Intercultural Communication and Adolescent Learners: A Corpus-based Approach to Online and Face-to-face Interaction

YEN-LIANG LIN

Thesis submitted to the University of Nottingham

for the Degree of Doctor of Philosophy

AUGUST 2013

Abstract

This study reports on a corpus analysis of samples of online and face-to-face intercultural communication among a group of British and Taiwanese adolescents, with the aim of exploring the particular lexical, grammatical and discourse features of the online and spoken discourse from three perspectives: a keyness approach, a discourse analytical perspective and a multi-word sequence perspective. Keyness approach brings together three levels of keyness analysis: keywords, semantic domains and parts-of-speech, and further highlights those linguistic features that deserve particular attention. Furthermore, a discourse analytical approach adds greater detail and depth of description of language patterning by examining the particular linguistic features in context. Such findings that pertain to discourse and pragmatic functions in context are not likely to be made when only keyness is examined. The third approach of this thesis focuses on recurrent multi-word sequences, paying particular attention to their discourse functions in online and spoken settings. It is evident that multi-word sequences often perform systematic discourse functions, even though they do not usually constitute complete grammatical or idiomatic structures. The approach also examines the developmental perspectives of multi-word sequences, showing that intercultural contact with native speakers of English fosters the longitudinal development of the use of sequences by the Taiwanese learners. The method here, which focuses on naturally occurring language output, diminishes the effects of the artificial contexts often created in language testing settings. In light of the potential significance of the research to EFL pedagogy, the thesis further reports on the extent to which EFL textbooks used in Taiwan represent the particular linguistic features identified in authentic intercultural communication. The research findings demonstrate the

pedagogical merit of the analyses of the three perspectives and thus help in the design of courses for adolescent intercultural interaction in both online and face-to-face settings.

Keywords: intercultural communication, corpus-based approach, computer-mediated communication, face-to-face interaction, adolescent learners

Parts of this thesis have been published as journal articles / book chapters / conference papers

Exploring recurrent multi-word sequences in EFL textbook dialogues and authentic discourse. *Language Teaching & Learning* (accepted for publication).

Vague language and interpersonal communication: An analysis of adolescent intercultural conversation. *International Journal of Society, Culture & Language* (accepted for publication).

Discourse functions of recurrent multi-word sequences in online and face-to-face intercultural communication. In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics* (pp. 105-129). London: Springer.

Lexical features of adolescent online intercultural communication. In M. Dąbrowska, J. Lesniewska, & B. Piąte (Eds.), *Languages, Literatures and Cultures in Contact: English and American Studies in the Age of Global Communication* (pp. 217-232). Krakow: Tertium.

Mind the gap! Textbook conversation vs. authentic intercultural interaction. In Y. Leung (Ed.), *Selected Papers from the 21st International Symposium on English Teaching* (pp. 42-54). Taipei: Crane Publishing. (Best Paper Award)

Analysing intercultural discourse in online and spoken communication: A Keyness Approach. Paper presented at the sixth Inter-Varietal Applied Corpus Studies (IVACS) conference, Leeds, UK.

Formulaic language in EFL textbook conversation and naturally occurring communication. Paper presented at the 2012 AAAL Conference, Boston, USA.

Intercultural Communication and Young Learners: an analysis of online and face-to-face interaction. Paper presented at the 44th BAAL Annual Meeting, Bristol, UK.

Keywords, Parts-of-speech and Semantic Domains in Online Intercultural Communication. Paper presented at the 14th Warwick International Postgraduate Conference, Warwick, UK.

Intercultural Exchange on Electronic Discussion Boards: the Development of Formulaic Sequences by Taiwanese Junior High School Students. Paper presented at 2011 International Conference on EFL Education, Changhua, Taiwan.

The Lexical Features of Adolescent Online Intercultural Communication. Paper presented at the Twelfth International Conference on English and American Studies, Krakow, Poland.

Acknowledgments

Undertaking this PhD has been a truly life-changing experience for me and it would not have been possible to complete this thesis without the great support that I received from many people. Most importantly, I would like to express my profound appreciation to my supervisors Svenja Adolphs and Louise Mullany for all their support and guidance throughout this process. Without their encouragement and ongoing feedback this PhD would not have been achievable.

My heartfelt gratitude is also expressed to the British Council who proposed the "Connecting Classrooms Project" that connects the adolescents from Hualien (Taiwan) and Cumbria. Deep appreciation particularly goes to all of the participants who took part in this project and kindly agreed for their online and spoken data to be used in this thesis. Without their participation this study would not have been possible. Additionally, many thanks to the teachers in Cumbria, Alison Philips and Moira Houlan, who serve as the co-ordinators in the intercultural exchange programme.

I would also like to thank the academic and administrative staff in the School of English at the University of Nottingham for their support and patience over the years. Special thanks go to Zoltán Dörnyei, Norbert Schmitt, Kevin Harvey, Michaela Mahlberg and Rachele De Felice who allowed me to attend their lectures. I am also very grateful to my friends and research colleagues who have been so supportive over the years: Anne Liu, Wenjuan Yuan, Wipapan Ngampramuan, Letty Chan, Pawel Szudarski, Daniel Hunt and Klaudia Lee deserve special mention. Finally, I owe a debt of gratitude to my family for encouraging me and providing financial support for me to study abroad.

Table of Contents

List of illustrations

CHAPTER 1

Introduction

1.1 Background and motivation

The emergence of English as a global language, technological developments and growing demand for intercultural interaction are making people around the world increasingly interconnected, and these also challenge the traditional contexts of English as a foreign language (EFL) teaching and learning. It is therefore imperative for EFL learners to become successful users of English (SUEs) (Prodromou, 2005) so that they can communicate effectively and appropriately with people who have a different cultural or linguistic background (Byram, 1997, 2012; Jackson, 2012; Spencer-Oatey & Franklin, 2009). Throughout my teaching career in Taiwan, I have been increasingly aware of such an impact of globalisation on foreign language education. In 2008 the British Council proposed an intercultural exchange project, *Connecting Classrooms*, with the primary aim of creating global partnerships between clusters of schools in the UK and others around the world, particularly in Asia. I have been involved in this intercultural project since the outset of the programme, when I was a junior high school[1] teacher in Hualien (Taiwan), and I also served as the main corresponding person in the partnership project between Hualien and Cumbria in the UK, helping to organise both online and face-to-face intercultural exchanges among teachers, students and Council officers from the two countries.

[1] Junior high schools are levels of schooling between elementary and high schools. In Taiwan a junior high school includes grades 7 through 9, consisting of students from ages 12 to 15.

