same observation has already been made in the comparison between Hong Kong students' oral presentations and native speakers' prepared speeches (see Chapter 4 Section 4.2.2.1.1). Therefore, the dominance of present tense constructions can be regarded as a characteristic of Hong Kong learners' English that appears across both speech and writing.

Table 31

Comparison on Mean Normalised Frequencies of Past Tense, Perfect Aspect and Present Tense in Hong Kong Students' Writing Data and Native Speakers' Writing Data (LOCNESS Essays)

Writing Data	Past Tense	Perfect Aspect	Present Tense
HK Learners	12.3	2.1	98.8
Native speakers	23.7	10.0	79.6

The prevalence of present tense and the dearth of past tense are a well-known feature of Hong Kong English (e.g. Gisborne, 2009, p. 161; Platt, 1982, p. 410). Many researchers have attempted to provide explanations for this characteristic usage among Hong Kong second language learners and one of the most popular reasons is the distinct morphological systems in Chinese and English. In Chinese, verbs are not inflected for tense, but in English, they are. This difference in the two language systems may pose considerable difficulty for Hong Kong students (Milton, 2001, p. 32; Webster, Ward & Craig, 1987, p. 69). Due to negative transfer, the learners may fail to inflect English verbs for past tense use, hence producing errors (Milton, 2001, p. 32). They may also avoid using past tense by referring rarely to past time (Milton, 2001, p. 32). Besides the influence of mother tongue, Milton (2001) proposes other reasons for the overuse of present tense and the underuse of past tense: institutional influence, which means that students have very few opportunities to practise using tense and aspect, other than the unmarked simple present, in their written assignments, and "syntactic, semantic, pragmatic and strategic considerations" (p. 32), which he does not specify.

In the present study, the frequent occurrence of present tense verbs in Hong Kong students' essays can be justified by the writing task instructions, which require the students to discuss some general issues such as world peace and university education. But considering the equally general nature of the native speakers' essay topics, the underuse of past tense in the learner writing data becomes a puzzle. Did the students simply omit the past tense morphemes? Or did they avoid past reference? In order to diagnose the problem, all verbs in the learner writing data have been carefully examined. There are altogether 1,110 present tense verbs (raw frequency) in the learner writing data, but only about 1% represent cases of missing past tense

inflection. In other words, such errors are very rare in the learner writing data of the present study. This finding matches Yang and Huang's (2004) observation that when admitted to university, Hong Kong students "have basically acquired the norms of English tense-aspect use" (p. 65), as the essays in the learner writing data were composed by secondary school leavers who wanted to apply for university admission and who sat for a public examination. The extremely low error rate shows that the underuse of past tense verbs in the Hong Kong students' written essays is not caused by the students' erroneous usage, but by their avoidance strategies.

Among all the past tense verbs (raw frequency = 137) used in the learner writing data, over 50% are modal verbs, the majority of which are *would* and *should*. In addition to past time reference, the past tense forms of modal verbs can indicate hypothetical meaning, tentativeness and mood (Quirk et al., 1985, pp. 231-236). An examination of sentences containing these modal verbs suggests that most of them have no relation at all to past time reference:

> In my opinion, talks and seminars about world peace should be held. In the talk or seminar, the content should mainly concern about the importance of world peace. Besides, I think the talk or seminar should more systematic and graphically, make it easy to understand, and make it more interesting in order to attract the young people to join. (EXMAF:184-254)

> As nearly all shops offer discounts for customers during Christmas holidays, people would like to buy gifts for their parents, friends and teacher to appreciate their concerns and cares on them. People with higher purchasing power would spend on the gifts lavishly. (EXMDA:3-51)

Another 13% of the past tense verbs in the learner writing data are cases of misuse: Present tense rather than past tense should be used. As a result, only the remaining past tense verbs, about one-third of the total number, are indeed referring to states or

actions in the past. When writing the essays, the students used these past tense verbs when they gave examples of past events to support their arguments:

> Firstly, young people should read the newspaper and watch the TV news everyday. From the newspaper and the TV news, we not only know the news of local place itself, but also realize what is happening around the world. For example, few days early, NATO decided to attack Yugoslavia to punish her suppressing the independent movement of other minority races. If we had never read the news, we could have never gotten the knowledge and happening in the world. (EXMAD:171-267)

> Also, the mobile phones produce annoyance in the schools and even in the examination rooms. Yesterday, when I was having my Chinese examination, a mobile phone from one of the candidates rang. It totally disturb me and I did not know where I was up to after this phone. (EXMCE:92-151)

In the first example, the student writer referred to a piece of international news to support the point that news exposure helped one understand global issues. In the second example, the student cited personal experience to illustrate how annoying mobile phones could be. In both excerpts, the students' supporting details are past events, which is reflected not only in the use of past tense verbs, but also in the use of time adverbials, *few days early* and *yesterday*. Since past tense verbs are a minority in the learner writing data, one can easily imagine that most often the Hong Kong students argued for their view within the present time frame:

> Moreover, the young can promote international understanding through the mass media such as television, radio, Internet. It can conduct the message in the world. This method are very effective to get the message quickly. In fact, there are many organization which walk to street to conduct the message of keeping friendly relationship to other countries. (EXMAJ:110-179)

> In audio centres, audiences are intended to have an appreciation of the music or to have relaxations. But sometimes, since some people may forget to turn the mobile telephones off, the shows or concerts may therefore interferred. (EXMCG:46-90)

These two students wrote on the same topics as the ones in the two previous excerpts (from EXMAD and EXMCE). Nevertheless, instead of using past events as support, these two students described what might happen in daily life. In this way, they could adopt the habitual present perspective and write in a timeless manner. Examples like these two are very common in the learner writing data.

As mentioned at the beginning of this section, the native-speaker writing data contains many more past tense verbs than the learner writing data does. A total of 285 past tense verbs (raw frequency) have been used in the native speakers' essays in the present study and nearly 40% of them refer to past events. Like the Hong Kong students, the native speakers used these past events as examples to support their arguments. But unlike the Hong Kong students, the native speakers did this more extensively. Not only did they give more such examples in their essays, but they also included more specific details of the past events:

> Whilst such behaviour cannot be condoned, it must be remembered that boxers do realise the risk of their chosen profession, just as other sportsmen do. When Jim Murray died, fresh calls were made for the abolition of boxing. However, when Ayrton Senna crashed and died at Imola in 1994, an equally tragic event, few calls were made to ban motorracing. On the contrary people were discussing new rules for the bodywork of the car, rather than why there was only one set of tyres on one of the most notorious corners in motor racing. Both sportsmen realised the risks that the were taking when setting out in their chosen profession, and sadly, both paid the ultimate price. (BRWBE:84-221)

> It has been shown to have been present in world war 2; when German troops killed Jews they did not hold themselves responsible as they had been told to do it, they had been set on the course by someone else. Similarly, the people in command in many scientific situations easily commit themselves to decisions because there is no immediacy for them, i.e. they rarely see their decisions face-to-face. Using the analogy of the persecution of the Jews in ww2 again, it was said that Goebbels was physically sick when he actually visited the concentration camps and saw what happened but could still order the camps to carry out their work because he had no sense of immediacy of the situation when he was sat in his office in Berlin. (BRWFE:1-144)

It can be illustrated from these two examples that native-speaker writers tend to elaborate on the past events mentioned in their essays, which is something that Hong Kong students are not able to do. This difference in writing skills may also contribute to the relatively low frequency of past tense verbs in Hong Kong students' writing.

As evident from the qualitative analysis of the learner writing data and the native-speaker writing data, Hong Kong students give fewer examples of past events, as well as less elaborated examples, to support their arguments than the native speakers do. Therefore, this difference in persuasive strategies can be one factor leading to the overuse of present tense and the underuse of past tense in Hong Kong students' written essays. Whereas the native speakers tend to cite definite incidents happening in the past to give specific support to their arguments, the Hong Kong students usually adopt a more general approach, appealing to the readers' common sense and creating scenarios from everyday experience. This strategy adopted by the learners enables them to write without any specific reference of time, explaining why the unmarked present tense is used most of the time but past tense and the perfective aspect, both of which require a clear indication of time, are used much less often. Further evidence to this explanation is the underuse of time adverbials (Feature 5, the tenth most underused feature) in Table 30, which hints that Hong Kong students tend

not to mark time information. Milton (2001) also observes that temporal adverbs are underused in Hong Kong students' writing (p. 86). As a result, the dominance of present tense verbs in Hong Kong students' expository or argumentative essays can be attributed to the particular persuasive strategy adopted by the students. Hong Kong students' preference for present tense can be considered relevant to their limited English proficiency, their less than satisfactory writing skills and the time constraint of the writing task.

5.2.2.1.2 Greater reliance on pronouns for nominal functions.

The second characteristic observed in Hong Kong students' written essays concerns the unusually frequent occurrence of third person pronouns. In Table 29, the feature of third person personal pronouns (Feature 8) is the second most overused feature. Pronouns are traditionally defined as substitutes for noun phrases (Wales, 1996, p. 1). Although the actual functions of pronouns are far more complex, this definition applies quite well to third person pronouns (Wales, 1996, p. 21). The nature of the third person pronouns is rather different from that of the first person and the second person pronouns:

> The 3rd person pronouns may also be used situationally, to refer to some person(s) or thing(s) whose identity can be inferred from the extralinguistic context... Much more frequently, however, the identity of the referents of 3rd person pronouns is supplied by the linguistic context. (Quirk et al., 1985, p. 347)

In other words, even though third person pronouns can be used for both situational reference, i.e. what Halliday and Hasan (1976) call "exophora" (p. 33), and textual reference, i.e. what Halliday and Hasan (1976) call "endophora" (p. 33), it is the

latter function that is more common in English. Given that the learner data under discussion is some written texts, but not spoken discourse, it can be quite safely assumed that the dominant function of the third person personal pronouns in such data is the endophoric one. Consequently, the overuse of these pronouns in Hong Kong students' essays suggests that noun phrases are more often substituted by these pronouns in the learners' writing than in the native speakers' writing, a view which is confirmed by some information in Table 30. In this table, total other nouns (Feature 16) and nominalizations (Feature 14) are the most underused and the seventh most underused syntactic features. Underused are not only these linguistic units functioning as heads of noun phrases, but also prepositional phrases, a common post-modifiers in noun phrases. The feature of total prepositional phrases (Feature 39) is the second most underused feature in the learner writing data of the present study. Therefore, the overuse of third person personal pronouns, together with the underuse of total other nouns, total prepositional phrases and nominalizations, shows that Hong Kong students, when compared with native speakers, rely more heavily on pronouns to fulfil nominal functions in writing.

Table 32 has been constructed to better illustrate the differential use of nominal structures between Hong Kong students and native speakers. This table recapitulates the mean normalised frequencies of major nominal structures including third person personal pronouns (Feature 8), pronoun *it* (Feature 9), nominalizations (Feature 14) and total other nouns (Feature 16) in the two sets of writing data. Following the practice of Biber (1988), the present study excludes the pronoun *it* from the count of third person personal pronouns. Thus what has been counted in Feature 8 is the subjective, objective, possessive and reflexive pronouns of *he*, *she* and *they* (see Appendix A). The pronoun *it* is counted as a separate feature. Likewise, the

frequency of total other nouns and the frequency of nominalizations are independent of each other, as the count of total other nouns excludes the count of nominalizations and that of gerunds (see Appendix A). It can be observed from Table 32 that, on the one hand, the number of third person personal pronouns in Hong Kong students' essays nearly doubles the number of these pronouns in native speakers' essays. On the other hand, there are far fewer nouns, including nominalizations, in the learners' essays than in the native speakers' essays. When the frequencies of all four features in Table 32 are added up, the sum in the learner data (256.1) and the sum in the native-speaker data (261.8) are very close to each other. It seems that Hong Kong students substitute more noun phrases with third person pronouns, hinting that their reliance on grammatical cohesion is greater than the native speakers'.

Table 32

Comparison on Mean Normalised Frequencies of Major Syntactic Features Serving Nominal Functions in Hong Kong Students' Writing Data and Native Speakers' Writing Data (LOCNESS Essays)

Writing Data	3rd Person Pronouns	Pronoun *it*	Nominalizations	Total Other Nouns
HK Learners	33.6	14.4	14.2	193.9
Native speakers	17.9	14.8	18.4	210.7

The Hong Kong students' heavier use of third person pronouns and their lower dependence on nouns suggest that they have a smaller vocabulary size than the native speakers. This is unsurprising, as it is generally believed, and empirically demonstrated too, that second language learners of English use a narrower range of

vocabulary than native speakers do (Nation, 2001, p. 178). In fact, evidence from Table 30 also lends support to this belief. The type/token ratio (Feature 43) of the learner data is smaller than the type/token ratio of the native-speaker data by 4.5. Since the essays in the two sets of data are comparable in length, a smaller type/token ratio means less varied vocabulary in Hong Kong students' writing. Hinkel (2002) has also observed that the second language learners in her study, i.e. Chinese, Japanese, Korean, Vietnamese, Indonesian and Arabic university students studying in America, overuse third person pronouns (subjective, objective, possessive and reflexive pronouns of *he*, *she*, *they* and *it*) in writing (p. 86). The researcher suggests that second language learners lack "the lexical repertoire" (p. 88) necessary for using lexical substitution and hence rely on third person pronouns for establishing cohesion.

A qualitative analysis of all the third person personal pronouns (raw frequency = 377) employed in Hong Kong students' written essays reveals that as much as 30% contain cohesive problems. The erroneous usage mainly falls into two categories: ambiguous ties and non-existent ties. Regarding the first category, two or more possible antecedents appear in the preceding text, creating ambiguity in the interpretation of the anaphor concerned:

> Many family problem raised because the wife and husband get married before they know each other deep. Consequently, they will break up because of the conflict in daily lives. And the children will be hurt seriously. So living together before marry can avoid this problem. They can even change their mind if they really can't live together harmony. (EXMEG:112-185)

In this excerpt, the third person plural pronouns in the last sentence, *they* and *their*, refer to the phrase *the wife and husband.* Due to the presence of another plural, animate noun phrase, *the children*, the readers may be confused as to what the underlined pronouns mean when they first read the passage. One fifth of the ambiguous ties can be attributed to the long distance (a distance of four sentences or above) between the third person personal pronouns and the intended antecedents:

> Although we are actually caught in a dilemma in considering whether buying mobile telephones or not, I personally strongly believe that they are the most annoying, unnecessary and time-wasting device. Just say how it's annoying first. In our daily lives, different special sound can be heard everywhere such as in shopping arcades, restaurants and MTR. It is centred around us just like an evil persuading us to enjoy it. In addition, not only simply the sound, but also various types of sound effects are troublesome. Let's imagine, they sound altogether like a noisy concert instead of a harmonious symphony.
>
> Moreover, if they are useful, we can still accept. However, they are totally unnecessary. (EXMCB:0-43)

In this extreme case, the two anaphors, the underlined *they*, and the intended antecedent, *mobile telephones*, are separated by six to seven sentences, as well as a paragraph boundary! According to Carroll (1999), long-distance anaphora can cause breakdowns in communication, as it does not conform to conventional usage and it places a heavy burden on the readers' working memory (p. 160). But ambiguous ties do not necessarily occur over a long distance; they may occur in successive sentences, too:

> Employers do not like their employees who are very lazy and do their personal things during working hours. They also hope their employees will put their heart to make a good job. (EXMBH:181-217)

In this example, the intended antecedents of the third person plural pronouns shift back and forth between the *employers* and the *employees*: The first *their* (in *their employees*) refers to the noun *employers* and the second *their* (in *their personal things*) refers to the noun phrase *their employees*. But what does the third pronoun *they* refer to? The expectation set up in the previous reference is that this *they* should refer to *their employees*. However, what the student writer really means is the *employers*. In order to resolve ambiguity like this, the readers can only resort to their common sense. Altogether 20% of all the third person personal pronouns in Hong Kong students' essays in the present study are ambiguous.

Another 10% of the third person personal pronouns in the learner writing data fall into the second category of cohesive problems, non-existent ties. The ties are considered non-existent because the antecedents of the anaphors do not appear at all. Sometimes the student writers simply forgot to put the antecedents in the preceding texts. But in some other cases, the intended antecedents, though present, do not match the anaphors and hence the cohesive ties cannot be established successfully. The mismatch can be either grammatical or semantic in nature, as shown in the following two examples:

> Everybody around the world can have Christmas, and they should. (EXMDB:27-38)

> Arguments between couples can make one to be more mature. Accepting one's view can shorter the distances between two of them. (EXMEH:44-70)

In the first example, the anaphor *they* and the antecedent *everybody* do not agree in number. The tie between the two is thus not valid. In the second example, the confusion arises from a part-whole relationship: The anaphor *them* actually refers not

to the bigger unit, *couples*, but the smaller component of this unit, i.e. individuals married as couples. Again, the cohesive tie is not set up successfully.

The finding that as much as 30% of third person personal pronouns in Hong Kong students' essays involve problematic use of grammatical cohesion is worth attention. Pronouns are one of the grammatical features that are usually introduced at the very early stage of formal instruction. According to Curriculum Development Council (2004), subjective and objective cases of personal pronouns, including all first person, second person and third person ones, are introduced at Key Stage 1, i.e. Primary 1-3, under the Hong Kong education system (p. 24). Most primary school English textbooks contain grammar drills on endophoric reference like the examples given in Curriculum Development Council (2004):

> I have a cat. It is very naughty.
> Mr. Chan is a good teacher. We like him very much. (p. 24)

Why do Hong Kong students, who have learnt this pronominal reference ever since the age of 6, still fail to master such English usage at the age of 18? This grammatical problem is not limited to the less proficient students. When the 52 essays in the learner writing data were divided into two halves according to their scores, it was discovered that the essays with higher scores contributed over 40% of the total number of ambiguous and non-existent cohesive ties. In other words, proficient student writers and less proficient student writers alike encounter problems in using third person personal pronouns for endophoric reference.

One possible explanation for Hong Kong students' less than perfect use of third person personal pronouns is interference from Chinese. Being a highly contextualised language, Chinese allows reference of previously mentioned concepts

to be omitted (Walls & Walls, 2009, p. 216; P.-C. Yip & Rimmington, 1997, p. 142). In both written Chinese and spoken Cantonese, omissions of subjects and objects whose meanings are obvious from the textual or situational context are common (Matthews & Yip, 1994, pp. 68, 77, 83; Tiee & Lance, 1986, pp. 7-8; Walls & Walls, 2009, p. 216; P.-C. Yip & Rimmington, 1997, pp. 137, 142). This kind of ellipsis, or what Tao and Healy (2005) call "zero anaphora", though possible in English, is much more heavily used in Chinese (pp. 101-102). Tiee and Lance (1986) observe that subject and object pronouns in Chinese, mainly employed for the purposes of clarity, emphasis and contrast, are used much less frequently than the corresponding pronouns in western languages (pp. 7-8). The fact that anaphoric reference is often not overtly marked by pronouns in Chinese may explain why Hong Kong students sometimes use English pronouns without clear antecedents: They assume that the readers can understand the meanings of the pronouns with the aid of the context and hence overlook the importance of establishing precise pronominal cohesive ties. The habit of not mentioning previously introduced noun phrases in Chinese may also help account for the overuse of third person personal pronouns in Hong Kong students' essays. Perceiving that there is no real communicative need to repeat given information in the texts, the students may prefer pronouns, which are function words with minimal meaning, to noun phrases, which involve content words, to fulfil the grammatical functions of subjects and objects in English. In sum, Hong Kong students' greater reliance on third person personal pronouns in the present study may reflect negative transfer from their mother tongue, in addition to their limited vocabulary knowledge.

5.2.2.1.3 Preference for an interactive style.

The final characteristic of Hong Kong students' written essays discussed in Section 5.2.2.1 is that Hong Kong students appear to adopt a more interactive tone than native speakers do. One can easily notice the dominance of pronoun categories in Table 29. The overuse of third person pronouns has been examined in Section 5.2.2.1.2. In this section, attention is directed to first person and second person pronouns. Second person pronouns (Feature 7) and first person pronouns (Feature 6) are the third most overused and the seventh most overused features in the essays written by Hong Kong students. In English, first person pronouns, e.g. *I* and *we*, refer to "the speaker(s)/writer(s) of the message" (Quirk et al., 1985, p. 339); second person pronouns, e.g. *you*, refer to "the addressee(s), but exclud[ing] the speaker(s)/writer(s)" (Quirk et al., 1985, p. 339). Both types of pronouns signal the language producers' involvement with the audience (Chafe, 1982, p. 46). The more frequent occurrence of these pronouns in the learner writing data indicates that Hong Kong students refer more often to themselves and their readers than native speakers do, which is a piece of strong evidence for their preference for a more interactive writing style.

Besides this primary evidence, two other signs can be found in Table 29 and Table 30. The feature of private verbs (Feature 56), i.e. verbs of intellectual states and acts in Biber (1988), is the tenth most overused feature in the learner writing data. Since these verbs refer to personal feelings or invisible cognitive processes, they go naturally with first person pronouns. Chafe (1982) classifies references to language producers' mental processes as a signal of involvement (pp. 46-47). In addition to the overuse of private verbs, another proof of Hong Kong students' interactive writing style is their underuse of agentless passive constructions. The

feature of agentless passives (Feature 17) is the fifth most underused feature in Table 30. Quirk et al. (1985) note that passive constructions are more commonly found in some objective, impersonal style of writing (p. 166). If a passive sentence conveys an impersonal flavour, a passive sentence without an agent can be said to represent an even more impersonal style as the agentive role is left out altogether. Therefore, the underuse of the impersonal agentless passives, together with the overuse of private verbs, reinforces the view that Hong Kong students favour a more involved approach when writing essays.

Interestingly, this higher involvement observed in Hong Kong students' essays does not seem to arise from the particular communicative situation. To recapitulate, the learner writing data in the present study contains some excerpts of essays written in an English writing examination, which is a high-stakes territory-wide formal assessment. Students sitting for the examination understood perfectly that the essays were to be read by nobody, but very few examiners, whom they could not possibly know. Consequently, there is no interpersonal function inherent in the real-life communicative situation. If there is no demand for involvement in the authentic situation, perhaps the more interactive nature of the learner writing data can be attributed to the communicative situation as specified in the task requirement. Among the five essay topics, three (university education, mobile phones, and Christmas celebrations in China) are strictly argumentative in nature, asking the students to argue for or against a general issue. One (world peace) requires the students to give explanations and suggestions; another one (living together before marriage) asks the students to analyse a phenomenon and share their views (see Appendix E). In other words, all the essay topics prompt the students to express personal opinions. In view of this, it is not surprising that first person pronouns and

private verbs occur frequently in the learner writing data. Nonetheless, the native speakers' essays used for comparison in the present study are argumentative in nature, too. In this sense, there is no difference between the task requirements for the two groups of writers. Consequently, there seems to be no justification, from the real and the constructed communication situations, for Hong Kong students' adoption of a more interactive style in the learner writing data.

Among the three overused features, second person pronouns, first person pronouns and private verbs, the high frequencies of the last two features can be more readily understood. First person pronouns and private verbs are common devices for expressing the writers' views and attitudes. Their overuse in the learner writing data can be explained by Hong Kong students' inclusion of more personal opinions in argumentative essays. What is more intriguing is the learners' overuse of second person pronouns (M = 12.9, SD = 19.2), which is six times more than the native speakers' use (M = 2.1, SD = 5.4). Under what circumstances do Hong Kong students use second person pronouns in examination essays? Further analysis of the learner writing data suggests two interesting usage patterns.

First, the second person pronouns in the learner writing data seem to distribute quite unevenly across the different topics (see Table 33). Many second person pronouns can be found in essays discussing university education and mobile phones; in contrast, very few second person pronouns are found in essays discussing world peace and Christmas in China. The uneven distribution in the learner writing data is more obvious, when Table 33 is contrasted with Table 34, which shows the distribution of second person pronouns in the native-speaker writing data. In Table 34, the average number of second person pronouns for each topic ranges from 0.1 to 1.6. There are three topics with an average of 0.1, three topics with an average of 0.5

or 0.6, and one topic with an average of 1.6. But in Table 33, the average number of second person pronouns ranges from 0 to 5.1. There are two topics with an average near 0, one topic with an average of 2.9, and two topics with an average near 5. It appears that the Hong Kong students use more second person pronouns in essays with topics more directly relevant to themselves and their daily lives (e.g. university education and mobile phones) but use fewer second person pronouns in essays with topics of less relevance (e.g. world peace and Christmas in China). Perhaps the amount of usage of second person pronouns is, to a certain extent, a topic-induced effect. It depends on the student writers' empathy with the subject matter. Nevertheless, this suggestion is merely tentative due to the small number of essays involved in each topic in the learner writing data.

Table 33

Raw Frequencies of Second Person Pronouns (N = 144) in Hong Kong Students' Essays of Different Topics

Essay Topic	No. of Essays	Total No. of 2nd Person Pronouns	No. of 2nd Person Pronouns Per Essay
World peace	10	0	0
University education	10	51	5.1
Mobile phones	12	59	4.9
Christmas in China	10	5	0.5
Living together before marriage	10	29	2.9
Total	52	144	

Table 34

Raw Frequencies of Second Person Pronouns (N = 26) in Native Speakers' Essays (LOCNESS Essays) of Different Topics

Essay Topic	No. of Essays	Total No. of 2nd Person Pronouns	No. of 2nd Person Pronouns Per Essay
Transport	8	4	0.5
Boxing	8	1	0.1
National Lottery	8	5	0.6
Computers and human brain	7	11	1.6
BSE* and beef	7	1	0.1
In vitro fertilization	9	1	0.1
Fox hunting	5	3	0.6
Total	52	26	

*BSE = bovine spongiform encephalopathy

The second observation on the use of second person pronouns in Hong Kong students' essays is that most of these pronouns are employed as generic reference. About 87% of the second person pronouns can be safely replaced by *we*, *they*, or *one*, other pronouns referring to people in general. For example:

> Looking outside the campus, you will probably see that many people on the street are talking with their mobile phone or it is not strange that you will hear more than once that there is a phone ringing when you are on the journey in a bus. (EXMCD:100-148)

> Even the one you live before marriage may not be the last one you be with in the rest of your life at the end, it would also be worthless. (EXMEH:101-132)

The fact that the students use a lot of second person pronouns for generic reference does not mean that they are unaware of the other options available for such reference. Among the above-mentioned 87% of the second person pronouns,

one-fifth are mixed with other generic pronouns, either within the same sentence or in the immediate neighbouring sentence:

> In fact, most of our skills come from our experience, so if you are willing to try, you would get a good prospect. (EXMBD:135-160)

> Also, you won't talk too much by using the telephone in the street as you need to pay $1 for every minutes! Thus, when people pick up the phone and dial it, they will just say what they want to say and what the main topic is, without saying any ridiculous things. (EXMCF:170-229)

In the first example, there is a change of generic pronouns from the first person plural *we* perspective to the second person perspective. In the second example, there is a transition from the *you* perspective to the third person plural *they* perspective. In both cases, it is in fact possible to maintain the same perspective and the resultant sentences will not differ much from the above sentences in meaning. The overuse of second person pronouns suggests that there seems to be some preference for the generic *you* among Hong Kong learners.

Since the generic use accounts for a large proportion of the second person pronouns in the learner writing data, its popularity among Hong Kong students requires further explanation. According to Quirk et al. (1985), when *we*, *you*, and *they* are used generically, they still "retain something of the specific meaning associated with the 1st, 2nd, and 3rd persons respectively" (p. 354). One distinctive feature of second person pronouns, as opposed to first person ones, is that they exclude reference to the writers. Perhaps the Hong Kong students use the generic *you* so as to distance themselves from some of the arguments in their essays. If this is the case, one may expect the generic *you* to co-occur frequently with ideas that the student writers disagree with:

> Even though you regard Christmas as a foreign custom, you still have no reason to prohibit the celebration of it think about the Buddhalism in China. (EXMDC:170-196)

Despite the existence of such usage, it constitutes only a small minority of the learner writing data. Furthermore, the generic *you* does appear in situations in which the students elaborate on their stances:

> The most important things for success are job experience, skills and new ideas. Practice makes perfect, if you have job experience, you would find that you can get a good job easily and your prospect will be better. Since you have job experience that the others do not have, you will have a comparative advantage. (EXMBD:41-108)

Therefore, it does not seem likely that the Hong Kong students use the generic *you* to differentiate opposite opinions from their own. Neither do they seem to use this pronoun to distance themselves from negative ideas, because, as evident from the following example, the generic *you* is used with positive and negative messages alike:

> In the past, if you cannot meet the basic requirements, eg. if you fail in English, you will not have a chance to study in university. But if you are good at sports, music, art etc, you still have a chance to succeed. (EXMBG:180-234)

Added to the above observations is the distribution of second person pronouns across different essay topics illustrated in Table 33. The tendency that more second person pronouns are found in essays discussing issues forming parts of the students' daily lives contradicts the hypothesis that the learners choose the second person pronouns as generic reference in order to distance themselves from the propositions.

Given the different pieces of evidence from the learner writing data, this hypothesis

appears fallible.

In fact, the finding that Hong Kong students like using second person pronouns is not unique to the present study. Milton's (2001) comparison of Hong Kong students' and British students' examination essays has also discovered that Hong Kong students tend to overuse *you* (p. 8). He attributes this to the learners' transfer of spoken English features to written English (p. 12). Milton's (2001) proposal can be one explanation to the exceptionally high frequency of second person pronouns in the present study. Quirk et al. (1985) point out that the generic *you* is the colloquial or informal equivalent of the more formal generic *one* (pp. 354, 387). Wales (1996) notes the popularity of the generic *you* in English speech, too (p. 45). Evidence supporting the spoken-language-transfer explanation for the overuse of second person pronouns can be found in the present study:

> Secondly, people also think that in university, student can meet many people and learn how to cooperate with others. It is good for them. Yes, that's right. But you haven't think that in university, you can only meet students and teachers. (EXMBF:54-110)

> I am sure that almost all of you have experiences that when you attend a meeting, seating on a bus, watching a movie in cinemas, and even listening to beautiful musics in concerts, the musics are always accompanied by those disguisting ringing sounds of mobile phones. (EXMCA:67-117)

> Suddenly, "Where does a song come from?" Sally asked. I don't know but I just know that Sally and I were disturbed by those people who have mobile phones. Actually, if you want to see film, why don't you turn off your phone or turn it in a vibration mode before come in a cinema? (EXMCJ:3-75)

In the learner writing data, around 20% of the total number of second person pronouns co-occur with typical features of spoken discourse including interactional responses (e.g. *Yes, that's right* in the first example), phrases used for addressing audience who is physically present (e.g. *all of you* in the second example) and direct questions (e.g. the last example). Reading the above excerpts, one cannot help feeling that the students are talking directly to some imaginary audience. It appears that transfer of spoken language use can be a good explanation for the overuse of second person pronouns in Hong Kong students' English writing. This negative transfer may be caused by the teaching practice in Hong Kong. Both debates and argumentative essays are parts of the senior secondary school English curriculum (Curriculum Development Council & Hong Kong Examinations and Assessment Authority, 2007, p. 23). Debates are one kind of popular classroom activities accompanying the teaching of argumentative writing in Hong Kong schools and debate speeches are also one specific type of argumentative writing in some English textbooks. The fact that argumentative speech and argumentative writing often accompany each other in the English classrooms may lead to the students' mixing features of the two genres.

Although there is some support for the explanation of spoken-language transfer, this cannot be the complete picture. For one thing, as mentioned in the last paragraph, only 20% of the second person pronouns are explicitly related to spoken language use. For another, far more second person pronouns have been found in Hong Kong students' writing than in their speech in the present study. The mean normalised frequency of second person pronouns in the learner writing data is 21.9 (SD = 19.2), but the mean normalised frequency of second person pronouns in the learner speech data is only 8.4 (SD = 10.5). If spoken-language transfer is an adequate explanation,

one will expect the Hong Kong students to use second person pronouns at least as frequently, if not more frequently, in speech as in writing. But it turns out that this is not the case. Consequently, spoken-language transfer cannot be the sole cause of Hong Kong students' overuse of second person pronouns in English essays.

To better explain the phenomenon, two other possible factors affecting Hong Kong students' use of second person pronouns are considered. The first one is associated with the teaching of English writing in Hong Kong. Wong's (1992) analysis on a popular series of Hong Kong secondary school English textbooks reveals that argumentative writing is not introduced until the students are promoted to Form 4 (pp. 128-129, 148). The approach adopted by the textbooks in the teaching of argumentative writing involves heavy emphasis on content and minimal advice on writing style (pp. 138, 148). What Wong (1992) observes may have a serious impact on Hong Kong students' writing. Adequate exposure to a genre can help students get a sense of what the genre should be like and explicit instruction can raise their awareness towards the genre-specific norms. Both of these improve second language learners' sensitivity and writing skills so that they can produce English writing fulfilling native speakers' expectation. Without adequate exposure and explicit instruction, the students may not be conscious of the differences existing among different genres. Second person pronouns, as well as first person pronouns, are actually quite common in the other written genres that Hong Kong students have learnt and practised since an early stage, e.g. personal letters and short stories. It is possible that the students, ignorant of the relatively impersonal tone in argumentative writing, retain the interactive style that they normally use for other written assignments when they write argumentative essays. As a result, the delayed introduction of argumentative writing and the lack of focus on writing style in school

textbooks may account for Hong Kong students' use of a more involved style in the learner writing data.

Although Wong (1992) does not contain any analysis on learner language data, she does speculate how the inadequacy of the English writing instruction may influence Hong Kong students' writing: Hong Kong students "will transfer the rhetorical skills they acquire in their Chinese language curriculum to help them with their needs for writing argumentative essays in English" (p. 149). In addition to English textbooks, Wong (1992) has also surveyed a series of Chinese textbooks. She has found that Chinese argumentative writing is introduced right at the beginning of the secondary school education, i.e. when the students are in Form 1 (pp. 159, 167-168, 241). What is more is that the Chinese textbooks adopt "a rhetorical/textual approach" (p. 244), which means that argumentative techniques are explicitly taught. To Wong (1992), the more advanced knowledge in Chinese argumentative writing enables the students to transfer some of the strategies used in Chinese argumentative essays to English essays in order to compensate for the inadequate training in English argumentative writing. Her comparison of 10 argumentative essays from the Chinese textbooks and 10 argumentative essays from the English textbooks also shows that there are more first person and second person pronouns in the Chinese essays than in the English essays, in terms of both the total number used and the number used per topic chain (pp. 332, 400, 432, 435, 469). However, the length of the Chinese essays is much longer than the length of the English ones and the topic chains are of varying lengths, too. One cannot readily assume, simply based on Wong's (1992) findings, that the overuse of first person and second person pronouns found in the learner writing data of the present study is caused by negative transfer from Chinese. As suggested by Chao's (2003) study on

Taiwanese students' Chinese and English writing, the relationship between one's L1 and L2 language systems is rather complex and "Chinese rhetorical influence is not necessarily strong enough to cause interference" (p. 130).

Whereas the first factor that may affect Hong Kong students' use of second person pronouns is associated with the specific educational context in Hong Kong, the second factor proposed concerns the acquisition of generic pronouns by second language learners in general because the overuse of second person pronouns is not unique to Hong Kong English language learners. In a study of writer/reader visibility, Petch-Tyson (1998) has discovered that Dutch, Finnish, French and Swedish learners of English all include more second person pronouns, though in varying degree, in argumentative writing than native speakers do (p. 112). She suggests that different cultural backgrounds, different persuasive strategies and different perceptions of the essay topics may be some reasons behind the higher writer/reader visibility in the learner essays (p. 117). Hinkel's (2002) comparison of expository and argumentative essays from second language learners and native speakers has also found that Chinese, Japanese, Korean, Vietnamese and Arabic students employ more second person pronouns than native speakers do (p. 86). The researcher comments that the Confucian culture in China, Japan and Korea and the Arab culture may contribute to the students' overuse of second person pronouns (p. 87). Cobb (2003) and van Rooy (2008) demonstrate respectively that Canadian French learners of English and South African learners of English overuse second person pronouns. Considering the findings of the above four studies and the present study, one can presume that the overuse of second person pronouns is a prevalent phenomenon among second language learners of English. It seems likely that there are some reasons, other than culture-specific ones, behind the general trend of second language learners' overuse

of second person pronouns.

The second factor that is proposed to explain the overuse of second person pronouns among Hong Kong students, as well as other second language learners, is the neglect of the generic use of personal pronouns in English language teaching. For one thing, personal pronouns are usually introduced at the early stage of English language acquisition, during which learners focus more on the here and now. When teachers first introduce these pronouns, specific reference, which is more relevant to the immediate context, always takes priority over generic reference. For another, grammarians tend to oversimplify the complexity involved in the different usage of English pronouns (Wales, 1996, pp. xi, 50). In particular, the generic plural personal pronouns, including *you*, "although extremely common, are often but briefly described by grammarians, if they are mentioned at all" (Wales, 1996, p. 59). If the generic use of personal pronouns is not given due recognition by linguists, it is no surprise that they are neglected by language teachers. Berry (2009) observes that the generic function of *you* is "frequently forgotten in pedagogical grammars of English" (p. 29). His survey of different pedagogical materials reveals that although generic *you* is mentioned in scientific grammars, it is not touched upon at all in grammar books for teachers (p. 33). Neither is it addressed in EFL classroom textbooks and grammar exercise books (p. 33). If students are not explicitly taught the generic use of personal pronouns, they are left to acquire it by themselves. But since many English language learners only have limited exposure to the English language, they may not fully understand the stylistic constraints associated with the generic use of personal pronouns. For example, the generic *you* is quite common in pedagogical materials (Berry, 2005, p. 88). If there is no explicit warning against its use in writing, second language learners may consider it perfectly natural in written

English.

Hong Kong students' preference for an interactive style in written essays manifests itself in the overuse of first person pronouns, second person pronouns and private verbs and the underuse of agentless passives. Various factors that may contribute to this higher involvement in the learner writing data have been discussed. Possible reasons behind Hong Kong students' usage include a greater emphasis on personal opinions in argumentative writing, an insensitivity towards the distribution of second person pronouns in speech and writing, a delayed introduction of argumentative writing in the English curriculum, a lack of focus on different language styles in the teaching of writing, and a negligence of the generic use of personal pronouns by the language experts and the teaching practitioners.

Ending Section 5.2.2.1, this last paragraph provides a summary for the discussion section of the characteristics of Hong Kong students' written English in expository or argumentative essays. This section, Section 5.2.2.1, has addressed Research Question 2.1 (How are Hong Kong students' written essays different from comparable native-speaker written production, in terms of the use of syntactic features?) by discussing three distinctive usage patterns in Hong Kong students' written essays: the dominance of present time frame, the greater reliance on pronouns for nominal functions, and the preference for an interactive style. When compared with the native speakers, Hong Kong students tend to use more present tense verbs, but fewer past tense verbs and perfective verbs. Although the nonmarking of past tense is often regarded as a feature of Hong Kong English, the low occurrence of past tense in the present study does not originate from errors. The underuse of time adverbials suggests that the learners avoid using other temporal

perspectives by describing events without specific reference of time. Analysis on the writing strategies adopted by the learners reinforces the belief that they favour the habitual present perspective. The Hong Kong students have also been found to use more third person personal pronouns, but fewer nouns, prepositional phrases and nominalizations. The learners' higher tendency to substitute noun phrases with third person pronouns hints that they have a smaller vocabulary size than the native speakers and that they rely more on grammatical cohesion than the native speakers do. The fact that one-third of the third person personal pronouns used in the learner writing data exhibit cohesive problems shows that the students' mother tongue may have an adverse effect on their use of these pronouns in English. In addition to third person pronouns, the Hong Kong students tend to overuse other personal pronouns. When compared with native speakers, the learners use more first person pronouns, second person pronouns and private verbs, but fewer agentless passive constructions. Through the analysis of the learners' use of second person pronouns, it is proposed that the more interactive tone in their writing is shaped by the students' insensitivity towards mode and genre differences, the local practice in the teaching of writing, and the little attention paid to pronouns in English language teaching.

5.2.2.2 Relationship between the deviated features and Biber's (1988) model.

The previous section, Section 5.2.2.1, has discussed the top 10 most overused and the top 10 most underused syntactic features in the learner writing data by identifying underlying relationships among the features on the basis of textual and communicative functions. In this section, these 20 cases of most heavily deviated usage are examined in relation to Biber's (1988) model on spoken and written

language differences. More specifically, they are categorised according to how they are represented on Biber's (1988) model in order to determine if they are indicative of any spoken-like language use in Hong Kong students' written essays. The findings will provide an answer for Research Question 2.2: Do the differences in the use of syntactic features observed suggest that Hong Kong students' written essays exhibit spoken-like characteristics?