1.1.1 The intercultural dimension in foreign language teaching

Issues such as the intercultural dimension of education and the rise of the Internet as a tool for intercultural communication have been increasingly emphasised as being important for foreign language teaching by governments worldwide. In the UK, for example, the Department for Education (formerly the Department for Education and Skills [DfES]) published *Putting the World into World-Class Education* (DfES, 2004, 2007) and *Developing a Global Dimension in the School Curriculum* (DfES, 2005), stressing the need for pupils to improve their knowledge and understanding of other cultures to prepare them for life and work as global citizens. This means equipping young people and adults with advanced language skills, with the means to communicate with others across the world and to understand cultures other than their own. The Association of International Educators (formerly the National Association for Foreign Student Affairs [NAFSA]) of the United States also proposed *An International Education Policy*, revealing the importance of cultural and foreign language study in primary and junior high schools and encouraging international partnerships that would facilitate internationalised curricula (NAFSA, 2007).

In addition, the *White Paper on Intercultural Dialogue* published by the Council of European Ministers of Foreign Affairs identifies intercultural dialogue as "a means of promoting awareness, understanding, reconciliation and tolerance, as well as preventing conflicts and ensuring integration and the cohesion of society" (Council of Europe, 2008, p. 8). The *Common European Framework of Reference for Languages* (CEFR; Council of Europe, 2001) also points out the importance of "interculturality" for language learners, which enables the individual to "develop

an enriched, more complex personality and an enhanced capacity for further language learning and greater openness to new cultural experiences" (p. 44). With regard to the teaching context in Taiwan, the recent *White Paper on International Education for Primary and Junior High Schools* published by the Ministry of Education in 2011 puts forward the same argument. Furthermore, the latest *General Guidelines of Grades 1-9 Curriculum* published by the Ministry of Education (2010) indicate that the ability to properly communicate with people of different linguistic and cultural backgrounds is an indicator of one of the core competencies that junior high school students should acquire. In this regard, English language teaching and learning in Taiwan are based on a view of language as communication and intercultural understanding.

This notwithstanding, research has reported that EFL teaching and learning in Taiwan are still restricted to classroom drills (Lu, 2003; Lin, 2009). As a result, learners have relatively limited opportunities to apply what they learn in the classroom in order to communicate socially with English-speaking people and, in turn, to develop intercultural competence in such interactions. In addition, this intercultural dimension does not occupy an important place in the competency indicators of the national EFL curriculum, although intercultural learning and understanding have been cited as among the most important educational goals in the General Guidelines (Ministry of Education, 2010). Only four out of the 86 competency indicators are related to culture understanding and learning, whereas 63 indicators are related to linguistic competence and the remaining 19 indicators pertain to language learning motivation and methods (Lin, 2009). In this regard, intercultural learning is not an area upon which focus is concentrated in Taiwan's EFL teaching practices.

In addition, a number of studies based in Taiwan (Chang, 2010; Lin, 2009; Tsi, 2002) have reported an inadequacy in the cultural and interpersonal components included in the EFL instruction and textbooks. That is, the teaching/learning materials used in the EFL classrooms are generally focused on traditional written grammar and vocabulary learning, in which most of the teachers spend most of the teaching time on these aspects; culture instruction and pragmatic functions of language, which help learners become aware of the appropriate language use in different contexts, are therefore often neglected, and consequently EFL learners might not acquire sufficient knowledge and skills for real-life communication in order to create and sustain successful interactions and build good relationships among interlocutors. Tsi (2002) investigated 154 secondary school EFL teachers' opinions with regard to culture instruction and authentic materials in their classrooms in Taiwan and uncovered a number of problems that hindered the teaching of culture. These include the inadequacy of inter- and cross-cultural components in the English course books and teacher's manuals, the difficulty of obtaining authentic and cultural materials and resources, the constraints of instructional time and the insufficient knowledge of teachers with respect to different cultures and pragmatic language use.

1.1.2 Previous research on intercultural exchange projects

In light of the importance of the intercultural dimension in language education, intercultural exchange and communication via Internet-based *Computer-Mediated Communication* (CMC) has been promoted to foreign language teachers as a means of providing pupils with opportunities to communicate with people from different countries (Dooly & O'Dowd, 2012; Kabata & Edasawa, 2011; Liaw & Master, 2010; Sasaki, 2010; Stickler & Emke, 2011; Warschauer & Kern, 2000).

One such example is the Connecting Classrooms Project, administered by the British Council. It aims to create global partnerships between clusters of schools in the UK and others around the world. In this project the participants join the online community and work with their international peers on collaborative curriculum projects, which enable them to interact across geographical boundaries to enhance their understanding of each other's societies, languages and cultures (British Council, n.d.).

In this regard, intercultural exchange provides learners with authentic input and opportunities to participate in the target social and cultural contexts. This type of learning derives in part from the socio-cultural perspective proposed by Vygotsky (1962), which illuminates the role of social interaction in creating an environment to learn language, learn about language and learn through language (Warschauer, 1997, p. 471). Language, in this approach, plays an important role in successful human learning and interaction. In addition, culture is regarded as inseparable from language; the two are intricately interwoven such that one shapes the other (Brown, 2007). Research has shown that online intercultural exchange is beneficial for both language and culture learning since online exposure offers language immersion in an authentic socio-cultural context in which the language is used, and students are given ample chance to encounter a considerable amount of authentic language through being able to read messages and respond to their audience (Liaw & Master, 2010; Montero, Watts, & Garcia-Carbonell, 2007; O'Dowd, 2007; Spencer-Oatey & Franklin, 2009; Warschauer & Kern, 2000).

Previous studies of intercultural exchange projects have highlighted the positive outcome in raising cultural awareness and understanding (Chang, 2010; Liaw,

2007; Liaw & Master, 2010; O'Dowd, 2007; Stickler & Emke, 2011), English learning motivation and engagement (Chang, 2010; Lu, 2003; Yang, 2011) and English language skills development (Jeng, 2010; Kabata & Edasawa, 2011; Liaw, 2007; Lu, 2003; Sasaki, 2010). The CANDLE project, for example, conducted by Liaw (2007) and Liaw and Master (2010) aimed to establish an innovative web-based environment to support students at a tertiary level to develop linguistic and intercultural competences. In this project EFL students in Taiwan read the culture-related articles offered by the researchers and then shared their responses to the articles with their English-speaking partners from the USA via online forums. The analysis of the students' forum entries found increases in the length and complexity of sentences in their writing and a reduction in grammatical errors. Additionally, the content analysis of the forum entries revealed different types of intercultural competences.

Kabata and Edasawa's (2011) key-pal project conducted between Japanese and Canadian university students indicated that students have ample opportunities for engaging in different aspects of language learning, including vocabulary, grammar and phrase/sentential expressions. The DfES (2004) also cited an intercultural exchange project involving early secondary school students in Slough, 90% of whose pupils are of Asian ethnic origin, which had initiated a link with a similar project in Delhi. It was found that through developing close links on a one-to-one basis via Internet and e-mail, pupils and teachers gained a global perspective both in subjects and in addressing moral issues (p. 12). Moreover, Sasaki's (2010) study on Japanese university students' EFL vocabulary development through e-mail interactions with a native English speaker found that students' repeated encounters and productive opportunities to use the target words played a vital role

in their vocabulary development. The communicative needs of online interaction also facilitated the learners' attempts to study new words by referring to all the resources available to them (e.g., dictionaries, classmates and teachers). In this regard, the use of CMC provides an authentic audience and opportunities to join in authentic language and cultural practice in the target foreign language, taking students beyond classroom cultures and learner-to-learner communication (Dooly & O'Dowd, 2012; Hanna & de Nooy, 2009; Montero et al., 2007).