To begin with, the top 10 most overused and the top 10 most underused features in Table 29 and Table 30 are reproduced in Table 35 and Table 36. Newly added to these two tables is information about how the syntactic features appear in Biber's (1988) model. In the second column, the numbers indicate the textual dimensions that the features belong to in the model. The numbers, 2, 4 and 6, are put in brackets because Dimension 2, Dimension 4 and Dimension 6 are not directly relevant to the distinctions between speech and writing. The mark N/A means that a feature does not belong to any of the six dimensions. In spite of the fact that Biber (1988) has counted 67 syntactic features, not every one of them has a role to play in his model. Nine features (e.g. gerunds) are not loaded on any of the six textual dimensions. In the third column, features belonging to Dimension 1, Dimension 3 and Dimension 5 are further classified in terms of the sides that the features lie in. Both ends of each dimension are specified and the underlining represents where a feature lies.

In Table 35, the top most overused syntactic feature, present tense, is a feature of Dimension 1 in Biber's (1988) model and its presence makes a text more involved, rather than informational. Therefore, the overuse of present tense hints that the Hong Kong students' essays are, to a certain extent, more involved than the native speakers' essays. Similarly, five other features in this table, second person pronouns, possibility modals, analytic negation, first person pronouns and private verbs, lie in

the involved side of Dimension 1. The overuse of these features all suggests that the learners' essays are more involved in some aspects. Regarding Dimension 3, the feature of total adverbs is loaded on the situation-dependent side. The overuse of this feature implies that the student writers employ more situation-dependent reference. According to Biber (1988), typical speech can be characterised by an involved focus, situation-dependent reference and non-abstract content (p. 162). Table 35 has demonstrated that among the top 10 most overused syntactic features, eight features are relevant to the three textual dimensions distinguishing speech from writing and the overuse of seven features indicates that the Hong Kong students' essays, being more involved and more reliant on situation-dependent reference, should be more spoken-like than the native speakers' essays.

Table 35

The Top 10 Most Overused Syntactic Features in Hong Kong Students' Writing Data, Classified According to Biber's (1988) Three Dimensions (1, 3 & 5)

Syntactic Feature		Dimension	Which Side the Feature Lies in the Dimension
3.	Present Tense	1	involved vs informational
8.	Third Person Pronouns	(2)	N/A
7.	Second Person Pronouns	1	involved vs informational
52.	Possibility Modals	1	involved vs informational
67.	Analytic Negation	1	involved vs informational
42.	Total Adverbs	3	explicit vs situation-dependent
6.	First Person Pronouns	1	involved vs informational
45.	Conjuncts	5	abstract vs non-abstract
15.	Gerunds	N/A	N/A
56.	Private Verbs	1	involved vs informational

Table 36

The Top 10 Most Underused Syntactic Features in Hong Kong Students' Writing Data, Classified According to Biber's (1988) Three Dimensions (1, 3 & 5)

Syntactic Feature		Dimension	Which Side the Feature Lies in the Dimension
16.	Total Other Nouns	1	involved vs informational
39.	Total Prepositional Phrases	1	involved vs informational
1.	Past Tense	(2)	N/A
2.	Perfect Aspect	(2)	N/A
17.	Agentless Passives	5	abstract vs non-abstract
43.	Type/Token Ratio	1	involved vs informational
14.	Nominalizations	3	explicit vs situation-dependent
10.	Demonstrative Pronouns	1	involved vs informational
63.	Split Auxiliaries	(4)	N/A
5.	Time Adverbials	3	explicit vs situation-dependent

The way to understand Table 36 is slightly different from the way to understand Table 35. For example, it can be seen from Table 36 that the top most underused syntactic feature, total other nouns, is a feature of Dimension 1 and its presence makes a text more informational. However, since Table 36 lists features that are underused by the Hong Kong students, the fact that the feature of total other nouns appears in Table 36 means that there are fewer nouns in the learners' essays, which in turn shows that the texts are less informational. In other words, the underuse of total other nouns represents the more involved nature of the learner writing data, as the involved quality and the informational quality of a text are presented as two ends of a continuum in Biber's (1988) model. In a similar vein, there are two other features in Table 36, total prepositional phrases and type/token ratio, lying in the informational side of Dimension 1. The underuse of these features again indicates

that the learners' essays are more involved in some ways. With reference to Dimension 3, one feature, nominalizations, is loaded on the explicit side. Its underuse hints that the students use less explicit reference and that there is more situation-dependent reference in the learners' essays. Regarding Dimension 5, the feature of agentless passives appears on the abstract side in Biber's (1988) model. Its underuse implies that the learners' essays are relatively non-abstract. Overall, Table 36 has illustrated that, seven out of the top 10 most underused syntactic features are related to Dimension 1, Dimension 3 and Dimension 5 in Biber's (1988) model. Moreover, the underuse of five features suggests that the Hong Kong students' written essays, having less informational concern, less explicit reference and less abstract content, should be more spoken-like than the native speakers' essays.

Taking both Table 35 and Table 36 into consideration, one can notice that, among the 20 cases of most deviated usage, 15 are connected to the three textual dimensions that can explain the variation across speech and writing. It seems reasonable to claim that sensitivity to mode differences is highly related to differences between Hong Kong students' and native speakers' writing. The mode differences are creating trouble for Hong Kong students. The students may not be able to use syntactic features in a way that is appropriate for the written mode and that meets native speakers' expectation. Furthermore, 12 cases of the deviated usage are indicative of typical characteristics of English speech, i.e. an involved focus, context-dependent reference and concrete content. These 12 cases are:

Overuse

- Present Tense
- Second Person Pronouns
- Possibility Modals
- Analytic Negation
- Total Adverbs
- First Person Pronouns
- Private Verbs

Underuse

- Total Other Nouns
- Total Prepositional Phrases
- Agentless Passives
- Type/Token Ratio
- Nominalizations

In response to Research Question 2.2, the analysis in this section does suggest that Hong Kong students' written essays exhibit some spoken-like characteristics.

In Section 5.2, Hong Kong students' written essays have been compared with similar native-speaker written production, written essays, on the level of syntactic features. Results of all 67 syntactic features have been presented. The discussion has focused on the top 10 most overused features and the top 10 most underused features. It has been found that Hong Kong students' essays differ from native speakers' ones in terms of the temporal perspective, the pronominal anaphora, and the writer/reader involvement. These differences can be traced to Hong Kong students' knowledge and understanding of the English language, their writing strategies, the language system of their mother tongue, and the educational practice in Hong Kong and worldwide. It has also been found that more than half of the syntactic features representing the most serious differential usage reflect a confusion over spoken and written language use. It is highly likely, from the analysis of syntactic features in Section 5.2, that Hong Kong students' written essays are more spoken-like than native speakers' essays.

5.3 Comparison on Mean Dimension Scores

Following a detailed analysis of the use of syntactic features in Hong Kong students' written essays, this section moves on to report findings on textual dimensions in the learner writing data. Hong Kong students' essays are compared and contrasted with native-speaker production in terms of the three textual dimensions that are found to distinguish between speech and writing in Biber's (1988) model. In Section 5.3.1, statistics on the dimension scores of both the learner writing data and the similar native-speaker writing data are first presented. In Section 5.3.2, these statistical results are then discussed with reference to Research Questions 2.3 and 2.4. A comparison is drawn to investigate how Hong Kong students' written essays deviate from the native speakers' ones. Another comparison is also made to determine if Hong Kong students' essays bear any resemblance to common native-speaker spoken genres. Section 5.3 should help the readers further understand whether Hong Kong students' written English exhibits characteristics typically found in English speech.

5.3.1 Results

This section reports the descriptive statistics concerning the dimension scores of the learner writing data and the corresponding native-speaker writing data (LOCNESS essays). Included in this Results section are (a) mean scores and standard deviations of all six textual dimensions of Hong Kong students' written essays and (b) mean scores and standard deviations of all six textual dimensions of native speakers' written essays. Detailed procedures for the computation of dimension scores can be found in Chapter 3 Section 3.5.3.2.

5.3.1.1 Dimension scores of Hong Kong learners' written essays.

Table 37 presents the mean dimension scores and the standard deviations of the six textual dimensions calculated from the learner writing data. The mean scores of all six dimensions identified in Biber's (1988) are included here for the readers' reference. But as mentioned in Chapter 3 Section 3.2.2, only Dimension 1, Dimension 3 and Dimension 5 are relevant to the distinction between speech and writing. Subsequent discussion on the findings will hence focus on these three dimensions. The mean score for Dimension 1 Involved versus Informational Production is 2.1. The mean score for Dimension 3 Explicit versus Situation-dependent Reference is 1.4. The mean score for Dimension 5 Abstract versus Non-abstract Information is 4.7. What is worth noticing is the unusual, large standard deviation figure of Dimension 1. It seems that the learners' essays are more varied along this particular textual dimension. This impression is confirmed by an examination of the statistics: The Dimension 1 scores of the 52 essays range from -19.8 to 29.0. The range of the Dimension 1 scores (48.9) is more than twice the range of the Dimension 3 scores (20.2) or the range of the Dimension 5 scores (21.4).

Table 37

Mean Dimension Scores and Standard Deviations of Hong Kong Students' Writing Data

Dimension	Mean Score	Standard Deviation
Dimension 1	2.1	11.4
Dimension 2	-3.7	2.6
Dimension 3	1.4	4.4
Dimension 4	3.7	5.1
Dimension 5	4.7	5.4
Dimension 6	-0.4	3.1

5.3.1.2 Dimension scores of native speakers' written essays.

Table 38 shows the mean dimension scores and the standard deviations of the argumentative essays written by British native speakers. Again scores of all six textual dimensions are presented, but our discussion will concentrate on Dimension 1, Dimension 3 and Dimension 5 only. The mean score for Dimension 1 Involved versus Informational Production is -5.6. The mean score for Dimension 3 Explicit versus Situation-dependent Reference is 2.2. The mean score for Dimension 5 Abstract versus Non-abstract Information is 3.6. Similar to the case in Table 37, the standard deviation of Dimension 1 in Table 38 is much larger than the rest of the standard deviations in the same table, suggesting that the native speakers' essays are also more varied along this dimension than along the other dimensions. Nonetheless, the standard deviation of Dimension 1 in Table 38 is not as large as the standard deviation of Dimension 1 in Table 37, implying that, relatively speaking, the native speakers' essays are less varied than Hong Kong students' essays on Dimension 1. Other standard deviations in Table 38 are similar in magnitude to their counterparts in Table 37.

Table 38

Mean Dimension Scores and Standard Deviations of Native Speakers' Writing Data (LOCNESS Essays)

Dimension	Mean Score	Standard Deviation
Dimension 1	-5.6	9.5
Dimension 2	-2.1	2.6
Dimension 3	2.2	4.3
Dimension 4	4.0	5.4
Dimension 5	3.6	5.9
Dimension 6	-1.2	3.0

5.3.2 Discussion

After the provision of statistical information in Section 5.3.1, this section studies Hong Kong students' written essays by placing the genre within Biber's (1988) theoretical framework. The following discussion tries to characterise these essays by considering their relationships with other native-speaker genres along the three textual dimensions most relevant to mode differences. To achieve this purpose, Figure 16, Figure 17 and Figure 18 have been constructed. Illustrating graphically the mean dimension scores reported in Table 37 and Table 38, these figures show clearly where the learner writing data and the comparable native-speaker writing data are situated along Dimension 1, Dimension 3 and Dimension 5 in Biber's (1988) model. In each figure, the vertical axis represents the dimension scores. On the right, Biber's (1988) model is reproduced. All 23 genres investigated by Biber (1988) are plotted vertically according to their respective mean dimension scores. Genres with very similar mean scores may be placed on the same horizontal level so as to avoid overlaps of different labels. In the middle part of each figure, the two genres investigated in the present study, the learners' essays (labelled as "HK Essays") and

the native speakers' essays (labelled as "British Essays"), are also positioned according to their mean scores. The decision to put them in the middle, rather than on the right, is mainly based on clarity of information presentation. On the left, the two vertical dotted lines represent the range of scores of all six spoken genres in Biber's (1988) model and the range of scores of all 17 written genres in the same model respectively.

In the following two sub-sections, Figure 16, Figure 17 and Figure 18 are explored in detail. The discussion in Section 5.3.2.1 is intended to answer Research Question 2.3 (How are Hong Kong students' written essays different from comparable native-speaker written production, in terms of the three textual dimensions identified in Biber's (1988) model?). The relative positions of Hong Kong students' essays and native speakers' essays in Figure 16, Figure 17 and Figure 18 are considered to find out how Hong Kong students' writing deviates from native speakers' expectation. Section 5.3.2.2 addresses Research Question 2.4 (Is there any similarity between Hong Kong students' written essays and common native-speaker spoken genres, in terms of the three textual dimensions identified in Biber's (1988) model?). Special attention is paid to those genres in Biber (1988) which are near the learner writing data in Figure 16, Figure 17 and Figure 18. Through the discussion in Section 5.3.2.2, it can be decided whether Hong Kong students' written English resembles real-life spoken language use of native speakers.

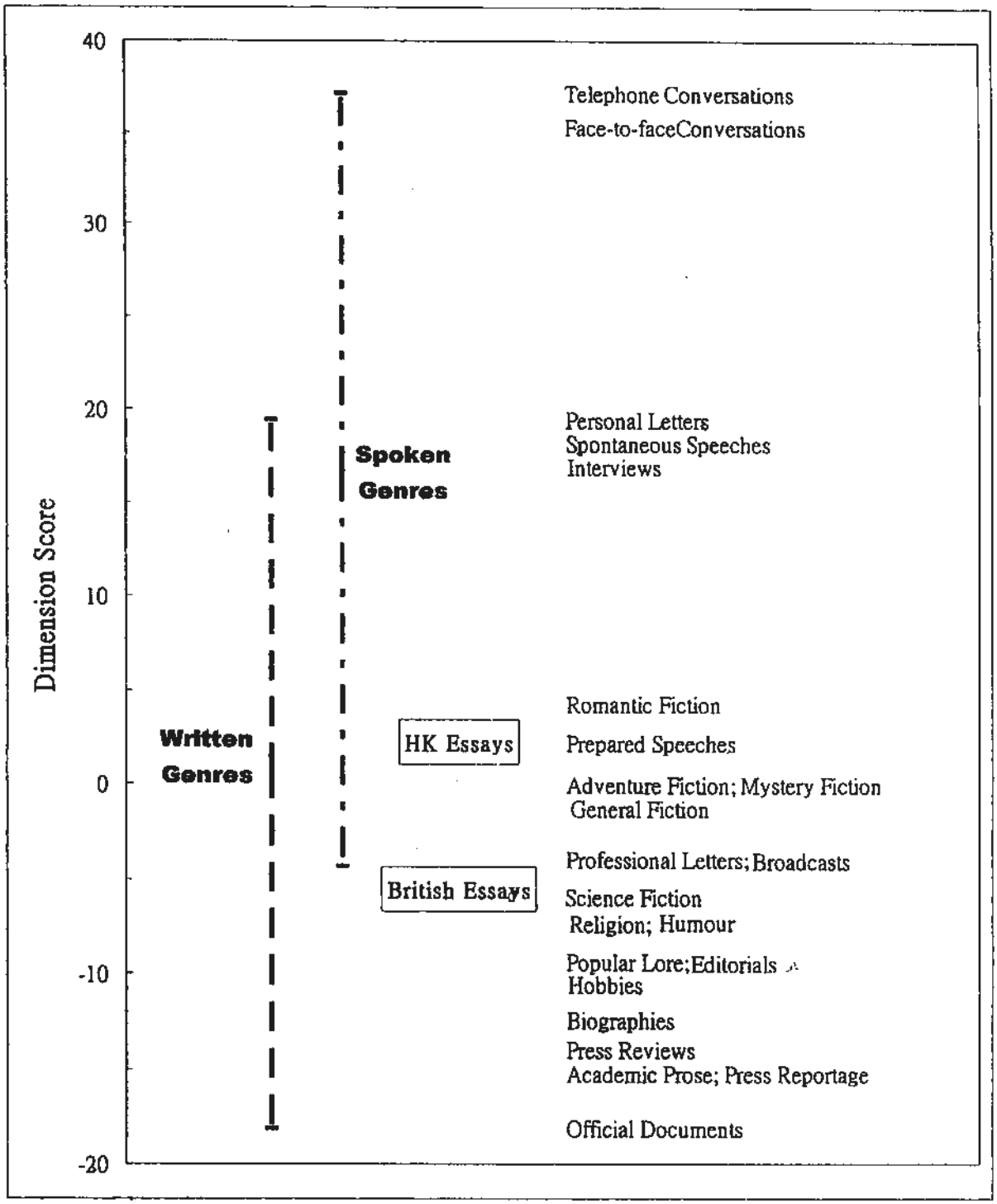

Figure 16. Distribution of the mean scores of the learner writing data, the native-speaker writing data (LOCNESS Essays) and Biber's (1988) different genres (p. 128) along Dimension 1 Involved versus Informational Production

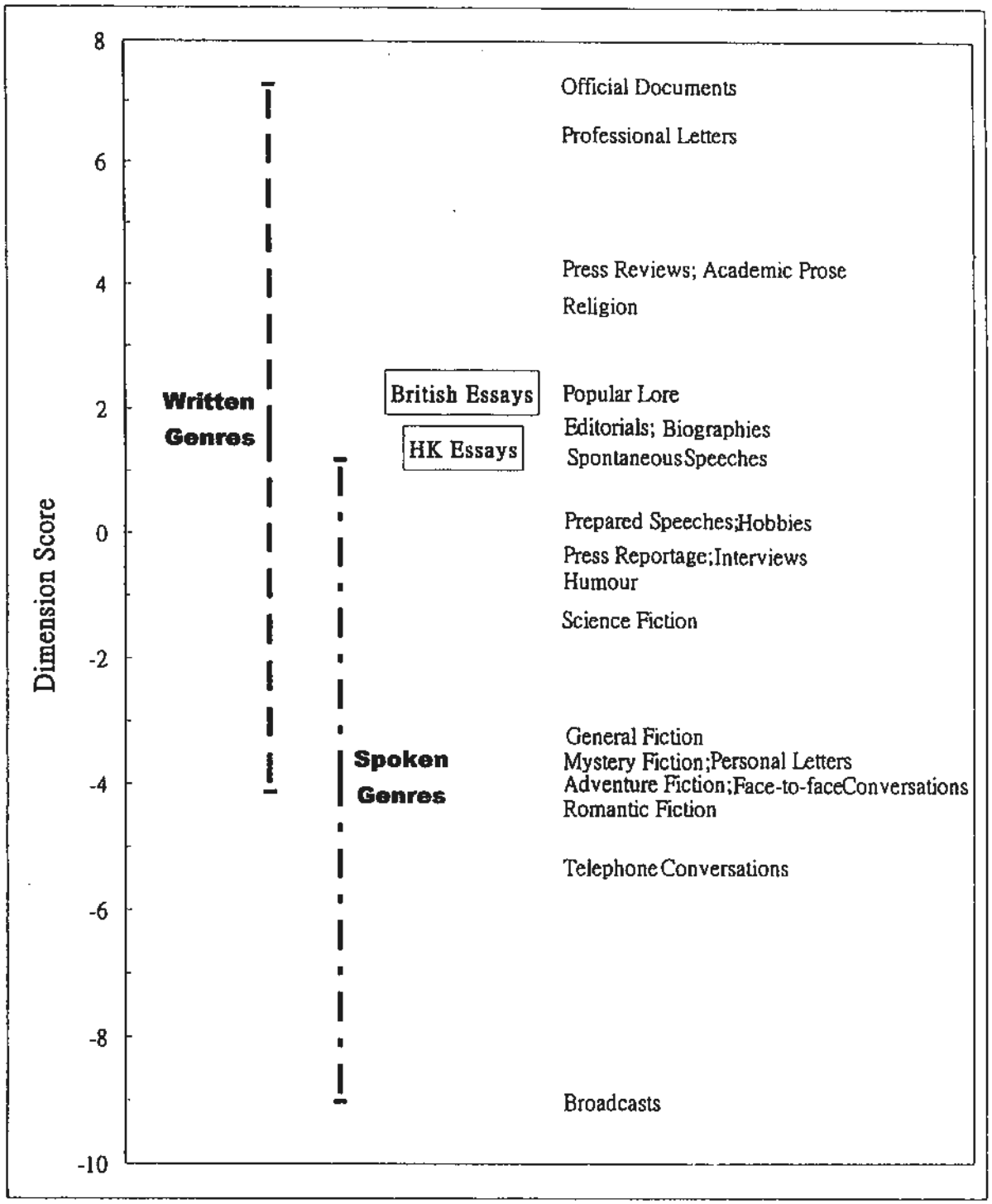

Figure 17. Distribution of the mean scores of the learner writing data, the native-speaker writing data (LOCNESS Essays) and Biber's (1988) different genres (p. 143) along Dimension 3 Explicit versus Situation-dependent Reference

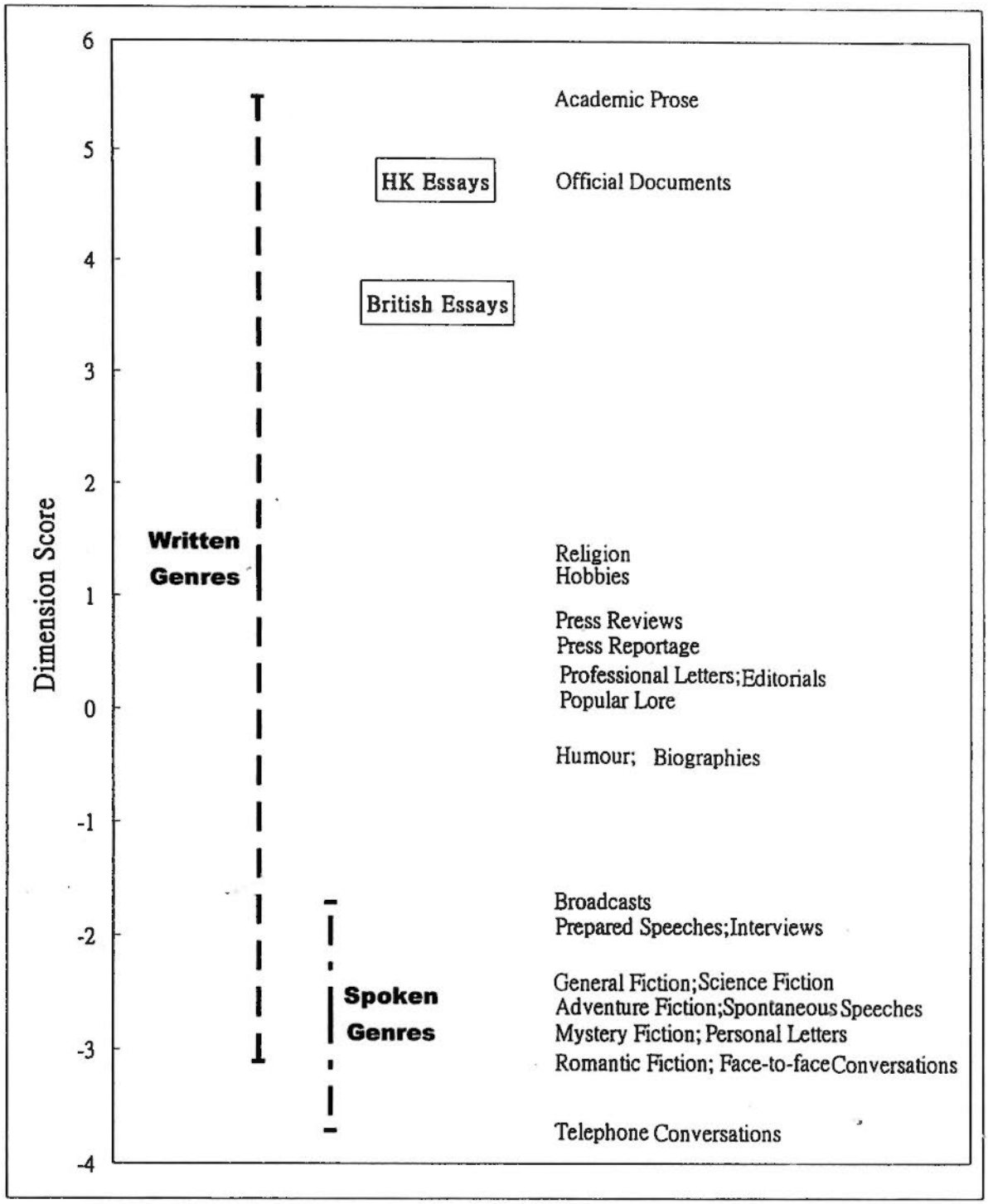

Figure 18. Distribution of the mean scores of the learner writing data, the native-speaker writing data (LOCNESS Essays) and Biber's (1988) different genres (p. 152) along Dimension 5 Abstract versus Non-abstract Information

5.3.2.1 Comparison between Hong Kong learners' written essays and native speakers' written essays.

How are Hong Kong students' written essays different from comparable native-speaker written production (written essays), in terms of the three textual dimensions identified in Biber's (1988) model? This section is going to provide an answer for this research question by examining the left and the middle parts of Figure 16, Figure 17 and Figure 18. In Figure 16, Dimension 1 Involved versus Informational Production is represented. Hong Kong students' essays have a larger mean score than the native speakers' essays. In Biber's (1988) model, the larger the Dimension 1 score, the more involved the discourse is; the smaller the score, the more informational the discourse is (p. 135). In other words, the larger score of the learner writing data shows that Hong Kong students' essays are more involved than the native speakers' ones. As illustrated by the two vertical dotted lines, there is a tendency for the native-speaker spoken genres to have higher Dimension 1 scores than the native-speaker written genres. Given this observation, it can be claimed that the score of the learners' essays deviates from that of the native speakers' essays by moving towards the end characterising native-speaker spoken genres. Moreover, whereas the score of the native speakers' essays falls outside the range of scores of native-speaker spoken genres, the score of the learners' essays falls within the range. Consequently, Figure 16 clearly demonstrates that Hong Kong students' essays are more spoken-like than the comparable native-speaker written production along Dimension 1.

As shown in Figure 17, Hong Kong students' essays have a smaller Dimension 3 score than the native speakers' essays. On Dimension 3 Explicit versus Situation-dependent Reference, a higher mean score indicates heavier reliance on

explicit reference and a lower mean score represents greater dependence on contextual reference (Biber, 1988, p. 142). Figure 17 thus suggests that the Hong Kong students rely more on situation-dependent reference. Since it is usually the native-speaker spoken genres which exhibit lower scores along this textual dimension, the score of the Hong Kong essays differs from that of the British essays by moving towards the end characterising native-speaker spoken genres. At the same time, the score of the learner writing data is also approaching the upper limit of the scores of the native-speaker spoken genres. As a result, when compared with the native speakers' essays, the learners' essays can be considered more spoken-like along Dimension 3. However, it should also be pointed out that the difference between the two sets of essays is quite small. The more spoken-like nature of Hong Kong students' essays on Dimension 3 is not as obvious as that observed in Dimension 1.

In Figure 18, Hong Kong students' essays are found to have a larger mean score than the native-speakers' essays along Dimension 5 Abstract versus Non-abstract Information. According to Biber (1988), this textual dimension distinguishes genres by their varying degree of abstract content. The higher the score, the more abstract information the genre contains (p. 151). The higher Dimension 5 score of the learner writing data reveals that Hong Kong students' essays are more abstract than the native speakers' essays. It can be observed from the diagram that common native-speaker spoken genres tend to have very low scores along this textual dimension. Based on this information, one can conclude that Hong Kong students' written essays are not more spoken-like than the native speakers' essays along Dimension 5.

In summary, when contrasted with comparable native-speaker written production by using Biber's (1988) theoretical framework, Hong Kong students' written essays are found to (a) adopt a more involved style, (b) use more situation-dependent reference and (c) contain more abstract information. An involved focus and a reliance on situation-dependent reference are typical characteristics of spoken discourse (Biber, 1988, p. 162). The analysis in this section demonstrates that Hong Kong students' essays are more spoken-like than native speakers' essays along Dimension 1 Involved versus Informational Production and Dimension 3 Explicit versus Situation-dependent Reference, but not Dimension 5 Abstract versus Non-abstract Information.

5.3.2.2 Comparison between Hong Kong learners' written essays and common native-speaker spoken genres.

The discussion in Section 5.3.2.1 has already shown that the learner writing data in the present study differ from similar native-speaker written production by being more spoken-like on two out of three relevant textual dimensions in Biber's (1988) model. This section further explores the spoken-like nature of Hong Kong students' written English by addressing Research Question 2.4: Is there any similarity between Hong Kong students' written essays and common native-speaker spoken genres, in terms of the three textual dimensions identified in Biber's (1988) model? The following discussion will draw on Figure 16, Figure 17 and Figure 18 again, but this time the focus is to identify the native-speaker genres which behave similarly to the learners' essays, so mainly the right hand side of the graphs will be considered.

On Dimension 1 Involved versus Informational Production, the label for Hong Kong students' essays is located in close proximity to romantic fiction, prepared speeches, adventure fiction, mystery fiction and general fiction (see Figure 16). The mean score of the learners' essays (2.1) is nearly identical to the mean score of prepared speeches (2.2), suggesting that Hong Kong students use as much involved language in their written essays as the native speakers do in prepared speeches. Other genres with similar mean scores to the learners' essays are all fiction: romantic fiction (4.3), adventure fiction (-0.0), mystery fiction (-0.2) and general fiction (-0.8). Despite its written nature, fiction is closely connected to speech because of the considerable amount of dialogue it contains. Tannen (1985) has commented that "imaginative literature has more in common with spontaneous conversation than with the typical written genre, expository prose" (p. 137). In view of this, Hong Kong students' essays can be considered quite similar to genres with a lot of written speech. Given the evidence in Figure 16, it seems reasonable to conclude that, in terms of the use of involved or informational language, Hong Kong students' essays closely resemble native speakers' spoken language produced under situations in which careful preparation is allowed and interpersonal communication is limited.

Regarding Dimension 3 Explicit versus Situation-dependent Reference, the mean score of Hong Kong students' essays (1.4) is very close to the mean score of spontaneous speeches (1.2). In Figure 17, the former is plotted right next to the latter. The similarity between the two scores indicates that the Hong Kong students employ as much context-dependent reference in their written essays as the native speakers do in spontaneous speeches, representing the spoken-like nature of the learner writing data. The two written genres with similar mean scores, editorials (1.9) and biographies (1.7), often rely on situation-dependent reference, too. Editorials usually

cover current social issues that are familiar to the general public. It is not difficult to understand that a fair amount of exophoric reference has to be used. Biographies, by definition, are people's life stories. Physical and temporal contexts are created as stories proceed. Explicit reference is not deemed necessary. Therefore, the position of Hong Kong students' essays in Figure 17 suggests that, in terms of the use of explicit and situation-dependent reference, Hong Kong students rely quite a lot on the second type of reference in written essays and their language use is very similar to native speakers' use in spontaneous speeches.

On Dimension 5 Abstract versus Non-abstract Information, there is no native-speaker spoken genre with a mean score similar to that of Hong Kong students' essays (see Figure 18). In the figure, the genre of Hong Kong students' essays is situated near official documents and academic prose. In other words, the Hong Kong students include as much abstract content in their written essays as the native speakers do in formal writing. As shown in Figure 18, this high level of abstract focus does not exist in native speakers' spoken language use.

Considering the above evidence from Figure 16, Figure 17 and Figure 18, Hong Kong students' written essays resemble native-speaker spoken language use on Dimension 1 Involved versus Informational Production and Dimension 3 Explicit versus Situation-dependent Reference: They are similar to prepared speeches in terms of the use of involved/informational language and they are also similar to spontaneous speeches in terms of the use of explicit/situation-dependent reference. The discussion in Section 5.3.2.1 has already shown that the language used in Hong Kong students' essays deviates from the language used in comparable native-speaker written production by shifting towards a direction of typical spoken language use on Dimension 1 and Dimension 3. This section, Section 5.3.2.2, adds further the

observation that such deviation causes the learners' essays to become highly similar to some real-life native-speaker spoken genres in certain aspects. As a result, the findings in these two sections show that Hong Kong students' written English exhibits an unusually high proportion of features which are typically expected to be found in native speakers' spoken English.

However, the above conclusion does not necessarily mean that the English that the Hong Kong students use in their essays is totally inappropriate for the written mode. As illustrated in Figure 16 and Figure 17, there is considerable overlap between the scores of the native-speaker written genres and those of the native-speaker spoken genres. In fact, along both Dimension 1 and Dimension 3, the mean score of the learners' essays still lies within the range of scores for written genres. Basically the language that the Hong Kong students use should still be considered acceptable for the written mode. What is problematic is that whereas the mean score of the native speakers' essays falls in a range that is exclusively for written genres, the mean score of the learners' essays falls in a range that is applicable to spoken genres too. Hong Kong students' essays deviate from the native speakers' expectation of what written essays should be like by containing more involved language and more situation-dependent reference, causing the resultant written products to sound more spoken-like than they should be. The findings in Section 5.3.2.1 and Section 5.3.2.2 clearly demonstrate that Hong Kong students are not totally insensitive to mode differences, but obviously they are not sensitive enough.

5.3.2.3 Spoken-like nature of Hong Kong learners' written essays.

Before the end of Section 5.3, an analysis of the composition of the Dimension 1 score and the Dimension 3 score of the learner writing data is provided. As mentioned in Section 5.3.2.1 and Section 5.3.2.2, Hong Kong students' written essays are found to be more spoken-like along Dimension 1 Involved versus Informational Production and Dimension 3 Explicit versus Situation-dependent Reference. A discussion on the composition of the scores of these two textual dimensions can help readers better understand the causes behind the spoken-like nature of Hong Kong students' written essays. Table 39 and Table 40 provide the composition information of Dimension 1 score and the composition information of Dimension 3 score respectively. To recapitulate, each dimension score was calculated from the standardised frequencies of all the syntactic features loaded on that particular dimension (see Chapter 3 Section 3.5.3.2). In the tables, the first column classifies the syntactic features into positive ones and negative ones according to how they are loaded on the dimension in Biber's (1988) model. The second column lists all the syntactic features loaded on the dimension. The third column and the fourth column show the standardised frequencies in Hong Kong students' written essays and the standardised frequencies in native speakers' written essays respectively. The last column compares the two sets of standardised frequencies by subtracting the frequencies in the fourth column from the frequencies in the third column. In both tables, the dimension score of a particular set of data equals the sum of standardised frequencies of all positive features minus the sum of standardised frequencies of all negative features.

Table 39

Comparison Between Standardised Frequencies of Syntactic Features on Dimension 1 in Hong Kong Students' Writing Data ($SFreq_{HK}$) and Standardised Frequencies of Syntactic Features on Dimension 1 in Native Speakers' Writing Data ($SFreq_{NS}$)

	Syntactic Features on Dimension 1	$SFreq_{HK}$	$SFreq_{NS}$	$SFreq_{HK}$ - $SFreq_{NS}$
Positive Features	Private verbs	-1.07	-1.35	+0.28
	That deletion	-0.41	-0.45	+0.04
	Contractions	-0.52	-0.67	+0.15
	Present tense verbs	0.62	0.05	+0.57
	Second person pronouns	0.22	-0.56	+0.78
	Do as pro-verb	-0.71	-0.60	-0.11
	Analytic negation	1.34	-0.14	+1.48
	Demonstrative pronouns	-0.22	0.41	-0.63
	General emphatics	-1.08	-0.99	-0.09
	First person pronouns	-0.50	-0.72	+0.22
	Pronoun *it*	0.57	0.63	-0.06
	Be as main verb	-1.60	-1.65	+0.05
	Causative subordination	0.66	0.39	+0.27
	Discourse particles	-0.48	-0.49	+0.01
	Indefinite pronouns	0.85	0.35	+0.5
	General hedges	-0.25	-0.02	-0.23
	Amplifiers	0.40	0.07	+0.33
	Sentence relatives	-0.03	1.02	-1.05
	Wh questions	0.12	-0.19	+0.31
	Possibility modals	4.13	1.25	+2.88
	Clause coordination	-0.52	-0.65	+0.13
	Wh clauses	0.66	0.30	+0.36
	Stranded prepositions	-0.54	-0.55	+0.01
Negative Features	Nouns	0.38	0.85	-0.47
	Word length	0.48	0.17	+0.31
	Prepositions	-1.35	-0.82	-0.53
	Type/token ratio	0.56	1.43	-0.87
	Attributive adjectives	-0.52	-0.63	+0.11

Table 40

Comparison Between Standardised Frequencies of Syntactic Features on Dimension 3 in Hong Kong Students' Writing Data ($SFreq_{HK}$) and Standardised Frequencies of Syntactic Features on Dimension 3 in Native Speakers' Writing Data ($SFreq_{NS}$)

	Syntactic Features on Dimension 3	$SFreq_{HK}$	$SFreq_{NS}$	$SFreq_{HK} - SFreq_{NS}$
Positive Features	*Wh* relative clauses on object positions	-0.72	-0.58	-0.14
	Pied-piping constructions	-0.64	-0.42	-0.22
	Wh relative clauses on subject positions	0.15	0.84	-0.69
	Phrasal coordination	1.22	1.06	+0.16
	Nominalizations	-0.40	-0.10	-0.3
Negative Features	Time adverbials	-0.93	-0.33	-0.6
	Place adverbials	-0.52	-0.39	-0.13
	Adverbs	-0.35	-0.70	+0.35

On Dimension 1, the score of Hong Kong students' written essays (2.1) is larger than the score of native speakers' written essays (-5.6). As mentioned in the last paragraph, the dimension score was calculated by subtracting the sum of standardised frequencies of all negative features from the sum of standardised frequencies of all positive features. Consequently, it is those positive features with larger standardised frequencies and those negative features with smaller standardised frequencies in the learner writing data that contribute to the larger Dimension 1 score of Hong Kong students' written essays. In Table 39, positive features with positive standardised frequency differences, in descending order of the magnitude of the difference, are:

- Possibility modals (Feature 52)
- Analytic negation (Feature 67)
- Second person pronouns (Feature 7)
- Present tense (Feature 3)
- Indefinite pronouns (Feature 11)
- *Wh* clauses (Feature 23)
- Amplifiers (Feature 48)
- Direct *wh*-question (Feature 13)
- Private verbs (Feature 56)
- Causative adverbial subordinators (Feature 35)
- First person pronouns (Feature 6)
- Contractions (Feature 59)
- Independent clause coordination (Feature 65)
- *Be* as main verb (Feature 19)
- Subordinator *that* deletion (Feature 60)
- Discourse particles (Feature 50)
- Stranded prepositions (Feature 61)

In other words, it is the overuse of these 17 features that makes Hong Kong students' written essays become more spoken-like than the native speakers' essays. In addition to these overuse cases, the underuse of the following three features (i.e. negative features with negative standardised frequency differences in Table 39) also accounts for the spoken-like nature of the learners' essays. These features, in descending order of the magnitude of the difference, are type/token ratio (Feature 43), total prepositional phrases (Feature 39), and total other nouns (Feature 16).

On Dimension 3, the score of Hong Kong students' written essays (1.4) is smaller than the score of native speakers' essays (2.2). What causes the smaller Dimension 3 score is positive features with smaller standardised frequencies and negative features with larger standardised frequencies in the learner writing data. Therefore, what contributes to the spoken-like nature of Hong Kong students' written essays on Dimension 3 is positive features with negative standardised frequency differences and negative features with positive standardised frequency differences in

the last column of Table 40. These features, arranged in descending order of the magnitude of the standardised frequency difference, are *wh* relatives on subject positions (Feature 31), total adverbs (Feature 42), nominalizations (Feature 14), pied-piping relative clauses (Feature 33) and *wh* relatives on object positions (Feature 32). In other words, the factors that contribute to the spoken-like nature of Hong Kong students' written essays on Dimension 3, in descending order of influence, are the underuse of *wh* relatives on subject positions, the overuse of total adverbs, the underuse of nominalizations, the underuse of pied-piping relative clauses and the underuse of *wh* relatives on object positions.

The analysis in this last section completes the discussion on the mean dimension scores of Hong Kong students' written essays. Earlier parts of Section 5.3.2 have confirmed the speculation resulting from the analysis of syntactic features in Section 5.2 that Hong Kong students' essays are more spoken-like than native speakers' essays and demonstrated that the spoken-like nature of the learners' writing manifests itself in the relative focus on involved and informational communicative purpose (Dimension 1) and the relative amount of explicit and contextual reference (Dimension 3). Through an analysis of the standardised frequencies composing the Dimension 1 and the Dimension 3 scores in this last part of Section 5.3.2, the specific features that constitute the spoken-like nature of Hong Kong students' writing have been revealed. By understanding the source of the problem, second language learners can become more competent writers and avoid violations of mode differences in their writing.

5.4 Summary

This chapter reports and discusses findings on the writing data of Hong Kong learners. It analyses Hong Kong students' written essays from two perspectives: the use of individual syntactic features and the overall textual dimensions. Major findings are summarised as follows:

Use of syntactic features

- When compared with native speakers, Hong Kong students overuse the following features most in their written essays:
  - Present Tense
  - Third Person Personal Pronouns
  - Second Person Pronouns
  - Possibility Modals
  - Analytic Negation
  - Total Adverbs
  - First Person Pronouns
  - Conjuncts
  - Gerunds
  - Private Verbs
- When compared with native speakers, Hong Kong students underuse the following features most in their written essays:
  - Total Other Nouns
  - Total Prepositional Phrases
  - Past Tense
  - Perfect Aspect
  - Agentless Passives
  - Type/Token Ratio
  - Nominalizations
  - Demonstrative Pronouns
  - Split Auxiliaries
  - Time Adverbials
- When these 20 cases of most deviated usage are all taken into consideration, three general patterns emerge:
  - Hong Kong students rely more heavily on present tense constructions, but avoid other temporal perspectives.