1.2 Rationale

While there is a burgeoning field of research looking at communicative competence in intercultural settings, little is known about the language use in adolescent learners of English who take part in both online and face-to-face intercultural exchanges. This present study investigates intercultural discourse by a group of adolescent Taiwanese learners of English interacting with adolescents based in the UK on electronic discussion boards and in face-to-face interaction. Given that the role of language use is crucial in intercultural interaction, it is extremely important that the naturally occurring discourse in different contexts of use by different groups of participants is studied in detail. As Boxer (2002) notes, the study of discourse across cultures "represents an especially important endeavor in modern times" because of its great potential "for miscommunication and misperception based on differing norms of interaction across societies and speech communities" (p. 150). This thesis will take a look at this important area, in which differing language patterning by Taiwanese and British participants is examined. In this case, a judicious use of automated corpus linguistic techniques, in conjunction with more established analytical frameworks of discourse analysis,

would greatly benefit such a study to search for particular patterns in the ways that the teenagers interact with each other. Although previous studies have extensively employed these techniques in the investigation of different contexts of language use, few have applied these methods for understanding intercultural discourse and further informing English teaching and learning by identifying the key domains that EFL learners use significantly differently from their native English-speaking interlocutors. As a result, a specialised corpus based on such interaction could prove to be of value as it represents the specific people using the language (i.e., adolescent EFL learners vs. natives of English) in specific contexts (i.e., two different modes of intercultural communication). General corpora may not be appropriate for this function on account of their internal composition (Flowerdew, 2004; Gavioli, 2005; Koester, 2010; McEnery et al., 2006), thus such a specialised corpus may offer a deeper insight into the language use in intercultural settings for the present research purpose and for educators concerned with implementing an intercultural perspective in EFL classrooms.

Although significant English language corpora (e.g., British National Corpus [BNC]), spoken English corpora (e.g., Cambridge and Nottingham Corpus of Discourse in English [CANCODE]), Taiwanese learner corpora (e.g., Taiwanese Learner Corpus of English [TLCE]) and English adolescent corpora (e.g., Corpus of London Teenagers [COLT]) have existed for some time, few, if any, corpora particularly based on adolescent intercultural communication have been established. In the current study the British and Taiwanese Teenage Intercultural Communication Corpus (BATTICC) was constructed by collecting the data from these teenagers' online correspondence on a discussion board and spoken data in face-to-face interaction. In addition, although several studies investigating

intercultural exchange projects have been carried out on the development of language skills and have indicated positive effects on foreign language learning, few focusing on the longitudinal process from different types of intercultural communication – CMC and face-to-face interaction – have been conducted. The design of diachronically-compiled corpora of intercultural discourse (i.e., BATTICC) would be a useful way forward in this respect, focusing on the developmental perspective of language use by Taiwanese EFL learners.

1.3 Aims and objectives

This thesis is a corpus-based study of BATTICC, and its overarching aim is as follows:

> To explore the particular linguistic features of British and Taiwanese adolescent discourse in online and face-to-face intercultural communication.

The study investigates the lexical, grammatical, discourse and pragmatic features of the online and spoken intercultural discourse of British and Taiwanese teenagers from three perspectives: a corpus-linguistics point of view (the keyness approach in particular), a discourse analytical perspective and a multi-word sequence perspective. The keyness method (Baker, 2006; Rayson, 2008; Scott, 2010) allows macroscopic analysis (the study of the characteristics of whole texts or varieties of language) to inform the microscopic level (focusing on the use of a particular linguistic feature) (Rayson, 2008, p. 39). It therefore helps to highlight the significant linguistic features that need to be investigated further. To that end, the following specific questions are addressed:

1. What topics are young people mainly concerned with in online and face-to-face intercultural communication?
2. What are the statistically significant differences in the use of lexical and grammatical categories between Taiwanese and British participants?

Although previous corpus studies have extensively employed keyness analysis in the investigation of different contexts of language use (see Chapter 4), few, if any, have applied this method for the purpose of English language teaching and learning. This study thus intends to demonstrate the pedagogical merit of keyness analysis, which will help to inform teachers and materials developers designing courses for adolescent intercultural interaction.

In the discourse analytical approach, initial quantitative analysis is employed to inform further qualitative analysis, concentrating specifically on the distinctive linguistic features of online and spoken discourse in the intercultural setting. The cultural and functional differences in language use between the two groups of learners in different communication modes are also explored, with analysis pursuing the following questions:

3. What are the distinctive linguistic features of online and spoken discourse by adolescents? To what extent do the British and Taiwanese participants employ them in intercultural communication? To what extent does spoken grammar exist in online discourse?

Another important linguistic aspect that I consider in the thesis is the use of multi-word sequences, which are defined as "frequently occurring contiguous

words that constitute a phrase or a pattern of use" (Greaves & Warren, 2010, p. 213). Previous research has indicated an observable tendency for particular items to co-occur in the written and spoken discourse of both native and non-native speakers of English, and for these co-occurrences to make up an appreciable proportion of authentic language use. This study particularly addresses functional and developmental perspectives of the use of multi-word sequences, as articulated in the following specific questions:

4. What are the high-frequency recurrent multi-word sequences in intercultural communication? Do they serve certain functions in the context?
5. To what extent does the use of multi-word sequences by Taiwanese learners develop over time in the one-year intercultural exchange?

In addition, in light of the potential significance of the research to EFL teaching and learning, I analyse the EFL textbooks used in Taiwanese junior high schools as they constitute the main and perhaps only source of language input that the Taiwanese learners receive. As such, I investigate the textbook dialogues and contrast them with the authentic intercultural communication, with analysis pursuing the following questions:

6. To what extent do the EFL textbooks used in Taiwanese junior high schools display the distinctive linguistic features of authentic intercultural discourse? How can corpus evidence support EFL teaching/learning materials development?

I concentrate particularly on the use of multi-word sequences and the spoken

grammar commonly found in BATTICC and further discuss what the role of textbooks might be in this context, and I also explore how corpus data can benefit learners for better intercultural communication. Based on the findings of the study, a sample of teaching/learning materials is further developed that can be introduced in EFL classrooms in Taiwanese junior high schools. Furthermore, some didactic and methodological advice could be added to the generalised policies in the recent *White Paper on International Education for Primary and Junior High Schools in Taiwan (Draft)* (Ministry of Education, 2011) to provide approaches that would foster both linguistic and intercultural competence in adolescent learners. It is hoped that this research will contribute to a greater understanding of adolescent intercultural discourse in both online and face-to-face interaction, and will provide evidence-based insights which help to inform EFL teaching and learning using authentic texts and the key elements identified in naturally occurring communication.