- Hong Kong students substitute third person personal pronouns for noun phrases more frequently.
- Hong Kong students adopt a more interactive style by using more first person and second person pronouns.

- Among the 20 cases of most deviated usage, 15 features are related to the differences between speech and writing. A total of 12 cases of overuse and underuse suggest that Hong Kong students' essays are more spoken-like than native speakers' essays.

Textual Dimensions

- On Dimension 1 Involved versus Informational Production:
  - Hong Kong students' essays are more involved than native speakers' essays.
  - Hong Kong students' essays clearly fall within the range for spoken genres, whereas native speakers' essays do not.
  - Hong Kong students' essays are very similar to native speakers' prepared speeches.
- On Dimension 3 Explicit versus Situation-dependent Reference:
  - Hong Kong students' essays contain more situation-dependent reference than native speakers' essays do.
  - Hong Kong students' essays approach the range for spoken genres, but native speakers' essays clearly fall outside it.
  - Hong Kong students' essays are very similar to native speakers' spontaneous speeches.

- On Dimension 5 Abstract versus Non-abstract Information:
  - Hong Kong students' essays are more abstract than native speakers' essays.
  - Both Hong Kong students' essays and native speakers' essays clearly fall outside the range for spoken genres.
- Hong Kong students' essays are more spoken-like than native speakers' essays on Dimension 1 and Dimension 3. Hong Kong students' use of language is also comparable to native speakers' spoken language use on these two textual dimensions.

In response to Research Question 2, this chapter has provided evidence that Hong Kong students' written English exhibits characteristics typically found in English speech.

CHAPTER 6 CONCLUSION

6.1 Summary of Findings from Speech Data and Writing Data

Chapter 4 and Chapter 5 have presented findings from Hong Kong students' spoken English and findings from Hong Kong students' written English respectively. In this last chapter of the thesis, findings from the two different modes are combined so that a conclusion of Hong Kong students' sensitivity to differences between speech and writing can be reached. In the present study, Hong Kong students' oral presentations and written essays have been found to deviate from comparable native-speaker production in terms of a number of syntactic features. In general, Hong Kong students favour the use of present tense structure, tentative style, simple noun phrase structure and explicit clause-relation marking in oral presentation. They prefer using present tense constructions, adopting pronouns for nominal functions and using an interactive tone in written essays. The students' choice of syntactic features is shaped by many factors; however, the influence of inadequate sociolinguistic competence is clearly visible in both the students' spoken and written English. The strongest evidence is the overuse of conjuncts in the learner speech data and the overuse of second person pronouns in the learner writing data. When compared with native speakers, Hong Kong students tend to employ more conjuncts to indicate clausal relationships in oral presentations. Some of the conjuncts found in the learner speech data, notably *however*, *on the other hand* and *therefore*, are linking devices biased towards the written mode (Altenberg, 1984, p. 44; Altenberg, 1986, p. 18). When compared with native speakers, Hong Kong students also tend to use more second person pronouns in written essays. The second person pronouns in the learner writing data have been used predominantly as generic reference and the

generic *you* is usually regarded as the colloquial equivalent of the generic *one* (Quirk et al., 1985, p. 387). Both the overuse of conjuncts in the learner speech data and the overuse of second person pronouns in the learner writing data produce concrete, preliminary evidence of Hong Kong students' violation of usage conventional to the production mode.

Another analysis concerning syntactic features is to determine how many features with the biggest differences, in terms of frequency of occurrence, between the learners' production and the native speakers' production are at the same time features distributing differently in spoken and written English. In the present study, 16 of the 20 cases of most deviated usage in the learner speech data and 15 of the 20 cases of most deviated usage in the learner writing data are relevant to the variation across speech and writing illustrated in Biber (1988). Regardless of the reasons behind, Hong Kong students' use of these 31 features, which differs considerably from native speakers' expectation, is likely to affect the listeners' and the readers' perception of the learner language's suitability for the production mode. In particular, six features in the learner speech data and 12 features in the learner writing data probably create an impression that Hong Kong students' English produced in one mode retains properties of English in another mode.

In addition to individual syntactic features, Hong Kong students' oral presentations and written essays have also been found to deviate from comparable native-speaker production in terms of textual dimensions. These textual dimensions are some abstract constructs that Biber (1988) proposed to explain how native speakers varied their use of syntactic features among different genres in speech and writing. Hong Kong students' oral presentations differ from comparable native-speaker oral production by behaving more written-like on two of the three

relevant textual dimensions. Similarly, Hong Kong students' written essays deviate from comparable native-speaker written production by behaving more spoken-like on two of the three textual dimensions. The analysis provides strong support for Hong Kong students' inadequate sensitivity to differences between spoken and written English. In spoken English, the students seem to include too much explicit reference and abstract information, both of which are more typical of written communication. In written English, they appear to employ too many devices for expressing involvement and contextual reference, both of which are more commonly found in oral communication.

To further prove that Hong Kong students' English produced in one mode bears close resemblance to English in another mode, the two sets of learner language data have been compared, on the level of textual dimensions, with native-speaker genres produced in another mode. Hong Kong students' oral presentations are similar to the native-speaker written genres of popular lore, editorials and biographies in terms of the use of explicit/situation-dependent reference and the genres of popular lore, humour and biographies in terms of the inclusion of abstract/non-abstract information. Hong Kong students' written essays resemble native speakers' prepared speeches in terms of the involved/informational focus and resemble native speakers' spontaneous speeches in terms of the use of explicit/situation-dependent reference. Whereas the analysis on textual dimensions mentioned in the previous paragraph shows that Hong Kong students violate conventions associated with the specific mode, the findings described in this paragraph suggest that very probably people who listen to Hong Kong students' spoken English or who read Hong Kong students' written English will sense the similarity between the learners' English and the native speakers' English in the opposite mode.

Overall, findings from Chapter 4 and Chapter 5 have supplied an affirmative answer to the chief research question: Hong Kong students' English produced in one mode does exhibit characteristics typically found in English in another mode. The present study demonstrates that Hong Kong students have limited sociolinguistic competence. They do not know how to adjust the proportion of different syntactic features in order to match the conventional usage in spoken and written English. As a result, their English does not suit the production mode very well. As revealed by the analysis on textual dimensions, the problem is more serious regarding their oral production. Although Hong Kong students' written English is spoken-like, it still lies within the range of usage associated with the written mode. Hong Kong students' spoken English not only sounds written-like, but also, to some extent, exceeds the range of usage concerning the spoken mode. In other words, Hong Kong students are less capable of producing suitable oral English than of producing suitable written English.

6.2 Implications of the Present Study

The implications of the present study mainly fall into two categories. First, the research findings suggest that there is some room for improvement in Hong Kong students' English proficiency. Some actions may be taken by the teaching profession to help the students. These pedagogical recommendations are presented in Section 6.2.1. Second, the results of the present study raise a few issues which may form important considerations for future studies on second language acquisition. These research-related implications are discussed in Section 6.2.2.

6.2.1 Pedagogical Implications

According to the findings of the present study, Hong Kong students encounter difficulties in producing English that is appropriate for the specific production mode. They do not possess adequate sensitivity to the conventional usage of syntactic features in spoken and written English. The discussion in Chapter 4 Section 4.2.2.1 and the discussion in Chapter 5 Section 5.2.2.1 have shown that, for every syntactic feature, the students' usage represents the interplay of various factors. There is simply no single solution to all the deviated usage. Nevertheless, aiming at fostering Hong Kong students' sociolinguistic competence, the following suggestions on the teaching of grammar, speaking and writing, may help the students use syntactic features in a more native-like manner.

6.2.1.1 Teaching of grammar.

The Hong Kong students under investigation in the present study have received formal English instruction for about 13 years and have achieved intermediate to advanced English proficiency. What the present study has revealed about this group of second language learners is that even though their use of syntactic features is largely grammatically accurate, it is not necessarily contextually appropriate. For example, the second person pronouns in the learner writing data are not erroneous; their frequent occurrence just appears unnatural in expository or argumentative essays. Traditionally, grammar pedagogy focuses primarily on accuracy. One important implication of the present study is that classroom instruction on syntactic features should take into consideration the concept of appropriateness, besides accuracy.

As pointed out in Andrews (2003), although English textbooks have already incorporated the communicative approach in the design of grammar exercises, most teaching practice in Hong Kong secondary schools does not truly conform to communicative language teaching principles (pp. 368-369). The teachers still tend to adopt a deductive approach to grammar pedagogy, stating the rules explicitly and then asking the students to do exercises (Andrews, 2003, p. 365). To help Hong Kong students improve their sociolinguistic competence, English teachers, especially those teaching senior students, should combine the teaching of grammar with the teaching of language skills. Instead of isolating the presentation of grammar, they should link the grammar items to the students' reading, writing, listening or speaking exercises so that the students can understand more about how different syntactic features are used in different contexts.

Furthermore, textbook writers should provide teachers and students with explicit advice on the stylistic constraints of different syntactic features. Since Hong Kong students' English exposure comes mainly from the educational context, they may fail to realise how different syntactic features are used in other English genres. Explicit warnings can help them avoid inappropriate use of syntactic features. The writers may consult some corpus-informed grammar books, e.g. Biber, Johansson, Leech, Conrad and Finegan (1999) and Carter and McCarthy (2006), to complement information in traditional grammar books. In this way, English textbooks can offer teachers and students useful information on both the theoretical aspect and the practical aspect of English syntactic features. In addition to textbook writers, the education authority should also update its knowledge on corpus-based grammar research. It should consider how the existing English curriculum may benefit from the new research findings, e.g. the distributional characteristics of different syntactic

features.

6.2.1.2 Teaching of speaking.

Findings from the learner speech data of the present study suggest that Hong Kong students' spoken English is highly similar to written English in certain aspects. In spite of the overlapping nature of spoken and written English, the spoken language does exhibit some unique characteristics. Sounding too bookish, Hong Kong students' spoken English may lead to ineffective communication and even leave a bad impression on the audience. One reason why Hong Kong students' use of syntactic features in spoken English is written-like is that traditional teaching of grammar emphasises the structure of written English. Little is said about the syntactic structure of spoken English. To improve Hong Kong students' oral English proficiency, textbook writers should highlight how spoken English is different from written English. As mentioned in Section 6.2.1.1, teachers can incorporate the teaching of grammar into the teaching of language skills. For example, after their students finish a listening comprehension exercise, the teachers can re-play some parts of the audio-materials and draw the students' attention to some constructions that are common in spoken English. Gradually, Hong Kong students' sensitivity to conventional spoken English usage can be increased.

Another suggestion that is conducive to Hong Kong students' spoken English development is to encourage the recording of students' speaking performance. Owing to the transient nature of speech, it is difficult to monitor the students' progress in speaking. On the one hand, teachers cannot possibly listen carefully to all 30 or 40 students simultaneously during the oral English class. On the other hand, upon the completion of an oral task, students quickly forget what they have said.

They have no clear idea of how well or how poorly they speak. Consequently, both teachers and students know little about the students' strengths and weaknesses in oral communication. One way to solve the problem is to have the students' speaking practice recorded. In doing so, the students are able to listen back to their own performance after the end of the oral English activity. This provides them with the opportunity to reflect on their speaking ability so that they can improve accordingly. Just as how peer review is done in writing, the teachers can give students some guidelines for evaluating their performance in the oral task. Moreover, the recordings allow the teachers to keep a "portfolio" of each student's speaking performance so that they can better understand each student's oral English proficiency as well as improvement over time. Such information certainly assists the teachers in planning and evaluating their teaching activities. In fact, language laboratories or computer laboratories are very common in Hong Kong secondary schools. There is not much technical difficulty associated with the recording of students' oral performance. The recording needs not be done in every speaking class; it may be done, for example, once a month or three times a semester, as deemed appropriate by the teachers. This method, which makes it possible to keep track of Hong Kong students' oral English performance, can help them improve their spoken English.

6.2.1.3 Teaching of writing.

The teaching of English writing in Hong Kong secondary school classrooms is characterised by a focus on content in the pre-task stage and a focus on grammar in the post-task stage. Wong (1992) has observed that English textbooks mainly offer advice on content when teaching argumentative writing (pp. 138, 148). I. Lee's (2008) survey shows that 94.1% of the written feedback in secondary school

students' compositions is form-focused and nearly all such feedback concerns grammar and vocabulary errors (pp. 76-77). However, the writing pedagogy for senior secondary school students should go beyond these two foci. Having mastered basic English writing skills, these students should learn how to become more effective writers. In the pre-writing stage, in addition to brainstorming ideas with their students, teachers can guide their students to analyse the communicative situation specified in the writing task and provide some advice on appropriate writing style. For example, when teaching argumentative writing, they can tell their students that despite the need for expressing personal stance, writers of argumentative essays should discuss the subject matter objectively and consider arguments from both sides so as to achieve the purpose of convincing readers. When marking their students' compositions, the teachers should not devote all their attention to proofreading. They should also evaluate the overall effectiveness of the whole piece of writing. Comments on organisation and language style are certainly beneficial to the students' development in English writing proficiency.

Secondary school English teachers' emphasis on grammatical accuracy in marking writing is heavily influenced by the examination-oriented culture in the Hong Kong educational system (I. Lee, 2008, p. 80). Perhaps the best way to change the existing pedagogical practice is to revise the assessment policy. For example, the criteria for marking essays in the UE examination are content, which involves relevant content, logical organization and well-constructed paragraphs, and language accuracy, which includes appropriate vocabulary, variety of sentence types, and accurate grammar and mechanics (Hong Kong Examinations Authority, 2001, p. 15). Although language appropriateness is not totally ignored, it only refers to vocabulary of appropriate variety and of an appropriate level of formality (Hong Kong

Examinations Authority, 2001, p. 15). The same marking scheme can still be found in the most recent UE examination (Hong Kong Examinations and Assessment Authority, 2009, p. 124). It seems that the examination authority only attaches importance to grammatical accuracy and linguistic variety in Hong Kong students' written language use. To provide an incentive for English teachers to pay more attention to language style in writing pedagogy, the examination authority should consider increasing emphasis on appropriate language style in the marking criteria. It can give the markers specific guidelines as to what kind of language style is expected in each essay topic and what kind of structure should be avoided. As suggested by Hughes (2003), the backwash of testing is not necessarily harmful; it can be beneficial to teaching, too (p. 2). A change in the marking criteria of the writing examination can improve the teaching of writing indirectly.

6.2.2 Research Implications

Besides pedagogical implications, the present study also has some implications which may help second language acquisition researchers to devise future research plans. Since the present study centres around Hong Kong English language learners, the discussion in this section addresses research on Hong Kong students' English before moving on to research concerning second language learners in general.

6.2.2.1 Features of Hong Kong learners' English.

The significance of the present study partly lies in its extensive survey of syntactic features in Hong Kong students' spoken and written English. As illustrated in the discussion in Chapter 4 Section 4.2.2.1 and Chapter 5 Section 5.2.2.1, patterns of overuse and underuse are influenced by decisions on research design, notably the

choice of the elicitation task and the choice of the target language norm. As a result, patterns identified in one study do not necessarily represent characteristics of Hong Kong students' English produced under different conditions. For example, in the present study, third person personal pronouns and total adverbs are both underused features in Hong Kong students' oral presentations. However, one should not readily claim that these two underuse patterns are typical in Hong Kong students' English. Actually, the two features have been found to be overused in Hong Kong students' written essays in the present study. What the above suggests is that, in order to understand more about the inherent properties of Hong Kong students' English, researchers should diversify the source of their learner language data. Existing studies on Hong Kong students' use of syntactic features are based primarily on written data, especially academic essays. More research should be conducted on other written genres, as well as on spoken data.

In the present study, some patterns of overuse and underuse in the learner speech data are identical to those in the learner writing data. These patterns, existing across two very different types of English language production, are likely to be found in other genres produced by Hong Kong students, too. The overuse patterns include present tense, second person pronouns, conjuncts and possibility modals; the underuse patterns include past tense, the perfect aspect, time adverbials, agentless passives and total prepositional phrases. These features can be some topics of investigation for researchers interested in Hong Kong second language learners' English. Furthermore, as the present study is based on language data produced by students completing their secondary school education, i.e. a group of learners whose English development has become relatively stable, the findings can provide insights into the variety of Hong Kong English in general.

6.2.2.2 Second language learners' sociolinguistic competence.

One rationale of the present study is to fill the existing research gap in second language acquisition literature and provide information on second language learners' sociolinguistic competence. Although the present study only investigates one specific group of second language learners, i.e. Chinese learners of English in Hong Kong, the results are of relevance to learners from other cultural backgrounds as well. The reason is that some patterns of overuse and underuse of syntactic features identified in the present study are equally found in English produced by other second language learners. For example, first person and second person pronouns are overused not only in Chinese learners' writing, but also in the writing of Japanese, Korean, Vietnamese, Dutch, Finnish, French and Swedish learners (Hinkel, 2002, p. 86; Petch-Tyson, 1998, p. 112). The perfect aspect and the passive voice are underused by Chinese, Japanese, Korean, Vietnamese, Indonesian and Arabic learners alike in writing (Hinkel, 2002, pp. 101, 113). It seems that second language learners of English, regardless of their mother tongue, all employ syntactic features quite differently from native speakers. It is possible that other groups of second language learners, similar to the Hong Kong students in the present study, are not very sensitive to the differences between spoken and written English. Future research should further pursue inquiries into second language learners' sociolinguistic competence. The similarities in overuse and underuse of syntactic features among different groups of second language learners may suggest that some common developmental patterns exist when second language learners develop their sensitivity to mode differences. It would be interesting to conduct longitudinal studies to find out how second language learners' sociolinguistic competence changes over time, the knowledge of which is crucial to the construction and the validation of

theoretical models of second language acquisition.

6.3 Limitations of the Present Study

The present study replicates one part of Biber's (1988) research methodology. The identification of syntactic features and the generation of statistics both closely follow Biber's (1988) practice. Consequently, some limitations of Biber (1988) also apply to the present study. These limitations, together with other limitations in the research design of the present study, are discussed in this section.

6.3.1 Biber's (1988) Methodology

There are two issues noticed during the implementation of Biber's (1988) methodology. The first issue concerns the searching of syntactic features. As noted in Chapter 3 Section 3.7, the instructions on the search of some features in Biber (1988), e.g. gerunds and emphatics, are not clear enough. This lack of clarity presents considerable difficulty for researchers who want to replicate the study. It causes confusion to the readers, too, because they cannot find out what has and what has not been counted for those features. Besides the unclear information, another limitation of Biber's (1988) search for syntactic features lies in his search method, which is limited in scope. It was noticed during the searching process of the present study that the search results of some features left out quite a number of instances of the features. One very good example is the feature of *be* as main verb, which, according to Biber (1988), includes only instances of *am, is, are, was, were, being, been, 'm, 're, isn't, aren't, wasn't,* and *weren't* that are followed directly by articles, demonstratives, quantifiers, numerals, possessive pronouns, address titles, prepositions and adjectives (see Feature 19 in Appendix A). At least three types of *be* as main verb are

ignored: (a) copular *be* followed by other word classes such as adverbs (e.g. *It is very difficult*); (b) copular *be* in the base form, e.g. when preceded by modal verbs (e.g. *It can be useful*); and (c) copular *be* followed by the possessive pronoun *her*, which is excluded from Biber's (1988) list of possessive pronouns (e.g. *It is her responsibility*). Biber's (1988) restricted search method reduces the accuracy in the search for some syntactic features, adversely affecting the validity of the data.

The second issue observed during the implementation of the present study is that Biber's (1988) formula for calculating dimension scores does not take into consideration the varying degrees of importance of different syntactic features. According to the formula, all features loaded on the same side of a textual dimension are given equal weighting. Nonetheless, the features were originally assigned different factor loadings in the factor analysis which generated the textual dimensions in Biber (1988) (see Appendix K). These loadings express the relative importance of the different syntactic features on the same dimension:

> a factor loading indicates the extent to which one can generalize from a factor to a particular linguistic feature, or the extent to which a given feature is representative of the dimension underlying a factor. The loading of a feature on a factor reflects the extent to which the variation in the frequency of that feature correlates with the overall variation of the factor; it indicates the strength of the co-occurrence relationship between the feature in question and the factor as a whole. (Biber, 1988, pp. 85-87)

Since different features on the same textual dimension share different amounts of variation with the dimension, it seems unlikely that they are of equal importance to the dimension. Biber's (1988) method for calculating dimension scores may understate the influence of some important features and overemphasise the influence

of some less important features.

6.3.2 Other Limitations

In addition to the limitations associated with Biber's (1988) methodology, there are two other limitations related to the standard of reference adopted in the comparison of the learner speech data. In the present study, the target language norm chosen for the comparison of Hong Kong students' oral presentations is prepared speeches in Biber (1988). As mentioned in Chapter 3 Section 3.4.2.1.1, these speeches, taken from the London-Lund Corpus of Spoken English, were produced by speakers who were much older than the Hong Kong students. Moreover, the speeches were delivered for a professional purpose. In other words, the learner speech data and the native-speaker speech data are different not only in the variable of nativeness, but also in other variables. This represents one limitation of the present study. Nevertheless, as observed by Granger (1998), there are limitations for most control corpora (p. 13). An ideal standard of reference, which is identical to the target data in every way except the variable under investigation, hardly exists in reality. What is most important is that differences between the control corpus and the learner corpus are duly acknowledged and their effects on the research results are taken into consideration in the interpretation of the findings.

Another limitation of the present study lies in the format of the prepared speech data. The native speakers' prepared speeches to which Hong Kong students' oral presentations were compared was adopted from Biber (1988) in the form of ready-made statistics. One drawback of this decision is that qualitative analysis of the prepared speeches has become impossible. Uncovering what is absent from the learner language data, such analysis can promote understanding on Hong Kong

students' underuse of syntactic features. It is especially desirable for a feature with miscellaneous syntactic functions like total adverbs in the present study. Quirk et al. (1985) describes adverbs as "the most nebulous and puzzling of the traditional word classes" (p. 438), owing to its members' heterogeneous nature. In Biber (1988), little is said about this feature and its functions: "The class 'other adverbs' has a much broader range of functions, which includes time and place reference in addition to specification of manner, etc." (p. 110). It is very hard to provide an explanation for the underuse of total adverbs in the learner speech data. For one thing, the multi-functional nature of adverbs makes it difficult to deduce the reason from the theoretical point of view. For another, the learner data itself hardly provides any clue as to what kind of adverbs is missing. In a case like this, qualitative analysis of the native-speaker data would be helpful. Therefore, the unavailability of textual data for native speakers' prepared speeches is a limitation in the present study. But since most other underused features in the learner speech data have quite clear syntactic functions, qualitative analysis of the native-speaker data is not essential. The limitation described in this paragraph thus does not exert serious influence on the analysis of Hong Kong students' oral presentations as a whole.

6.4 Suggestions for Further Research

In view of the issues mentioned in this last chapter of the thesis, some directions for future research are suggested. Biber (1988) was conducted more than 20 years ago. During these 20 years, there have been a lot of advances in computer technology and corpus linguistics. As there is some room for improvement in the retrieval of syntactic features in Biber (1988), it would be meaningful to conduct another similar multi-dimensional study by using more powerful corpus tools.

Nowadays, with corpora like ICE-GB, which is grammatically and syntactically annotated, the searching of syntactic features can be much more accurate and convenient. Besides verifying Biber's (1988) claims, the new study can extend knowledge on the distribution of syntactic features across speech and writing by including features which could not be retrieved automatically and were hence excluded in Biber (1988), e.g. *it*-clefts and fronted *that*-clauses (p. 221). It can also take into consideration other levels of linguistic features such as discourse features counted in Louwerse et al. (2004). Conducting a more comprehensive and accurate survey of English linguistic features can help differences between speech and writing to be better modelled in the future.

Regarding research on second language acquisition, the present study has shown that second language learners' sociolinguistic competence is a research area worthy of attention and that the research methodology used in this study has the potential to be applied to other varieties or other registers. Further research on sensitivity to mode differences can be conducted by using data from other second language learners of English, e.g. other Southeast Asian learners and European learners, or data from other genres, e.g. group discussions and narratives. Future research can also study how different task conditions may affect second language learners' performance, as Ädel (2008) has shown that the use of involvement features in learners' writing varies in timed and untimed essays, as well as in essays with and without secondary sources (p. 46). In addition to sensitivity to spoken and written English, future studies can examine other aspects of sociolinguistic competence like learners' awareness of the differences between formal and informal discourse. Investigating second language learners' sociolinguistic competence can improve current understanding on how learners use the target language in different

contexts, hence providing insights into the learners' communicative ability.

Unlike the phonological aspect of Hong Kong English, the morphosyntactic aspect of Hong Kong English is an under-researched area (Gisborne, 2009, p. 151). Although the present study has included an extensive survey of syntactic features, the data only represents Hong Kong English produced by speakers who are in their late teens. More studies should be conducted on the use of syntactic features by other speakers of Hong Kong English, e.g. those who are required to use English in the workplace. Ideally, a survey of syntactic features can be carried out on a representative sample of Hong Kong English. Through these studies, similarities and differences between Hong Kong English and other varieties of English can be revealed. As a result, Hong Kong English can be better established as a new English variety.

REFERENCES

Ädel, A. (2008). Involvement features in writing: Do time and interaction trump register awareness? In G. Gilquin, S. Papp & M. B. Díez-Bedmar (Eds.), *Linking up contrastive and learner corpus research* (pp. 35-53). Amsterdam: Rodopi.

Akinnaso, F. N. (1982). On the differences between spoken and written language. *Language and Speech, 25,* 97-125.

Altenberg, B. (1984). Causal linking in spoken and written English. *Studia Linguistica, 38,* 20-69.

Altenberg, B. (1986). Contrastive linking in spoken and written English. In G. Tottie & I. Bäcklund (Eds.), *English in speech and writing: A symposium* (pp. 13-40). Uppsala: Almqvist and Wiksell.

Andrews, S. (2003). "Just like instant noodles": L2 teachers and their beliefs about grammar pedagogy. *Teachers and teaching: Theory and practice, 9*(4), 351-375.

Bachman, L. F. (1990). *Fundamental considerations in language testing.* Oxford: Oxford University Press.

Beaman, K. (1984). Coordination and subordination revisited: Syntactic complexity in spoken and written narrative discourse. In D. Tannen (Ed.), *Coherence in spoken and written discourse* (pp. 45-80). Norwood, N.J.: Ablex.

Berry, R. (2005). Who do they think 'we' is? Learners' awareness of personality in pedagogic grammars. *Language Awareness, 14,* 84-96.

Berry, R. (2009). You could say that: The generic second-person pronoun in modern English. *English Today, 25*(3), 29-34.

Biber, D. (1986). Spoken and written textual dimensions in English: Resolving the contradictory findings. *Language, 62*(2), 384-414.

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

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison.* Cambridge: Cambridge University Press.

Biber, D. (2001). Dimensions of variation among eighteenth-century speech-based and written registers. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 200-214). Harlow: Longman.

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

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

Biber, D., & Finegan, E. (2001). Diachronic relations among speech-based and written registers in English. In S. Conrad & D. Biber (Eds.), *Variation in English: Multi-dimensional studies* (pp. 66-83). Harlow: Longman.

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

Biber, D., & Vásquez, C. (2008). Writing and speaking. In C. Bazerman (Ed.), *Handbook of research on writing: History, society, school, individual, text* (pp. 538-548). New York: Routledge.

Bolton, K., Nelson, G., & Hung, J. (2002). A corpus-based study of connectors in student writing. *International Journal of Corpus Linguistics, 7*(2), 165-182.

Boyle, J., & Boyle, L. (1991). *Common spoken English errors in Hong Kong.* Hong Kong: Longman.

Bunton, D. (1989). *Common English errors in Hong Kong*. Hong Kong: Longman.

Bunton, D. (1991). A comparison of English errors made by Hong Kong students and those made by non-native learners of English internationally. *Institute of Language in Education Journal, Special Issue No. 2,* 9-22.

Canale, M. (1983). From communicative competence to communicative language pedagogy. In J. C. Richards & R. W. Schmidt (Eds.), *Language and communication* (pp. 2-27). London: Longman.

Carroll, D. W. (1999). *Psychology of language* (3rd ed.). Pacific Grove, CA: Brooks/Cole Publishing Company.

Carter, R., & McCarthy, M. (2006). *Cambridge grammar of English: A comprehensive guide: Spoken and written English grammar and usage.* Cambridge: Cambridge University Press.

Chafe, W. L. (1982). Integration and involvement in speaking, writing, and oral literature. In D. Tannen (Ed.), *Spoken and written language: Exploring orality and literacy* (pp. 35-53). Norwood, N.J.: Ablex.

Chafe, W. L. (1985). Linguistic differences produced by differences between speaking and writing. In D. R. Olson, N. Torrance & A. Hildyard (Eds.), *Literacy, language, and learning: The nature and consequences of reading and writing* (pp. 105-123). Cambridge: Cambridge University Press.

Chafe, W., & Danielewicz, J. (1987). Properties of spoken and written language. In R. Horowitz & S. J. Samuels (Eds.), *Comprehending oral and written language* (pp. 83-113). San Diego: Academic Press.

Chan, A. Y. W. (2004a). Noun phrases in Chinese and English: A study of English structural problems encountered by Chinese ESL students in Hong Kong. *Language, Culture and Curriculum, 17*(1), 33-47.

Chan, A. Y. W. (2004b). Syntactic transfer: Evidence from the interlanguage of Hong Kong Chinese ESL learners. *The Modern Language Review, 88*(1), 56-74.

Chan, B. (1991). A study of errors made by F6 students in their written English with special reference to structures involving the transitive verb and the passive construction. *Institute of Language in Education Journal, Special Issue No. 2,* 43-51.

Chao, Y.-C. (2003). *Contrastive rhetoric, lexico-grammatical knowledge, writing expertise, and metacognitive knowledge: An integrated account of the development of English writing by Taiwanese students* (Doctoral dissertation, University of Auckland, New Zealand). Retrieved from Digital Dissertation Consortium. (UMI No. 3119448)

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Québec replications of three European studies. *The Canadian Modern Language Review, 59*(3), 393-423.

Collot, M., & Belmore, N. (1996). Electronic language: A new variety of English. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 13-28). Amsterdam: John Benjamins.

Colombi, M. C., & Schleppegrell, M. J. (2002). Theory and practice in the development of advanced literacy. In M. J. Schleppegrell & M. C. Colombi (Eds.), *Developing advanced literacy in first and second languages: Meaning with power* (pp. 1-19). Mahwah, N.J.: Lawrence Erlbaum Associates.

Cramer, D. (2003). *Advanced quantitative data analysis.* Philadelpha: Open University Press.

Crossley, S. A., & Louwerse, M. (2007). Multi-dimensional register classification using bigrams. *International Journal of Corpus Linguistics, 12*(4), 453-478.

Curriculum Development Council. (2004). *English language curriculum guide (primary 1-6)*. Hong Kong: Government Logistics Department.

Curriculum Development Council, & Hong Kong Examinations and Assessment Authority. (2007). *English language curriculum and assessment guide (secondary 4-6)*. Hong Kong: Government Logistics Department.

DeVito, J. A. (1966). Psychogrammatical factors in oral and written discourse by skilled communicators. *Speech Monographs, 33,* 73-76.

DeVito, J. A. (1967). Levels of abstraction in spoken and written language. *Journal of Communication, 17,* 354-361.

Drieman, G. H. J. (1962). Differences between written and spoken language. *Acta Psychologica, 20,* 36-57, 78-100.

Ellis, R. (1994). *The study of second language acquisition*. Oxford: Oxford University Press.

Færch, C., & Kasper, G. (1983). Plans and strategies in foreign language communication. In C. Færch & G. Kasper (Eds.), *Strategies in interlanguage communication* (pp. 20-60). London: Longman.

Fung, L., & Carter, R. (2007). Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics, 28*(3), 410-439.

Gilquin, G., & Paquot, M. (2008). Too chatty: Learner academic writing and register variation. *English Text Construction, 1*(1), 41-61.

Gisborne, N. (2009). Aspects of the morphosyntactic typology of Hong Kong English. *English World-Wide, 30*(2), 149-169.

Granger, S. (1998). The computer learner corpus: A versatile new source of data for SLA research. In S. Granger (Ed.), *Learner English on computer* (pp. 3-18). London: Longman.

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

Green, C. F. (1991). Teacher perceptions of the relative gravity of errors in written English. *Institute of Language in Education Journal, Special Issue No. 2,* 69-76.

Greenbaum, S. (1993). The tagset for the International Corpus of English. In C. Souter & E. Atwell (Eds.), *Corpus-based computational linguistics* (pp. 11-24). Amsterdam: Rodopi.

Greenbaum, S., & Nelson, G. (1999). Elliptical clauses in spoke and written English. In P. Collins & D. Lee (Eds.), *The clause in English: In honour of Rodney Huddleston* (pp. 111-126). Amsterdam: John Benjamins.

Greenbaum, S., & Svartvik, J. (1990). The London-Lund Corpus of Spoken English. In J. Svartvik (Ed.), *The London-Lund Corpus of Spoken English: Description and research* (pp. 11-59). Lund: Lund University Press.

Gumperz, J. J., Kaltman, H., & O'Connor, M. C. (1984). Cohesion in spoken and written discourse: Ethnic style and the transition to literacy. In D. Tannen (Ed.), *Coherence in spoken and written discourse* (pp. 3-19). Norwood, N.J.: Ablex.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. London: Longman.

Halliday, M. A. K., McIntosh, A., & Strevens, P. (1964). *The linguistic sciences and language teaching*. London: Longmans.

Halpern, J. W. (1984). Differences between speaking and writing and their implications for teaching. *College Composition and Communication, 35,* 345-357.

Hinkel, E. (2002). *Second language writer's text: Linguistic and rhetorical features*. Mahwah, N.J.: Lawrence Erlbaum Associates.

Hong Kong Examinations and Assessment Authority. (2008). Examination statistics: 2008 HKCEE analysis of results of candidates in each subject [Table]. Retrieved from http://www.hkeaa.edu.hk/DocLibrary/HKCEE /Release_of_Results/Exam_Report/Examination_Statistics/ceexamstat08_5.pdf

Hong Kong Examinations and Assessment Authority. (2009). *HKALE. AS use of English 2009*. Hong Kong: Hong Kong Examinations and Assessment Authority.

Hong Kong Examinations Authority. (1999). *HKALE question papers: AS use of English 1999*. Hong Kong: Hong Kong Examinations Authority.

Hong Kong Examinations Authority. (2000). *HKALE question papers: AS use of English 2000*. Hong Kong: Hong Kong Examinations Authority.

Hong Kong Examinations Authority. (2001). *HKALE question papers: AS use of English 2001*. Hong Kong: Hong Kong Examinations Authority.

Hughes, A. (2003). *Testing for language teachers*. Cambridge: Cambridge University Press.

Hyland, K. (2004). Pattern of engagement: Dialogic features and L2 undergraduate writing. In L. J. Ravelli & R. A. Ellis (Eds.), *Analysing academic writing: Contextualized frameworks* (pp. 5-23). London: Continuum.

Hyland, K., & Milton, J. (1997). Qualification and certainty in L1 and L2 students' writing. *Journal of Second Language Writing, 6*(2), 183-205.

Johansson, C. (1993). Whose and of which with non-personal antecedents in written and spoken English. In C. Souter & E. Atwell (Eds.), *Corpus-based computational linguistics* (pp. 97-116). Amsterdam: Rodopi.

Johansson, C., & Geisler, C. (1998). Piped piping in spoken English. In A. Renouf (Ed.), *Explorations in corpus linguistics* (pp. 67-82). Amsterdam: Rodopi.

Johnstone, B. (2002). *Discourse analysis*. Malden, Mass.: Blackwell Publishers.

Khattree, R., & Naik, D. N. (2000). *Multivariate data reduction and discrimination with SAS software.* Cary, N.C.: SAS Institute Inc.

Krashen, S. D. (1982). *Principles and practice in second language acquisition.* Oxford: Pergamon.

Krauthamer, H. S. (1999). *Spoken language interference patterns in written English.* New York: Peter Lang.

Kroll, B. M. (1981). Developmental relationships between speaking and writing. In B. M. Kroll & R. J. Vann (Eds.), *Exploring speaking-writing relationships: Connections and contrasts* (pp. 32-54). Illinois: National Council of Teachers of English.

Lee, D. (2005, June 22). David Lee's personal homepage. Retrieved from http://personal.cityu.edu.hk/~davidlee/devotedtocorpora/home/phd.htm

Lee, I. (2008). Understanding teachers' written feedback practices in Hong Kong secondary classrooms. *Journal of Second Language Writing, 17*, 69-85.

Littlewood, W., & Li, D. (2006). The sociolinguistic awareness of tertiary level students in Hong Kong and Mainland China. *Language Awareness, 15*(2), 97-109.

Lorr, M. (1983). *Cluster analysis for social scientists.* San Francisco: Jossey-Bass.

Louwerse, M. M., McCarthy, P. M., McNamara, D. S., & Graesser, A. C. (2004). Variation in language and cohesion across written and spoken registers. In K. Forbus, D. Gentner & T. Regier (Eds.), *Proceedings of the 26th Annual Meeting of the Cognitive Science Society* (pp. 843–848). Mahwah, NJ: Erlbaum.

Matthews, S., & Yip, V. (1994). *Cantonese: A comprehensive grammar*. London: Routledge.

McCrostie, J. (2008). Writer visibility in EFL learner academic writing: A corpus-based study. *ICAME Journal, 32*, 97-114.

Miller, J. E. (1994). Speech and writing. In R. E. Asher et al. (Eds.), *The Encyclopedia of language and linguistics (Vol. 8)* (pp. 4301-4306). Oxford: Pergamon Press.

Milton, J. (2001). *Elements of a written interlanguage: A computational and corpus-based study of institutional influences on the acquisition of English by Hong Kong Chinese students.* Hong Kong: Language Centre, Hong Kong University of Science and Technology.

Nation, I. S. P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nelson, G., Wallis, S., & Aarts, B. (2002). *Exploring natural language: Working with the British component of the International Corpus of English.* Amsterdam: John Benjamins Publishing Company.

Ochs, E. (1979). Planned and unplanned discourse. In T. Givón (Ed.), *Discourse and syntax* (pp. 51-80). New York: Academic Press.

O'Donnell, R. C. (1974). Syntactic differences between speech and writing. *American Speech, 49,* 102-110.

Petch-Tyson, S. (1998). Writer/reader visibility in EFL written discourse. In S. Granger (Ed.), *Learner English on computer* (pp. 107-118). London: Longman.

Platt, J. T. (1982). English in Singapore, Malaysia and Hong Kong. In R. W. Bailey & M. Gorlach (Eds.), *English as a world language* (pp. 384-414). Ann Arbor: University of Michigan Press.

Quirk, R., Greenbaum, S., Leech, G., & Svartvik, J. (1985). *A comprehensive grammar of the English language*. London: Longman.

Read, J. (2000). *Assessing vocabulary.* Cambridge: Cambridge University Press.

Riddle, E. (1986). The meaning and discourse function of the past tense in English. *TESOL Quarterly, 20*(2), 267-286.

Roberts, C., & Street, B. (1997). Spoken and written language. In F. Coulmas (Ed.), *The handbook of sociolinguistics* (pp. 168-186). Oxford: Blackwell Publishers.

Scott, M. (2008). WordSmith Tools (Version 5) [Computer software]. Liverpool: Lexical Analysis Software.

Shaw, P., & Liu, E. T.-K. (1998). What develops in the development of second-language writing? *Applied Linguistics, 19*(2), 225-254.

Szpara, M. Y., & Wylie, E. C. (2007). Writing differences in teacher performance assessments: An investigation of African American Language and Edited American English. *Applied Linguistics, 29*(2), 244-266.

Tannen, D. (1982). Oral and literate strategies in spoken and written narratives. *Language, 58*, 1-21.

Tannen, D. (1985). Relative focus on involvement in oral and written discourse. In D. R. Olson, N. Torrance & A. Hildyard (Eds.), *Literacy, language, and learning: The nature and consequences of reading and writing* (pp. 124-147). Cambridge: Cambridge University Press.

Tao, L., & Healy, A. F. (2005). Transfer of reference tracking strategies from Chinese to English. *Journal of Psycholinguistic Research, 34*(2), 99-131.

Tiee, H. H.-Y., & Lance, D. M. (1986). *A reference grammar of Chinese sentences with exercises*. Tucson, Ariz.: University of Arizona Press.