1.4 Outline of the thesis

The thesis consists of eight chapters. Following on from this introduction, Chapter 2 reviews relevant theoretical ground by providing an account of the discourse in intercultural communication, computer-mediated communication and spoken communication. A number of previous studies on these three issues will also be discussed. Chapter 3 describes methodological issues, including the project background, participants, data collection and procedure and analysis of BATTICC, a corpus that was constructed specifically for this study. This is followed by Chapters 4, 5 and 6, where I present in turn the analysis of online and spoken intercultural discourse from three points of view: a keyness perspective, a

discourse analytical approach and a multi-word sequence point of view. The findings of the analysis will be discussed in turn and the pedagogical implications will be considered. Chapter 7 then reports on the extent to which EFL textbooks used in Taiwanese junior high schools represent the particular linguistic features identified in authentic intercultural communication and what the role of textbooks might be in this context. Finally, Chapter 8 brings together the three perspectives of analysis on online discourse, spoken discourse and textbook conversation by looking at the findings, limitations and pedagogical implications of the thesis.

CHAPTER 2

Language and Intercultural Communication: An Overview

2.1 Introduction

This chapter reviews the theoretical foundations and previous studies of three perspectives: intercultural communication, computer-mediated communication (CMC) and face-to-face interaction, focusing particularly on the language use and marked linguistic features in different communication settings. Before beginning the review, it should be noted that the term *intercultural* rather than *cross-cultural* is employed, although they are sometimes used interchangeably. Nevertheless, there have been several attempts to define the difference between the two terms. According to Cheng (2012):

> Cross-cultural communication compares native discourse across cultures (for example, management meetings of Japanese and those of Americans), whereas intercultural communication involves an investigation of the discourse of people of different cultural and linguistic backgrounds interacting either in a lingua franca or in the native language of one of the participants. (p. 148)

In this study communication involved Taiwanese and British participants using English as the main language in their cultural exchange; the term *intercultural communication* is therefore used throughout this thesis. In this chapter, section 2.2 begins with a definition and discussion of the origins of intercultural communication and its importance in foreign language teaching and learning.

Section 2.3 then looks at the language in CMC by providing some examples of online communication. Emphasis is given to the unique linguistic elements that feature in online discourse. Finally, section 2.4 deals with the significant linguistic features of spoken discourse that are not typically included in traditional written grammar.

2.2 Intercultural communication (IC)

The term *intercultural* literally refers to the concept of "between cultures" (Spencer-Oatey & Franklin, 2009, p. 3), and as noted above, intercultural communication in this thesis can be generally defined as an exchange of information or ideas between people from different linguistic or cultural backgrounds. Kecskes (2012) points out that interculturality is a "situationally emergent and co-constructed phenomenon" that is created in the process of communication in which "cultural norms and models brought into the interaction from prior experience of interlocutors blend with features created ad hoc in the interaction" (p. 69). Žegarac (2007) also identifies the intercultural situation as one in which "the cultural distance between the participants is significant enough to have an adverse effect on communicative success, unless it is appropriately accommodated by the participants" (p. 41). It appears that intercultural communication should be undertaken on the basis of "respect for individuals and equality of human rights as the democratic basis for social interaction" (Byram, Gribkova, & Starkey, 2002, p. 9).

2.2.1 The origin of IC

The growing need for intercultural communication can be traced to the mid-1940s

when World War II ended. A large number of US government officials, diplomats, business leaders and other Americans were assigned to work in several newly independent and developing countries, and they gradually found that the efficacy of communication was impeded as a result of their lack of knowledge of foreign cultures and communication styles (Kumaravadivelu, 2008; Rogers, Hart, & Miike, 2002). The Foreign Service Institute (FSI) therefore began offering the officials programmes with the aim of developing their competence in intercultural communication. The anthropologist Edward T. Hall and the linguist George Trager were integral in such programmes. They jointly wrote *The Analysis of Culture* (Hall & Trager, 1959) as a training manual, presenting "a matrix for mapping a foreign culture along certain dimensions the most important of which was communication" (Kumaravadivelu, 2008, p. 212). The most influential book in this field was then published, namely *The Silent Language* (Hall, 1959), which was "the founding document of the new field of intercultural communication", and Hall is generally acknowledged to be the founder of the formal study of this field (Rogers et al., 2002, p. 11).

2.2.2 From communicative competence to being intercultural speakers

Since the 1970s *communicative competence*, coined by Hymes (1972), has been claimed to be one of the most important competence indicators in second or foreign language teaching. It clearly demonstrated "a shift of emphasis among linguists, away from the study of language as a system in isolation, a focus seen in the work of Chomsky (1965), towards the study of language as communication" (Usó-Juan & Martínez-Flor, 2008, p. 158). The concept has been further developed by Canale and Swain (1980) and Canale (1983) as consisting of grammatical competence, discourse competence, sociolinguistic competence and

strategic competence. Grammatical competence refers to the mastery of the linguistic code of language; discourse competence concerns the ability to associate meanings with acceptable spoken or written texts in various genres; sociolinguistic competence refers to an understanding of appropriate utterances in terms of the context in which they are uttered; and strategic competence refers to verbal or non-verbal strategies that communicators adopt to initiate, terminate, maintain, repair and redirect communication (Brown, 2007; Canale & Swain, 1980; Canale, 1983; Savignon, 2001). It seems, therefore, that knowledge of linguistic forms, meanings and functions are all required to meaningfully and effectively achieve communicative purposes.

However, because of the impact of English as an international language, Alptekin (2002) questions the appropriateness of the communicative competence model, and he asserts that "with its standardised native speaker norms, the model is found to be utopian, unrealistic, and constraining" (p. 57):

> It is utopian not only because native speakership is a linguistic myth, but also because it portrays a monolithic perception of the native speaker's language and culture, by referring chiefly to mainstream ways of thinking and behaving. It is unrealistic because it fails to reflect the lingua franca status of English. It is constraining in that it circumscribes both teacher and learner autonomy by associating the concept of authenticity with the social milieu of the native speaker. (p. 57)

Based on his comments, it can be seen that such a model seems inadequate for foreign language learners in intercultural settings. Therefore, he further suggests

that *intercultural communicative competence* should be developed by equipping the learners with linguistic and cultural behaviours, an awareness of differences and the strategies for coping with these differences. Usó-Juan and Martínez-Flor's (2006) current model of communicative competence also highlights the importance of the intercultural component given the increasing recognition that is nowadays ascribed to cultural aspects.

As a consequence, it is increasingly important to develop language learners as intercultural speakers who have "an ability to interact with "others", to accept other perspectives and perceptions of the world, to mediate between different perspectives, to be conscious of their evaluations and differences" (Byram et al., 2001, p. 5). Similarly, House (2007) defined an intercultural speaker as "a person who has managed to settle for the in-between, who knows and performs in both his and her native culture and in another one acquired at some later date" (p. 19). Nonetheless, Guilherme (2000) cautions that the intercultural speaker is not a cosmopolitan being who is floating over cultures, but someone who is committed to turning intercultural encounters into intercultural relationships.

In language teaching, it is imperative to equip language learners with the competencies that help them reach communicative goals in collaboration with diverse interlocutors from different cultural backgrounds. As a result, Byram et al. (2002) suggest that the aims of the intercultural dimension in language teaching are:

> To give learners intercultural competence as well as linguistic competence; to prepare them for interaction with people of other cultures; to enable them

to understand and accept people from other cultures as individuals with other distinctive perspectives, values and behaviours; and to help them to see that such interaction is an enriching experience. (p. 10)

2.2.3 Intercultural competence in foreign language learning

It is commonly accepted that becoming an intercultural speaker is much more complex than just realising that there are Self and Others (Skopinskaja, 2009). It requires certain attitudes, knowledge and skills to be promoted in addition to learners' linguistic, sociolinguistic and discourse competence. A model of intercultural competence widely employed in foreign language education is the one proposed by Byram (1997, 2000, 2012). This model provides concrete curricular objectives for the foreign language classroom, including the following five elements: attitudes, knowledge, skills of interpreting and relating, skills of discovery and interaction, and critical cultural awareness/political education. Fantini (2000, 2012) also describes five constructs that should be developed for successful intercultural communication: awareness, attitudes, skills, knowledge and language proficiency.

In both well-known models, the attitudes seem to be the foundation of intercultural competence, including "curiosity and openness, readiness to suspend disbelief about other cultures and belief about one's own" (Byram et al., 2002, p. 12). That is, the learners demonstrate a willingness to engage with otherness without prejudice and an interest in discovering other perspectives of the home and target cultures. Another crucial factor is knowledge: "not primarily knowledge about a specific culture, but rather knowledge of how social groups and identities function and what is involved in intercultural interaction" (p. 12). This also forms

part of the classification adopted by the Common European Framework of Reference (CEFR), including, for example "knowledge of the world", "socio-cultural knowledge" and "intercultural awareness", the last aspect receiving the least attention in national curricula (LACE, 2007, p. 25).

In addition, since intercultural speakers need to know how misunderstandings may occur and how they can resolve them, the skills of interpreting and relating are essential, namely, "the ability to interpret a document or event from another culture, to explain it and relate it to documents from one's own" (p. 13). Moreover, skills of discovery and interaction, the "ability to acquire new knowledge of a culture and cultural practices and the ability to operate knowledge, attitudes and skills under the constraints of real-time communication and interaction" are equally important (p. 13) in that language learners acquire the skills of finding out new knowledge and integrating it with what they already have. These skills echo Fantini's (2000) "awareness (of self and others)" involving exploring, experimenting and experiencing; this process is reflective and introspective. In turn, such skills "can be optionally expressed or manifested both to the self and to others" (p. 29) and are also thought of as "the keystone on which effective and appropriate interactions depend" (p. 28).

Last but not least, learners need a critical awareness of themselves and their values, as well as those of other people, namely critical cultural awareness/political education: "an ability to evaluate critically and on the basis of explicit criteria perspectives, practices and products in one's own and other cultures and countries" (Byram et al., 2002, p. 13). It is further asserted that the purpose of this is not to try to change learners' values, but to "make them explicit

and conscious in any evaluative response to others" (ibid.). Kirkpatrick (2007) also recommends that interculturally competent people should "try and ensure that any judgments ... can be supported rationally" (p. 15).

It appears that interculturally competent learners should be encouraged not only to observe similarities and differences between the cultures, but also to be able to analyse them from the viewpoint of others, thus establishing a relationship between their own and other systems (Skopinskaja, 2009). Sercu (2005) further asserts that the five aspects of intercultural competence "should not be considered as isolated components, but rather as components that are integrated and intertwined with the various dimensions of communicative competence" (p. 3).

2.3 Computer-mediated communication (CMC)

With networking tools becoming increasingly advanced, it is much easier to link people virtually in different parts of the world because individuals exchange messages and share information through the Internet in a variety of ways. The term Computer-Mediated Communication (CMC) has been widely known since the 1990s, referring to "a wide range of technologies that facilitate both human communication and the interactive sharing of information through computer networks" (Barnes, 2003, p. 4), and applied linguists are increasingly concerning themselves with the influence of computers and the Internet on language use. This section will firstly illustrate a number of distinctive features of the language in different modes of CMC, and the approaches to CMC discourse analysis will then be examined.

2.3.1 The language of CMC

It is a popular perception that "electronic discourse is writing that very often reads as if it were being spoken – that is, as if the senders were writing talking" (Davis & Brewer, 1997, p. 2), and consequently the language of CMC is often not as grammatically correct, complex and coherent as standard written language. Early on, a number of writers named it *written speech* (e.g., Elmer-Dewitt, 1994) or *visible conversation* (e.g., Colomb & Simutis, 1996). Crystal (2006) proposed the term *Netspeak*, defined as a type of language "displaying features that are unique to the Internet ... arising out of its character as a medium which is electronic, global, and interactive" (p. 20). Crystal (2011) then found *Internet linguistics* the most convenient and satisfactory name for the study of all manifestations of language on the Internet as it provides the required focus, rather than describing human interaction in general. Herring (2013) labels the set of features that characterise the grammar of electronic language *e-grammar*, which exhibits patterns that vary according to technological and situational contexts in text-based CMC.

The various types of economical language use in CMC, such as abbreviations, acronyms and ellipsis, have been one of its most remarked features. For example, some individual words are reduced to two or three letters (e.g., *pls* for *please*, *thx/tx* for *thanks*, *msg* for *message*; some appear because of their obvious rebus-like potential (e.g., *NE1* for *anyone*, *B4* for *before* and *l8r* for *later*). The acronyms can also be sentence-length, such as *AYSOS* [*Are you stupid or something?*], *GTG* [*got to go*], and *WDYS* [*What did you say?*] (Crystal, 2006, 2011). This distinctive language feature of CMC seems to be common in different modes of CMC, and it is likely made by users to economise on typing effort,

mimic spoken language features or express themselves creatively (Herring, 2003, 2012). Crystal (2006) further remarked that this phenomenon might be even more abbreviated as a result of limited character space or smaller screen size in mobile communication.