UCL/CECL Locness corpus. (n.d.). Retrieved from http://www.fltr.ucl.ac.be/fltr /germ/etan/cecl/Cecl-Projects/Icle/locness1.htm

University Centre for Computer Corpus Research on Language. (n.d.). *UCREL CLAWS7 Tagset.* Retrieved from http://ucrel.lancs.ac.uk/claws7tags.html

Van Rooy, B. (2008). A multidimensional analysis of student writing in Black South African English. *English World-Wide, 29*(3), 268-305.

Wales, K. (1996). *Personal pronouns in present-day English.* Cambridge: Cambridge University Press.

Wallis, S. (n.d.). ICE-GB Tagger (Version 1.1) [Computer software]. Retrieved from http://www.ucl.ac.uk/english-usage/staff/sean/

Walls, Y. L., & Walls, J. W. (2009). *Using Chinese: A guide to contemporary usage.* Cambridge: Cambridge University Press.

Webster, M., & Lam, W. C.-P. (1991). Further notes on the influence of Cantonese on the English of Hong Kong students. *Institute of Language in Education Journal, Special Issue No. 2,* 35-42.

Webster, M. A., Ward, A., & Craig, K. (1987). Language errors due to first language interference (Cantonese) produced by Hong Kong students of English. *Institute of Language in Education Journal, 2,* 63-81.

Winch, C., & Gingell, J. (1994). Dialect interference and difficulties with writing: An investigation in St. Lucian primary schools. *Language and Education, 8*(3), 157-182.

Wong, Y.-K. (1992). *Instruction in argumentative writing in Hong Kong secondary schools: A contrastive study of Chinese-English rhetoric (volumes I and II)* (Unpublished doctoral dissertation). Georgetown University, Washington, DC, USA.

Yang, S., & Huang, Y. Y. (2004). The impact of the absence of grammatical tense in L1 on the acquisition of the tense-aspect system in L2. *IRAL, 42*, 49-70.

Yip, V., & Matthews, S. (1991). Relative complexity: Beyond avoidance. *CUHK Papers in Linguistics, 3,* 112-124.

Yip, P.-C., & Rimmington, D. (1997). *Chinese: An essential grammar*. London: Routledge.

APPENDICES

Appendix A Biber's (1988) Search Methods for Syntactic Features (pp. 222-245)

Syntactic Feature	Search Method
1. Past Tense	• Any past tense verb form in dictionary; or • Any word longer than 6 letters and ending in *-ed*
2. Perfect Aspect	• Any word string beginning with the auxiliary verb *have* (any form, excluding *'s*) and ending with a verb in past tense/past participle form ▪ With 0, 1 or 2 adverb(s) in-between; or ▪ With a noun/pronoun in-between (for inverted forms in questions)
3. Present Tense	• Any base form of verbs in dictionary, excluding infinitives; or • Any third person singular present verb form in dictionary
4. Place Adverbials	• *aboard, above, abroad, across, ahead, alongside, around, ashore, astern, away, behind, below, beneath, beside, downhill, downstairs, downstream, east, far, hereabouts, indoors, inland, inshore, inside, locally, near, nearby, north, nowhere, outdoors, outside, overboard, overland, overseas, south, underfoot, underground, underneath, uphill, upstairs, upstream, west*
5. Time Adverbials	• *afterwards, again, earlier, early, eventually, formerly, immediately, initially, instantly, late, lately, later, momentarily, now, nowadays, once, originally, presently, previously, recently, shortly, simultaneously, soon, subsequently, today, tomorrow, tonight, yesterday*
6. First Person Pronouns	• *I, me, we, us, my, our, myself, ourselves* (plus contracted forms)
7. Second Person Pronouns	• *you, your, yourself, yourselves* (plus contracted forms)
8. Third Person Personal Pronouns	• *she, he, they, her, him, them, his, their, himself, herself, themselves* (plus contracted forms)
9. Pronoun *it*	• *it*
10. Demonstrative Pronouns	• *that/this/these/those* (excluding *that* as relative pronoun) followed by ▪ A verb (in any form); or ▪ Clause punctuation (".", "!", "?", ":", ";", "–"); or ▪ A tone-unit boundary (in spoken data); or ▪ *Who/whom/whose/which/and* • *that's* • *that* after a tone-unit boundary (in spoken data)
11. Indefinite Pronouns	• *anybody, anyone, anything, everybody, everyone, everything, nobody, none, nothing, nowhere, somebody, someone, something*
12. Pro-verb *do*	• Any form of the verb *do* except ▪ When followed by 0 or 1 adverb and a verb (i.e. when *do* is an auxiliary verb); or ▪ After a punctuation mark, a tone-unit boundary (in spoken data), or *who/whom/ whose/which* (i.e. when *do* is used in questions)

Appendix A (contd)

Syntactic Feature	Search Method
13. Direct *wh*-questions	● Any word string of *what/where/when/how/ whether/ why/whoever/whomever/whichever/ wherever/whenever/whatever/however* followed by an auxiliary verb (any form, but not in contraction) ■ After clause punctuation (".", "!", "?". ":", ";", "–") or a tone-unit boundary (in spoken data)
14. Nominalizations	● Any word ending in *-tion, -ment, -ness,* or *-ity* (plus plural forms)
15. Gerunds	● Any participle form serving a nominal function
16. Total Other Nouns	● Any noun in dictionary, excluding the ones counted in Feature 14 and Feature 15
17. Agentless Passives	● Any word string beginning with the auxiliary verb *be* (any form, excluding *be* and *'s*) and ending with a verb in past tense/past participle form ■ With 0, 1 or 2 adverb(s) in-between; or ■ With a noun/pronoun in-between (for inverted forms in questions)
18. *By*-passives	● The same search method as Feature 17 above, but with the whole word string followed by the word *by*
19. *Be* as Main Verb	● Any form of *be* (excluding *be* and *'s*) followed by ■ An article, a demonstrative (*this, that, these, those*), a quantifier (*each, all, every, many, much, few, several, some, any*), or a numeral; or ■ *My/our/your/his/their/its* (plus contracted forms); or ■ A preposition, an adjective, or an address title
20. Existential *there*	● Any word string beginning with *there* and ending with any form of *be* (excluding *be* and *'s*) ■ With 0 or 1 word in-between ● *there's*
21. *That* Verb Complements	● *that* after *and/nor/but/or/also*/a punctuation ■ When followed by an article, a demonstrative (*this, that, these, those*), a quantifier (*each, all, every, many, much, few, several, some, any*), a numeral, a pronoun, *there*, a plural noun, a proper noun, or an address title (i.e. *that*-clause in clause-initial positions) ● *that* after a verb counted in Features 55-58 ■ When not followed by a verb, *and*, clause punctuation (".", "!", "?". ":", ";", "–") or a tone-unit boundary (in spoken data) ● Any word string beginning with a verb counted in Features 55-58 and ending with *that* ■ With a preposition after the verb; and ■ With a noun before *that*; and ■ With any number of words (excluding noun) in-between the preposition and the noun ● *that* after a tone-unit boundary (in spoken data)
22. *That* Adjective Complements	● *that* after an adjective (with or without a tone-unit boundary in-between)
23. *Wh* Clauses	● *who/whom/whose/which/what/where/when/how/ whether/why/whoever/whomever/whichever/ wherever/whenever/whatever/however* after a verb counted in Features 55-57 ■ But not followed by an auxiliary verb (to exclude *wh* questions)

Appendix A (contd)

Syntactic Feature	Search Method
24. Infinitives	• Any word string beginning with *to* and ending with a verb in base form ▪ With 0 or 1 adverb in-between
25. Present Participial Clauses	• Any verb in present participle form after a punctuation mark or after a tone-unit boundary (in spoken data) ▪ When followed by a preposition, a pronoun, or an adverb; or ▪ When followed by an article, a demonstrative (*this, that, these, those*), a quantifier (*each, all, every, many, much, few, several, some, any*), or a numeral; or ▪ When followed by *who/whom/whose/ which/what/where/when/how/ whether/why/whoever/whomever/ whichever/ wherever/whenever/whatever/ however*
26. Past Participial Clauses	• Any verb in past tense/past participle form after a punctuation mark or after a tone-unit boundary (in spoken data) ▪ When followed by a preposition or an adverb
27. Past Participial *whiz* Deletion Relatives	• Any verb in past tense/past participle form after a noun or after a quantifier pronoun (*everybody, somebody, anybody, everyone, someone, anyone, everything, something, anything*) ▪ When followed by a preposition, an adverb or the verb *be* (any form, excluding *be* and *'s*)
28. Present Participial *whiz* Deletion Relatives	• Any verb in present participle form after a noun
29. *That* Relative Clauses on Subject Position	• Any word string beginning with a noun and ending with a verb (in any form) ▪ With *that* (or a tone unit boundary plus *that*) following the noun; and ▪ With 0 or 1 adverb in-between *that* and the verb
30. *That* Relative Clauses on Object Position	• *that* (or a tone unit boundary plus *that*) after a noun ▪ When followed by an article, a demonstrative (*this, that, these, those*), a quantifier (*each, all, every, many, much, few, several, some, any*), or a numeral; or ▪ When followed by a subjective pronoun (excluding *you*) or a possessive pronoun (excluding *her*); or ▪ When followed by an adjective, a plural noun, a proper noun, a possessive noun, or an address title
31. *Wh* Relatives on Subject Position	• Any word string that is not part of an indirect *wh* question and that begins with a noun and ends with a verb (in any form) ▪ With a *wh* pronoun (*who, whom, whose, which*) following the noun; and ▪ With 0 or 1 adverb in-between the *wh* pronoun and the verb
32. *Wh* Relatives on Object Position	• *wh* pronoun (*who, whom, whose, which*) after a noun ▪ When not followed by an adverb or a verb; and ▪ When the word string is not part of an indirect *wh* question
33. Pied-piping Relatives Clauses	• *wh* pronoun (*who, whom, whose, which*) after a preposition
34. Sentence Relatives	• *which* after comma or a tone-unit boundary (in spoken data)
35. Causative Adverbial Subordinators	• *because*

Appendix A (contd)

Syntactic Feature	Search Method
36. Concessive Adverbial Subordinators	• *although, though*
37. Conditional Adverbial Subordinators	• *if, unless*
38. Other Adverbial Subordinators	• *since, while, whilst, whereupon, whereas, whereby, inasmuch as, forasmuch as, insofar as, insomuch as, as long as, as soon as* • *so that, such that* ▪ when not followed by a noun or an adjective
39. Total Prepositional Phrases	• *against, amid, amidst, among, amongst, at, besides, between, by, despite, during, except, for, from, in, into, minus, notwithstanding, of, off, on, onto, opposite, out, per, plus, pro, re, than, through, throughout, thru, to, toward, towards, upon, versus, via, with, within, without*
40. Attributive Adjectives	• An adjective after an adjective; or • A noun after an adjective; or • Any adjective not counted in Feature 41
41. Predicative Adjectives	• An adjective after the verb *be* (any form, excluding *be* and *'s*) ▪ When not followed by an adjective, an adverb, or a noun; or ▪ When followed by an adverb, which is not followed by an adjective or a noun
42. Total Adverbs	• Any adverb in dictionary; or • Any word longer than 5 letters and ending in *–ly* • Excluding the ones counted in Feature 4, Feature 5, Feature 46, Feature 47 and Feature 48
43. Type/Token Ratio	• The number of different lexical items in the first 400 words of each text divided by 4
44. Mean Word Length	• Average length (in terms of number of letters) of all the words in a text
45. Conjuncts	• *alternatively, altogether, consequently, conversely, eg, e.g., else, furthermore, hence, however, i.e., instead, likewise, moreover, namely, nevertheless, nonetheless, notwithstanding, otherwise, rather, similarly, therefore, thus, viz.* • *in + comparison/contrast/particular/addition /conclusion/consequence/sum/summary/any event /any case/other words* • *for + example/instance* • *by + contrast/comparison* • *as a + result/consequence* • *on the + contrary/other hand* • *that is/else/altogether* ▪ When after a punctuation mark or a tone-unit boundary (in spoken data); and ▪ When followed by a comma or a tone-unit boundary (in spoken data) • *rather* ▪ When after a punctuation mark or a tone-unit boundary (in spoken data); and ▪ When followed by a comma or a tone-unit boundary (in spoken data) or a word that is not an adjective or an adverb
46. Downtoners	• *almost, barely, hardly, merely, mildly, nearly, only, partially, partly, practically, scarcely, slightly, somewhat*

Appendix A (contd)

Syntactic Feature	Search Method
47. Hedges	● *at about, something like, more or less, almost, maybe* ● *sort of, kind of* ■ When not after an article, a demonstrative (*this, that, these, those*), a quantifier (*each, all, every, many, much, few, several, some, any*), or a numeral ■ When not after an adjective or a possessive pronoun (*my, our, your, his, their, its*, plus contracted forms) ■ When not after *what/where/ when/how/ whether/why/whoever/ whomever/whichever/ wherever/whenever/ whatever/however*
48. Amplifiers	● *absolutely, altogether, completely, enormously, entirely, extremely, fully , greatly, highly, intensely, perfectly, strongly, thoroughly, totally, utterly, very*
49. Emphatics	● *for sure, a lot, such a, just, really, most, more* ● *real, so* ■ when followed by an adjective ● When the verb *do* is followed by a verb
50. Discourse Particles	● *well, now, anyway, anyhow, anyways* ■ When after a clause punctuation (".", "!", "?". ":", ";", "–") or a tone-unit boundary (in spoken data)
51. Demonstratives	● *that, this, these, those* ■ Excluding the ones counted in Feature 10 and *that* as relative, complementizer, or subordinator
52. Possibility Modals	● *can, may, might, could* (plus contractions)
53. Necessity Modals	● *ought, should, must* (plus contractions)
54. Predictive Modals	● *will, would, shall* (plus contractions)
55. Public Verbs	● Any present and past tense form of ■ *acknowledge, admit, agree, assert, claim, complain, declare, deny, explain, hint, insist, mention, proclaim, promise, protest, remark, reply, report, say, suggest, swear, write*
56. Private Verbs	● Any present and past tense form of ■ *anticipate, assume, believe, conclude, decide, demonstrate, determine, discover, doubt, estimate, fear, feel, find, forget, guess, hear, hope, imagine, imply, indicate, infer, know, learn, mean, notice, prove, realize, recognize, remember, reveal, see, show, suppose, think, understand*
57. Suasive Verbs	● Any present and past tense form of ■ *agree, arrange, ask, beg, command, decide, demand, grant, insist, instruct, ordain, pledge, pronounce, propose, recommend, request, stipulate, suggest, urge*
58. *Seem* and *appear*	● Any present and past tense form of *seem* and *appear*

Appendix A (contd)

Syntactic Feature	Search Method
59. Contractions	● Any contractions on pronouns or auxiliary forms (negation) ● Any 's suffix on nouns when followed by ■ A verb or a preposition; or ■ An adverb plus a verb; or ■ An article, a demonstrative (*this, that, these, those*), a quantifier (*each, all, every, many, much, few, several, some, any*), or a numeral; or ■ A possessive pronoun (*my, our, your, his, their, its*, plus contracted forms); or ■ An adjective before a clause punctuation (".", "!", "?", ":", ";", "–") or a tone-unit boundary (in spoken data)
60. Subordinator *that* deletion	● Any verb listed under Feature 55, Feature 56 and Feature 57 when followed by ■ An optional tone-unit boundary plus a demonstrative pronoun (*this, that, these, those*) or a subjective pronoun (*I, we, he, she, they*, plus contracted forms); or ■ A pronoun/noun plus a verb; or ■ Any word string beginning with an adjective, an adverb, an article, a demonstrative (*this, that, these, those*), a quantifier (*each, all, every, many, much, few, several, some, any*), a numeral, or a possessive pronoun (*my, our, your, his, their, its*, plus contracted forms) and ending with a verb (in any form) ◆ With a noun before the verb; and ◆ With 0 or 1 adjective before the noun
61. Stranded Prepositions	● Any preposition followed by a punctuation a mark or a tone-unit boundary (in spoken data)
62. Split Infinitives	● Any word string beginning with *to* and ending with a verb in base form ■ With 1 or 2 adverb(s) in-between
63. Split Auxiliaries	● Any word string beginning with an auxiliary verb and ending with a verb in base form ■ With 1 or 2 adverb(s) in-between
64. Phrasal Coordination	● *and* linking ■ 2 adjectives; or ■ 2 adverbs; or ■ 2 verbs; or ■ 2 nouns
65. Independent Clause Coordination	● *and* after a comma or a tone-unit boundary (in spoken data) ■ When followed by a subjective pronoun, a demonstrative pronoun (*this, that, these, those*), *so*, *then*, or *there* plus any form of *be* (excluding *be* and 's) ● *and* after a clause punctuation (".", "!", "?", ":", ";", "–") ● *and* when followed by ■ *who/whom/whose/which/what/where/ when/how/ whether/why/whoever/ whomever/whichever/ wherever/whenever whatever/however*; or ■ an adverbial subordinator (listed in Features 35-8), a discourse particle (listed in Feature 50), or a conjunct (listed in Feature 45)
66. Synthetic Negation	● *no* followed by a quantifier (*each, all, every, many, much, few, several, some, any*), an adjective, or a noun ■ Excluding *no* as a response ● *neither, nor*
67. Analytic Negation	● *not* (plus contracted forms)

Appendix B Letter and Consent Form for Teachers

Dear English Teachers,

I am a year-two PhD research student in the English Department of the Chinese University of Hong Kong and I am writing to seek your support in a research project.

The purpose of my study is to examine the features of Hong Kong students' spoken English. To achieve this aim, I need samples of oral English from Form 6 and Form 7 secondary school students. I would be grateful if you can consent to my data collection in your school.

My research procedures include:
(1) tape-recording your students' speech in your school oral examination;
(2) asking your students to fill in a short questionnaire which asks for information about their age, sex and grades in the English subject in HKCEE.

This is an anonymous study. The students' information will be kept confidential and the name of your school will not be revealed. The data collected will be used for research purpose only. Excerpts from the students' speech may appear in presentations or publications related to this research project.

This study will help English teachers in Hong Kong better understand their students' standard and needs. So your support is extremely important and your cooperation would be much appreciated.

If you consent to my data collection in your school, please complete the attached consent form. Should you have any enquiries, please feel free to contact me (amychui@cuhk.edu.hk).

Thank you very much for your cooperation!

Yours faithfully,

Amy Chui.

CONSENT FORM

I understand that Ms Amy Chui is conducting a research project about Hong Kong students' spoken English and I have read the information outlined in the previous page.

* *Please tick the appropriate box.*

☐ I consent to her data collection in my class/ school.

☐ I do not consent to her data collection in my class/ school.

Signature: _____ Date: _____

THANK YOU!

Appendix C Letter and Consent Form for Students

Dear Students,

I am a research student in the Chinese University of Hong Kong and I am writing to invite your participation in a research project.

The purpose of my study is to examine the features of Hong Kong students' spoken English. To achieve this aim, I need samples of oral English from Form 6 and Form 7 secondary school students.

If you agree to participate, what you will do is to:
(3) let me tape record your speech in your school oral examination;
(4) fill in a short questionnaire which asks for information about your age, sex and grades in the English subject in HKCEE.

Your information will be kept confidential and your name will not be revealed. The data collected will be used for research purpose only. Excerpts from your speech may appear in presentations or publications related to this research project.

Your participation is totally voluntary. There is no penalty for refusing to participate. As a participant, you may withdraw at any time without penalty.

This study will help teachers better understand the standard and the needs of Hong Kong students. So your participation is extremely important and your cooperation would be much appreciated.

If you agree to participate in this project, please complete the attached consent form and questionnaire. Should you have any enquiries, please feel free to contact me (amychui@cuhk.edu.hk).

Thank you very much for your cooperation!

Yours faithfully,

Amy Chui.

CONSENT FORM

I understand that Ms Amy Chui is conducting a research project about Hong Kong students' spoken English. I understand that participation in this project is voluntary. I also understand that my data will be kept confidential and used for research purpose.

I agree to participate in this research project. I give my permission to Ms Chui to tape record my speech and use my data for research purpose.

Signature: _____ Date: _____

QUESTIONNAIRE

Sex: M / F * *Please circle as appropriate.*

Age: _____

HKCEE results:

English Language (overall grade): _____

Paper 3 Speaking (sub-grade): _____

In which year did you take HKCEE and obtain the above results? _____

** For students taking HKCEE in 2006 or the years before:*

Which syllabus did you choose? Syllabus A / Syllabus B

THANK YOU!