Novel expressions, such as emoticons or smileys, are another feature commonly appearing in CMC owing to the constraints of technology: the lack of simultaneous feedback, emotional facial expressions, gestures and conventions of body posture and distance (Crystal, 2006, p. 32). That is, the receiver might not be able to send the electronic equivalent of a simultaneous nod, an *uh-uh*, or any of the other audio-visual reactions or emotions that play such a critical part in face-to-face communication (ibid.). Emoticons, short for *emotion icons,* referring to "a visual representation constructed through the use of a series of typographic symbols" (Garrison, Remley, Thomas, & Wierszewski, 2011, p. 112), are therefore commonly employed by CMC users. The most commonly used ones include :-) for pleasure, :-(for sadness, ;-) for winking, :-o for shocked, :~-(for crying and so on. Riordan and Kreuz's (2010) analysis of five contemporary corpora of CMC illustrated that 0.39% of punctuation across all corpora belonged to emoticons and indicated that such expressions are potentially helpful in CMC. However, they have sometimes been criticised as "an unnecessary and unwelcome intrusion into a well-crafted text" (Provine, Spencer, & Mandell, 2007, p. 305). Other attacks are sometimes on the ambiguity of emoticons in that they might forestall a gross misperception of a speakers' intent and lead to misunderstanding if they are not employed appropriately (Garrison et al., 2011; Riordan & Kreuz, 2010).

Moreover, Crystal (2006) notes the existence of several types of "prosody and paralanguage" available online for emotional expression (p. 37), since in text-based CMC, phonology is largely irrelevant; "typography and orthography take over the functions of sound" (Herring, 2013). Examples of such language use, as Crystal illustrated, include repeated letters (e.g., *aaaahhhh, ooooop, soooo*), repeated punctuation marks (e.g., *no more!!!!, hey!!!!!, see what you started??????*) and asterisks (e.g., the *real* answer) for emphasis, capitalised words (e.g., *I SAID NO*) for "shouting" and letter spacing (e.g., *W H Y, N O T*) for "loud and clear" (Crystal, 2006, p. 39). It can be noted that the language of CMC has a wealth of cues, providing information and expressing emotional intimacy, which were seen to be unique to face-to-face communication, to compensate for missing visual and aural cues. Riordan and Kreuz's (2010) corpus study reveals that the role of cues represented in their data is mainly to disambiguate a message (36%), to regulate the interaction (24%), to express affect (15%), and to strengthen the message content (10%). Such strategies also demonstrate the ability of users to adapt the computer medium to their expressive needs.

2.3.2 Different forms of CMC

Forms of CMC can be commonly categorised as either synchronous or asynchronous according to the medium on which CMC is being done. Asynchronous communication does not require that users be logged on at the same time in order to send and receive messages. Typical examples of asynchronous CMC include e-mail and electronic discussion forums (which can be set up in many virtual learning environments, such as MOODLE). In contrast, synchronous CMC takes place in real time. That is, participants communicate with each other at the same time or with a very short delay (Abrams, 2003; Crystal,

2006, 2011; Herring, 2003). Such instant messaging systems are available everywhere, including Windows Live Messenger, Yahoo Instant Messenger, Google Talk, Skype, and so forth. Also, with the rapid development of Internet connections, simultaneous communication is not restricted to text: voice/video-based CMC is becoming increasingly popular.

A number of studies have been conducted to compare asynchronous and synchronous CMC. For example, Perez (2003) examined whether two different modes of CMC, namely asynchronous (e-mail exchanges) and synchronous (online chatting), would affect foreign language learners' language productivity by calculating the number of words produced. Though the students produced a greater number of new words in synchronous chatroom discussions (mean=95) than in e-mail journals (mean=85), the difference was not statistically significant ($p>.05$). In this regard, synchronous CMC seems to be more effective in increasing language productivity. Regarding asynchronous interaction, nevertheless, the participants indicated that they "felt more relaxed... [and] had more time to think and elaborate while writing their weekly e-mail message" (p. 94). As Herring (2003) indicated, asynchronous CMC "permits users to take their time in constructing and editing messages" (p. 618).

Additionally, Abrams (2003) compared the oral language production of different groups of learners who participated in synchronous and asynchronous CMC. In terms of amount of speech produced during oral discussions, participants in the synchronous CMC group significantly outperformed their peers in the asynchronous CMC group when measuring the number of communicative units

(c-units[2]) and number of words that the participants produced. This seems to indicate that learners in the asynchronous CMC group were less motivated to participate in the discussions. Abrams (2003) explained that the members of the asynchronous CMC group sometimes had to wait several days before other members of their discussion group contributed their own comments (p. 164). It seems that responding to asynchronous messages may take from seconds to months due to the recipient's computer access, their habit of using computers and so on (Crystal, 2006), and such delays are highly likely to interrupt the discursive momentum and could reduce motivation. In reference to syntactic complexity, however, asynchronous CMC was more helpful to the subsequent language performance – a higher register (such as subordinate, relative, and infinitive clauses) was frequently used. Moreover, when learners "hark back" to previous comments, scaffold others' ideas and language, and react to others' messages explicitly, it indicates a "more cohesive discussion and reflects more sophisticated interpersonal communication skills" (p. 165).

In sum, the aforementioned previous research studies have indicated that asynchronous communication allows more time for learners to ruminate on and respond to their online interlocutors, though it might lack some of "the most fundamental properties of conversation, such as turn-taking, floor-taking and adjacency-pairing" (Crystal, 2006, p. 154). On the other hand, synchronous communication is more comparable to real-life communication with no intervals of delay and fast responses that shorten the time needed for communication, facilitating the amount of output.

[2] C-units are "isolated phrases not [necessarily] accompanied by a verb, but they have a communicative value" (Crookes, 1990, p. 184).

2.3.3 From CMC to face-to-face (FTF) interaction

While much has been discussed about the nature and implications of both CMC and FTF communication, it is interesting to examine the extent to which CMC affects follow-up face-to-face interaction, as has been presented in a range of previous research. Dietz-Uhler and Bishop-Clark (2001) conducted a study assessing the effects of synchronous and asynchronous CMC on subsequent FTF discussions. In their study, the participants were asked to read a short article before they were randomly divided into three groups: a synchronous (Internet chat) group, an asynchronous (Internet discussion board) group and a control group. The first two groups engaged in an online discussion about the article they read and face-to-face communication followed, while the control group had no online discussion but instead immediately began a face-to-face discussion. The results showed a positive effect of CMC on subsequent face-to-face discussions, significantly improving the confidence and enjoyment of the participants and helping them to express various perspectives.

Similarly, a recent study was conducted by Jeng (2010) on Taiwanese first-year undergraduate students, examining whether their performance in synchronous CMC can be transferred to its follow-up oral face-to-face discussion in terms of accuracy, lexical complexity, fluency and syntactic complexity. Though the results did not show statistical significance in such transferability, when comparing language output between the CMC and its follow-up FTF interaction, the differences between the interactions in all language areas were significant, except for syntactic complexity. Concerning accuracy, the averaged mean score of error rate was lower in synchronous CMC (39%) than in FTF discussion (59%), p=0.04 (p <.05), so the learners were apparently able to produce much more accurate

utterances in the synchronous CMC context than in the FTF interactions. In lexical complexity, the percentage of less frequent words was also higher in CMC (11.70%) than FTF (9.05%), p=0.03 (p <.05). Discourse in CMC therefore seems to display greater lexical range. As for fluency, including repetition, self-repair and incomplete sentences, relatively few markers were found in the CMC scripts, while such markers inevitably recurred in the face-to-face discussion (p=0.00). With regard to syntactic complexity, the only insignificant result, p=0.08 (p>.05), the research revealed a tendency to use simpler structures in synchronous CMC contexts.

Though little research has revealed statistical significance in the transferability of CMC and FTF interaction, CMC still has a positive effect on subsequent face-to-face discussions, that is, CMC seems to be "a stepping stone" to face-to-face communication. This might result from the integrated and higher-level competence needed to interact face-to-face, requiring each individual "to decode input, process it, and simultaneously plan his or her output, as well as make immediate decisions about style, register, cultural referents, pronunciation, lexicon, and syntax, both in listening and speaking" (Abrams, 2003, p. 158). As a result, it is accepted that CMC may well help individuals prepare for the following face-to-face discussions and feel less inhibited and less constrained; they are then likely to feel more comfortable and confident in expressing their thoughts, ideas and opinions in spoken communication. We now turn our attention to the literature review of spoken discourse, focusing on the particular linguistic features resulting from its real-time and interpersonal nature.

2.4 Spoken discourse in FTF interaction

Spoken discourse usually occurs in FTF interaction, which takes place in real time and is usually unplanned. Cutting (2011) notes that spoken language is a reflection of "the process of language construction", whereas written language is "a revised and polished product" in that writers usually have more opportunities to plan and structure their discourse than speakers; as a result, spoken discourse is often not as sophisticated as its written equivalent, that is, it often has "lower lexical density" and "less intricate grammar" than written discourse (ibid, p. 158-163). A number of scholars, such as Biber et al. (1999), Leech (2000), Carter and McCarthy (2006), Thornbury and Slade (2006), Mumford (2009) and Cutting (2011), have examined the grammar of spoken language and they all seem to agree that speech and writing basically share the same underlying grammatical system, but the system is adapted in a variety of dynamic and often resourceful ways to meet the specific situations in which each medium is applied. This section considers two core natures of spoken language, namely spontaneity and interpersonal interactiveness, and illustrates the unique features that distinguish it from written language.

2.4.1 Spoken discourse and real-time communication

One such feature is the nature of spontaneity in real-time conversation in which speakers do not often construct over-elaborate patterns, clauses or sentences. Conversations therefore often consist just of words or phrases, incomplete clauses or indeterminate sentence structures, since they are unplanned. For example, speakers might abandon or restart an utterance, or it is sometimes completed by other interlocutors, or its non-completion is sometimes caused by the interruption

of other speakers or situations. Carter and McCarthy (2006, p. 168) also note that in real-time speech, "utterances are linked ... as if in a chain" and thereby coordinating conjunctions (i.e., *and, or* and *but*) and simple subordinating conjunctions (i.e., *so* and *because*) are commonly used by speakers. Clause complexes therefore need reassessment since in spoken language clauses that are traditionally restricted to a subordinate function often have the capacity to function as main clauses as well (see McCarthy, 2006).

In addition, word order is more flexible in speech as "it is constructed in real time and follows the order of ideas emerging from a speaker, which may override grammar rules" (Carter & McCarthy, 2006, p. 172). Headers and tails may alter the word order of traditional written grammar, showing that spoken discourse is much more flexible than written forms are (Carter, Hughes, & McCarthy, 2011; Mumford, 2009). The term *headers* refers to fronting "adjuncts, objects and complements, and noun phrases before the pronoun" (Cutting, 2011, p. 160), for example, "*the teacher*, he is very nice". This is often used to emphasise what the speaker thinks to be particularly important. Tails, on the other hand, normally occur after clauses, which are commonly used to clarify or make explicit something in the main clause (Carter et al., 2011), for example, "They're really nice, *my teachers*". In this case, the noun phrase *my teachers* clarifies or repeats the referent of the pronoun *They* in the sentence that comes before it. The position of adverbials is another example of spoken feature. For instance, in casual conversation adverbials may occur after tags, as in "Spanish is more widely used isn't it *outside of Europe*?" (example presented by McCarthy, 2006, p. 38). The point made here is that the ordering of elements in spoken discourse may not conform to the norms of written language because of the constraints of

spontaneity and the need for "clear acts of topicalization and suchlike to appropriately orientate the listener" in face-to-face interaction (ibid.).

Hesitation devices are also extremely frequent in natural spoken discourse as "speakers attempt to keep the floor while formulating their next utterance" (Gilmore, 2004, p. 369). These are also called *dysfluencies* (Thornbury & Slade, 2006; Cutting, 2011). These occur when "the need to keep talking ... threatens to run ahead of mental planning" (Biber et al., 1999, p. 1048). In this regard, pauses, repeating and recasting can be commonly found in real-time conversation. The vocalisation *er* or *erm* typically features in pauses in unplanned speech. As shown in Carter and McCarthy's (2006) analysis, *er* represents the 17th most common item in native-speaker speech (p. 12). Repeating words or phrases also seems to be one way in which speakers can buy more time for thought. Recasting frequently occurs under the pressure of real time in that speakers backtrack and reformulate words and phrases (ibid.).

Carter and McCarthy (2006) claim that spoken language "foregrounds choices which reflect the immediate social and interpersonal situation" (p. 164). That is, conversation normally takes place in a shared context and is highly interactive, typically being co-constructed by the interlocutors and involving dynamic and unplanned turn-takings. In this regard, speakers need to adopt ways to organise their discourse and further signal to the listeners what is happening. For example, at the beginning or transition points of speakers' turns, certain words or phrases such as *yeah, oh, well, great, so, all right, you know, I mean,* etc. are used frequently as interjections or discourse markers (DMs), which function to "link segments of the discourse to one another in ways which reflect choices of

monitoring, organisation and management exercised by the speaker" (ibid., p. 208). In this regard, DMs seem to act as "punctuation for speech", which can be used to signal and signpost for the speaker (Carter, 2008, p. 15). This coherence-based point of view is concordant with Schiffrin's (1987) definition of DMs as "sequentially dependent elements which bracket units of talk" (p. 31). Fraser (1999) also states that DMs "impose a relationship between some aspect of the discourse segment they are a part of ... and some aspect of a prior discourse segment" (p. 938). It appears that DMs are crucial in real-time FTF communication. Without them, however, speakers are not able to use cues in organising the discourse and to indicate degrees of formality and people's feelings towards the ongoing interaction (Carter & McCarthy, 2006, p. 212), although they are often semantically and grammatically optional, that is, they occasionally can be excluded from utterances without syntactic and semantic consequences (Fraser, 1999; Fung & Carter, 2007; Hellermann & Vergun, 2007; Schiffrin, 1987).