(Chinese Version)

各位同學，你好！

本人是香港中文大學的研究生，現正進行一項關於香港中學生英語說話能力的研究計劃，需要收集中六、中七學生的英語樣本作爲研究資料，希望同學能參加此項計劃。

同學只需讓本人在同學進行校內英語會話考試時錄音，然後再填寫一份簡單問卷 (問卷內容包括你的年齡、性別及中學會考英文科的成績)。

本研究以不記名方式進行，同學所提供的個人資料將會被保密。所有收集所得的資料只會被用作學術研究用途，而語言資料有可能在與此計劃有關的報告中被引用。同學的參與全屬自願性質，絕對不會影響同學的考試成績，而中途退出亦不會有任何處分。

本人希望透過是項研究讓老師們更清楚了解香港學生的英語水平及學習需要，所以同學的參與、合作是十分重要的。

如同學同意參加此項計劃，請填妥附上的同意書及問卷。如有任何查詢，可電郵至 amychui@cuhk.edu.hk。

多謝合作！

同意書

本人已知悉有關香港中學生英語說話能力研究計劃的詳情，並明白是次參與全屬自願性質，而本人的資料將會被保密及被用作學術研究用途。

本人同意參加此項研究計劃。本人同意被錄音，亦同意資料被用作學術研究用途。

簽名: _____ 日期: _____

問卷

* *請圈示*

性別： 男 / 女

年齡：_____

香港中學會考成績：

英文科 (總分等級)：_____

卷三 說話能力(等級)：_____

你是在哪一年參加香港中學會考，並獲得以上成績 ？ _____

* *在 2006 年或以前參加香港中學會考的學生，*

請說明應考的課程： 課程甲 / 課程乙

多謝合作！

Appendix D Titles of 16 Articles Used for Oral Presentations (Learner Speech Data)

1.	Musical history	(2007 past paper, Set 16 Candidate D)
2.	Opera for the elderly	(2007 past paper, Set 16 Candidate C)
3.	The power of the brush	(2007 past paper, Set 16 Candidate B)
4.	Traditional religions	(2007 past paper, Set 16 Candidate A)
5.	Pollution solutions	(2007 past paper, Set 15 Candidate D)
6.	Dark day for Hong Kong	(2007 past paper, Set 15 Candidate C)
7.	Children's cry for clean air	(2007 past paper, Set 15 Candidate B)
8.	Action Blue Sky	(2007 past paper, Set 15 Candidate A)
9.	Healthy options	(2007 past paper, Set 14 Candidate D)
10.	No meat, thank you!	(2007 past paper, Set 14 Candidate C)
11.	Food made to order	(2007 past paper, Set 14 Candidate B)
12.	No chemical allowed	(2007 past paper, Set 14 Candidate A)
13.	Extraordinary exercise	(2007 past paper, Set 13 Candidate D)
14.	Dangerous games	(2007 past paper, Set 13 Candidate C)
15.	Gambling with childhood	(2007 past paper, Set 13 Candidate B)
16.	Sporting salvation	(2007 past paper, Set 13 Candidate A)

* The articles were all taken from

Hong Kong Examinations and Assessment Authority. (2007). *HKALE. AS use of English 2007*. Hong Kong: Hong Kong Examinations and Assessment Authority.

Appendix E Essay Topics Used in Learner Writing Data

1. You are taking part in an essay competition organized by the United Nations. The title of the essay is 'World peace and international understanding in the 21st century'. Write the essay. State briefly why world peace is important, and suggest three ways in which young people can help to promote international understanding. (1999 HKALE, Question 1)

2. Your English teacher has explained to your class that learning to construct a strong argument against a popular belief is a good way of developing your argumentative writing skills. Since most people believe that a university education is essential for a successful life, your teacher has asked you to write an essay entitled 'A university education is not essential for success'. Write the essay, developing three or four good arguments that strongly support the statement. (1999 HKALE, Question 2)

3. 'The mobile telephone is the most annoying, unnecessary and time-wasting device ever invented'. Write an essay in which you strongly support this statement. Provide at least three reasons in support of your argument. Give your essay a title. (2000 HKALE, Question 2)

4. Christmas celebrations have now become very popular in China. In many large cities, shopping malls are full of tall Christmas trees and men dressed as Santa Claus. Write an essay EITHER for OR against the proposition that 'Christmas should not be celebrated in China because it is a foreign custom'. Give reasons for your views. (2001 HKALE, Question 2)

5. In many parts of the world, it is very common for couples to live together before they decide to get married. Write an essay discussing BOTH the positive and negative aspects of living together before marriage. Give your own opinion about this issue. (2001 HKALE, Question 3)

* The questions were all taken from

Hong Kong Examinations Authority. (1999). *HKALE question papers: AS use of English 1999*. Hong Kong: Hong Kong Examinations Authority.
Hong Kong Examinations Authority. (2000). *HKALE question papers: AS use of English 2000*. Hong Kong: Hong Kong Examinations Authority.
Hong Kong Examinations Authority. (2001). *HKALE question papers: AS use of English 2001*. Hong Kong: Hong Kong Examinations Authority.

Appendix F Essay Topics Used in Native-speaker Writing Data (LOCNESS essays)

Topics	No. of essays
Transport	8
Boxing	8
National Lottery	8
Computers and the human brain	7
BSE* and beef	7
In vitro fertilization - genetic engineering	9
Fox hunting	5
TOTAL	**52**

*BSE = bovine spongiform encephalopathy

Appendix G Adjustments in Search Methods

Syntactic Feature	Search Method in the Present Study	Reason for the Change
21. *That* Verb Complements	• *that* after any form of the verbs listed in Features 55-58 (i.e. public verbs, private verbs, suasive verbs, *seem* and *appear*) • Any sentence containing verbs that are listed in Features 55-58 and that are not counted in the above method and not counted in Feature 23 and Feature 26 is checked manually to see if there is any other *that* complement not following the verb directly	The original search method in Biber (1988) cannot be carried out because it allows an unspecified number of words in one of the search strings.
30. *That* Relative Clauses on Object Position	• Relative pronoun *that* after a noun ▪ When followed by an article, a demonstrative (*this, that, these, those*), a quantifier (*each, all, every, many, much, few, several, some, any*), or a numeral; or ▪ When followed by a subjective pronoun (excluding *you*), a possessive pronoun (excluding *her*); or ▪ When followed by an adjective or a noun	(The change made has been highlighted in the left box.) The original search method in Biber (1988) is to use plural nouns, proper nouns, possessive nouns and address titles, some of which are not clearly defined.
59. Contractions	• Any contractions on pronouns, auxiliary verbs and verbs (including negation)	The search method in Biber (1988) tries to distinguish *'s* as contractions from *'s* possessive suffix by using a complex algorithm. Since the ICE tagset contains simple features for contractions, they are used instead.
61. Stranded Prepositions	• Any preposition not followed by ▪ A noun/pronoun ▪ An article/adjective plus a noun ▪ An article plus an adjective plus a noun	The original search method in Biber (1988) makes use of tone-unit boundaries to locate stranded prepositions in spoken data. But my spoken data is not marked for such boundaries.
63. Split Auxiliaries	• Any word string beginning with an auxiliary verb and ending with a verb in any form ▪ With 1 or 2 adverb(s) in-between	(The change made has been highlighted in the left box.) This original search method in Biber (1988) is to use *VB*, a verb in base form. But for split auxiliaries, the main verbs can be in base form (e.g. *does seriously affect*), in *–ing* form (e.g. *is seriously affecting*), or in past participial form (e.g. *has seriously affected*). The example given in Biber (1988) is also in past participial form. Therefore, the original *VB* in Biber (1988) should be a typo.

Appendix H Differences in Mean Normalised Frequency (M Diff.) and Differences in Standard Deviation (SD Diff) Between (a) Learner Speech Data and Biber's (1988) Spontaneous Speeches (p. 268); (b) Learner Speech Data and Biber's (1988) Prepared Speeches (p. 269)

Syntactic Feature	Comparison with Spontaneous Speech		Comparison with Prepared Speech	
	M Diff.	SD Diff.	M Diff.	SD Diff.
1. Past Tense	-48.2	-29.5	-32.6	-11.7
2. Perfect Aspect	-4.9	+0.3	-8.6	-5.5
3. Present Tense	+21.7	+5.9	+31.6	+1.1
4. Place Adverbials	-0.9	+1 1	-1.0	+0.6
5. Time Adverbials	-2.3	+0.9	-4.3	-0.8
6. First Person Pronouns	-21.2	-4.1	-2.6	+4.4
7. Second Person Pronouns	-6.2	-2.0	+3.3	+5.6
8. Third Person Pronouns	-2.4	+3.1	-7.6	+6.9
9. Pronoun *it*	+3.3	+9.1	+8.0	+11.0
10. Demonstrative Pronouns	-4.2	+2.1	-2.0	+3.4
11. Indefinite Pronouns	+2.3	+3.9	+2.4	+4.4
12. Pro-verb *do*	-4.0	-2.6	-2.0	-1.0
13. Direct *wh*-questions	-0.5	+0.5	+0.2	+1.1
14. Nominalizations	+0.5	+2.6	-1.9	+5.2
15. Gerunds	-1.5	+2.0	-2.3	+3.1
16. Total Other Nouns	+31.8	+15.6	+0.4	+10.8
17. Agentless Passives	-2.7	+0.2	-6.1	+0.5
18. *By*-passives	0.0	+0.4	-0.1	+0.3
19. *Be* as Main Verb	-20.4	+1.9	-15.7	+4.0
20. Existential *there*	+1.6	+2.6	+2.1	+4.3
21. *That* Verb Complements	+2.6	+4.0	+2.7	+4.7
22. *That* Adjective Complements	-0.3	-0.2	-0.5	+0.1
23. *Wh* Clauses	-0.2	+1.5	+0.8	+2.6
24. Infinitives	+5.4	+4.8	+4.3	+5.4
25. Present Participial Clauses	-0.1	+0.3	-0.1	+0.1
26. Past Participial Clauses	+0.2	+1.3	+0.2	+1.3
27. Past Participial *whiz* Deletion Relatives	-0.3	+0.6	-0.4	+0.1
28. Present Participial *whiz* Deletion Relatives	-0.2	+0.2	-0.8	+0.6
29. *That* Relative Clauses on Subject Position	+0.5	+1.3	+0.4	+1.0
30. *That* Relative Clauses on Object Position	-1.4	-1.4	-1.6	-1.4
31. *Wh* Relatives on Subject Position	-2.7	+0.9	-0.6	+1.5
32. *Wh* Relatives on Object Position	-2.0	-1.3	-2.4	-1.0
33. Pied-piping Relatives Clauses	-0.5	-0.5	-1.0	-1.0

Appendix H (contd)

Syntactic Feature	Comparison with Spontaneous Speech		Comparison with Prepared Speech	
	M Diff.	SD Diff.	M Diff.	SD Diff.
34. Sentence Relatives	-0.2	+0.1	0.0	+0.4
35. Causative Adverbial Subordinators	+4.4	+4.4	+5.9	+5.0
36. Concessive Adverbial Subordinators	+0.6	+1.6	+0.6	+1.4
37. Conditional Adverbial Subordinators	-0.8	+1.3	-0.1	+2.5
38. Other Adverbial Subordinators	+0.4	+1.7	+0.4	+1.3
39. Total Prepositional Phrases	-36.1	-2.2	-54.1	+5.9
40. Attributive Adjectives	-1.5	+11.8	-6.2	+13.6
41. Predicative Adjectives	-0.8	+3.3	0.0	+3.8
42. Total Adverbs	-24.7	+6.2	-21.5	+7.7
43. Type/Token Ratio*	+7.8	+0.8	+3.7	+1.3
44. Mean Word Length#	+0.2	-0.1	0.0	0.0
45. Conjuncts	+2.9	+4.0	+2.8	+4.1
46. Downtoners	-0.8	+0.1	-0.6	+1.1
47. Hedges	-0.1	+0.3	+0.2	+0.7
48. Amplifiers	-0.9	+2.4	+1.1	+3.5
49. Emphatics	-4.6	+0.5	-3.6	-0.4
50. Discourse Particles	-3.6	-2.4	-2.4	-2.4
51. Demonstratives	-0.7	+5.5	-2.0	+5.3
52. Possibility Modals	+7.1	+9.4	+8.1	+9.4
53. Necessity Modals	+3.1	+4.3	+1.9	+3.0
54. Predictive Modals	-4.5	+3.3	-0.4	+6.4
55. Public Verbs	-9.2	-3.9	-3.1	+1.1
56. Private Verbs	-7.4	+1.9	-3.4	+3.0
57. Suasive Verbs	-2.0	+0.6	-2.6	-1.0
58. *Seem* and *appear*	+0.1	+0.7	0.0	+0.7
59. Contractions	-7.6	+0.5	-3.1	-0.4
60. Subordinator *that* deletion	-0.3	+2.1	+3.4	+4.3
61. Stranded Prepositions	-4.4	-2.8	-3.6	-1.4
62. Split Infinitives	+0.1	+0.7	+0.1	+0.7
63. Split Auxiliaries	-2.2	+0.8	-3.4	+0.6
64. Phrasal Coordination	+4.3	+3.9	+4.9	+4.6
65. Independent Clause Coordination	-5.4	-2.7	+0.9	+5.4
66. Synthetic Negation	-1.0	+0.7	-1.2	+0.6
67. Analytic Negation	+1.6	+1.3	+2.3	+3.9
Distance Index	**85.3**	**42.5**	**79.6**	**36.1**

* The type/token ratio was calculated by dividing the number of different words by the number of tokens and then multiplying the resultant fraction by 100. The numbers in the "M Diff." columns refer to the differences between the two type/token ratios concerned.

The mean word length refers to the average number of orthographic letters in a word.

Appendix I Mean Normalised Frequencies (M) and Standard Deviations (SD) of the 67 Syntactic Features in Biber's (1988) Prepared Speeches (p. 269)

Category	Syntactic Feature	M	SD
Tense and aspect markers	1. Past Tense	48.3	22.5
	2. Perfect Aspect	11.3	9.1
	3. Present Tense	70.5	21.1
Place and time adverbials	4. Place Adverbials	1.9	2.1
	5. Time Adverbials	7.1	4.4
Pronouns and pro-verbs	6. First Person Pronouns	41.8	21.4
	7. Second Person Pronouns	5.1	4.9
	8. Third Person Personal Pronouns	37.1	15.3
	9. Pronoun *it*	8.9	4.4
	10. Demonstrative Pronouns	6.9	2.1
	11. Indefinite Pronouns	1.5	1.2
	12. Pro-verb *do*	2.4	2.3
Questions	13. Direct *wh*-questions	0.3	0.6
Nominal forms	14. Nominalizations	20.6	11.5
	15. Gerunds	5.1	1.9
	16. Total Other Nouns	189.1	21.6
Passives	17. Agentless Passives	9.6	3.9
	18. *By*-passives	0.2	0.4
Stative forms	19. *Be* as Main Verb	30.5	5.8
	20. Existential *there*	3.1	1.4
Subordination features	21. *That* Verb Complements	7.0	4.5
	22. *That* Adjective Complements	0.6	0.7
	23. *Wh* Clauses	0.2	0.4
	24. Infinitives	16.2	6.6
	25. Present Participial Clauses	0.2	0.6
	26. Past Participial Clauses	0.0	0.0
	27. Past Participial *whiz* Deletion Relatives	0.9	1.4
	28. Present Participial *whiz* Deletion Relatives	1.4	1.2
	29. *That* Relative Clauses on Subject Position	0.3	0.7
	30. *That* Relative Clauses on Object Position	1.6	1.4
	31. *Wh* Relatives on Subject Position	2.4	2.3
	32. *Wh* Relatives on Object Position	2.5	1.8
	33. Pied-piping Relatives Clauses	1.1	1.6

Appendix I (contd)

Category	Syntactic Feature	M	SD
Subordination features (contd)	34. Sentence Relatives	0.1	0.3
	35. Causative Adverbial Subordinators	1.6	1.7
	36. Concessive Adverbial Subordinators	0.1	0.4
	37. Conditional Adverbial Subordinators	2.4	1.1
	38. Other Adverbial Subordinators	0.8	1.4
Prepositional phrases, adjectives, and adverbs	39. Total Prepositional Phrases	112.6	12.5
	40. Attributive Adjectives	48.9	13.6
	41. Predicative Adjectives	3.6	1.6
	42. Total Adverbs	62.2	9.8
Lexical specificity	43. Type/Token Ratio*	49.0	3.4
	44. Mean Word Length#	4.4	0.2
Lexical classes	45. Conjuncts	0.5	0.7
	46. Downtoners	1.5	0.9
	47. Hedges	0.2	0.6
	48. Amplifiers	3.1	2.1
	49. Emphatics	4.8	3.2
	50. Discourse Particles	2.4	2.4
	51. Demonstratives	12.9	4.2
Modals	52. Possibility Modals	5.6	3.1
	53. Necessity Modals	2.6	2.9
	54. Predictive Modals	5.0	3.0
Specialized verb classes	55. Public Verbs	7.9	4.8
	56. Private Verbs	17.6	4.9
	57. Suasive Verbs	3.3	3.5
	58. *Seem* and *appear*	0.5	0.7
Reduced forms and dispreferred structures	59. Contractions	13.3	10.4
	60. Subordinator *that* deletion	1.9	1.6
	61. Stranded Prepositions	3.7	1.9
	62. Split Infinitives	0.0	0.0
	63. Split Auxiliaries	5.3	2.9
Coordination	64. Phrasal Coordination	1.1	1.2
	65. Independent Clause Coordination	8.6	3.4
Negation	66. Synthetic Negation	1.8	1.4
	67. Analytic Negation	8.4	4.3

* The type/token ratio, calculated by dividing the number of different words by the number of tokens and then multiplying the resultant fraction by 100, refers to the average number of different words per 100 tokens.

The mean word length refers to the average number of orthographic letters in a word.

Appendix J Mean Dimension Scores and Standard Deviations of Biber's (1988) Prepared Speeches (p. 125)

Dimension	Mean Score	Standard Deviation
Dimension 1	2.2	6.7
Dimension 2	0.7	3.3
Dimension 3	0.3	3.6
Dimension 4	0.4	4.1
Dimension 5	-1.9	1.4
Dimension 6	3.4	2.8

Appendix K Factorial Structure of Biber's (1988) Six Textual Dimensions* (pp. 102-103)

Table K1

Factorial Structure of Dimension 1 Involved Versus Informational Production

Positive Features		Negative Features	
Private verbs	.96	Nouns	-.80
That deletion	.91	Word length	-.58
Contractions	.90	Prepositions	-.54
Present tense verbs	.86	Type/token ratio	-.54
Second person pronouns	.86	Attributive adjectives	-.47
Do as pro-verb	.82		
Analytic negation	.78		
Demonstrative pronouns	.76		
General emphatics	.74		
First person pronouns	.74		
Pronoun *it*	.71		
Be as main verb	.71		
Causative subordination	.66		
Discourse particles	.66		
Indefinite pronouns	.62		
General hedges	.58		
Amplifiers	.56		
Sentence relatives	.55		
Wh questions	.52		
Possibility modals	.50		
Clause coordination	.48		
Wh clauses	.47		
Final prepositions	.43		

*** Features not used in the calculation on dimension scores are not included.**

Table K2

Factorial Structure of Dimension 2 Narrative Versus Non-narrative Concern

Positive Features		Negative Features
Past tense verbs	.90	-- no negative features --
Third person personal pronouns	.73	
Perfect aspect verbs	.48	
Public verbs	.43	
Synthetic negation	.40	
Present participial clauses	.39	

Table K3

Factorial Structure of Dimension 3 Explicit Versus Situation-dependent Reference

Positive Features		Negative Features	
Wh relative clauses on object positions	.63	Time adverbials	-.60
Pied piping constructions	.61	Place adverbials	-.49
Wh relative clauses on subject positions	.45	Adverbs	-.46
Phrasal coordination	.36		
Nominalizations	.36		

Table K4

Factorial Structure of Dimension 4 Overt Expression of Persuasion

Positive Features		Negative Features
Infinitives	.76	-- no negative features --
Predictive modals	.54	
Suasive verbs	.49	
Conditional subordination	.47	
Necessity modals	.46	
Split auxiliaries	.44	

Table K5

Factorial Structure of Dimension 5 Abstract Versus Non-abstract Information

Positive Features		**Negative Features**
Conjuncts	.48	-- no negative features --
Agentless passives	.43	
Past participial clauses	.42	
By-passives	.41	
Past participial *whiz* deletions	.40	
Other adverbial subordinators	.39	

Table K6

Factorial Structure of Dimension 6 On-line Informational Elaboration

Positive Features		**Negative Features**
That clauses as verb complements	.56.	-- no negative features --
Demonstratives	.55	
That relative clauses on object positions	.46	
That clauses as adj. complements	.36	