The use of discourse markers amongst native English speakers and secondary school pupils in Hong Kong is examined and compared by Fung and Carter (2007) based on a pedagogic sub-corpus from CANCODE. They show evidence that discourse markers serve as "useful interactional manoeuvres" to organise and structure speech on interpersonal, referential, structural and cognitive levels (p. 410). This notwithstanding, Evison (2008) notes that the range of pragmatic functions encoded in discourse marking by the L2 speakers is narrower than that of the L1 speakers. It might not therefore be easy for foreign language learners to use language in culturally, socially and situationally appropriate ways. Fung and Carter (2007) further draw on Wierzbicka's work in the field of cross-cultural pragmatics (e.g., 1991) and stress that DMs are useful conversational devices, "not just for

maintaining discourse cohesiveness and communicative effectiveness, but also for interpersonal and cross-cultural interaction" (Fung & Carter, 2007, p. 433). Given the interpersonal nature of spoken discourse, it is to this kind of feature that we now turn.

2.4.2 Spoken discourse and interpersonal communication

It is accepted that maintaining good relations between the speaker and hearer is important in casual face-to-face conversation, particularly in intercultural communication. O'Keeffe et al. (2007) use the term "relational language" to refer to language that serves to "create and maintain [a] good relationship between the speaker and hearer" (p. 159). One such device is vague language, which is found to be particularly common in daily conversation as speakers are often cautious not to sound over definite, which might be perceived as threatening or over-educated (Carter & McCarthy, 2006; Cheng & Warren, 2003; O'Keeffe et al., 2007). It involves the use of word or phrases such as *thing, stuff, stuff like that, or something, or anything, and everything, and so on, and things like that, kind of* and *sort of*. One of the main functions of vague language proposed by O'Keeffe et al. (2007) is to "hedge assertions or to make them fuzzy by allowing speakers to downtone what they say" (p. 177). In this regard, vague language softens expressions, so the speakers, as Carter and McCarthy (2006) note, "do not appear too direct or unduly authoritative and assertive"; the use of vague expressions is therefore a conscious choice by speakers and is not a product of "careless thinking or sloppy expression" (p. 202). The other function is to "indicate assumed or shared knowledge and mark in-group membership" (O'Keeffe et al., 2007, p. 177). In other words, it is "a marker of intersubjectivity" (Overstreet & Yule, 2002, p.

787). The speakers, in this case, do not necessarily convey precise and concrete information, and the hearers would know what their vague expressions refer to.

Situational ellipsis is another linguistic feature resulting from the nature of shared contexts in conversation. It means "not explicitly referring to people and things which are in the immediate situation" (Carter & McCarthy, 2006, p. 181). That is, in many situations some items, such as subject pronouns and verb complements, seem to be redundant because they are "recoverable from the immediate context, either the linguistic context or the situational one" (Thornbury & Slade, 2006, p. 83), and consequently they are often deliberately omitted. Mumford (2009) states that "rather than being impolite or casual, ellipsis is actually more appropriate than full forms in certain situations" (p. 141). Although situational ellipsis does not often conform to written grammar norms, it is a pervasive and natural feature of spoken English.

In addition, as noted in the previous section, speakers in face-to-face communication regularly use some words or phrases, i.e., DMs, to indicate their intentions regarding organising, structuring and monitoring the discourse, such as *well, right, I mean, you know* and *as I was saying*. These words or phrases also have important interpersonal functions in FTF communication, being used to indicate shared knowledge, attitudes of the speaker and responses like agreement, confirmation and acknowledgement (Fung & Carter, 2007). For example, *you know*, which is the most common chunk, is "an important signal of (projected or assumed) shared knowledge between speakers and listener, as well as being a topic-launcher" (O'Keeffe et al., 2007, p. 173).

This section has presented some lexical, grammatical and discourse features of spoken language, which have distinctive special qualities that distinguish them from the features of traditional written grammar. However, it is reported that our EFL classrooms have long been prioritising formal written grammar; spoken language is often considered to represent nonstandard forms of target language, and is therefore neglected in classroom teaching and EFL learning materials (Cullen & Kuo, 2007; McCarthy, 2006). As McCarthy (2006) claims, there can be "little hope for a natural spoken output on the part of language learners if the input is stubbornly rooted in models that owe their origin and shape to the written language" (p. 29).

In this case, learners will probably be heard as "at best rather formal and at worst pedantic and bookish" if they apply the written interaction model in actual intercultural communication in casual settings (McCarthy, 2006, p. 33). As a result, corpus linguists have proposed that the authentic data in a corpus can provide an empirical basis for language description by showing how language is actually used in natural contexts. By bringing to light features of language use, corpus-based and corpus-informed approaches can help improve syllabus and teaching materials design in English language teaching.

2.5 Summary

The literature reviewed in this chapter has come from three perspectives of language-in-use and intercultural communication. It firstly provided an account of the intercultural aspect of language teaching and learning (section 2.1) and further concentrated particularly on the distinctive linguistic features of CMC (section 2.2)

and spoken discourse (section 2.3), which are fundamental to the next stage of the investigations on the two different modes of intercultural communication. These ideas will be incorporated into the analytical framework for the present research that will be put forward in chapters 4, 5 and 6. Before the outset of the analysis, Chapter 3 will consider methodological issues for the research, including the background of the intercultural exchange project, participants, data collection and data analysis. It is to the data and methodology of empirical analysis of online and spoken data that we now turn.

CHAPTER 3

Data and Methodology

3.1 Introduction

Since this thesis aims to examine the particular linguistic patterns of online and face-to-face communication by adolescents, an analysis based on naturally occurring[3] samples of language data is of great importance. Teubert (2005) notes that "[the] corpus is considered the default resource for almost anyone working in linguistics" and "no introspection can claim credence without verification through real language data" (p. 1). That is, working with real language data provides insights into how language patterns are actually used and structured in different contexts of use. However, little effort has been made to date to offer systematic descriptions of naturally occurring intercultural interaction involving Taiwanese and English adolescent participants in both online and spoken settings. To embark on such descriptions, a specialised corpus can therefore be of value as it represents the language use of specific people in specific contexts, which helps to elucidate the connections between linguistic patterning and contexts of use (Koester, 2010, p. 67).

This chapter considers methodological issues for the research, beginning with the background of the intercultural exchange project, participants and procedure (3.2). I will then present a more detailed discussion of the definition of a corpus and the development of my own corpus BATTICC (3.3). Following on from this, three

[3] Discourse that is naturally occurring refers to language data that is not produced through the instigation of the researcher (Wood & Kroger, 2000).

perspectives of data analysis will be presented (3.4), including keyness analysis, a discourse analytical approach and multi-word sequence analysis.

3.2 Project overview

3.2.1 Project background

Hualien County Government in Taiwan and the Collaborative Venture of Cumbrian Secondary Schools in the UK have been a part of the global partnership programme "British Council Connecting Classrooms Project" since 2009. The project was initially funded by the British Council and is now jointly administered by the two bodies. The aim of the project is to create global partnerships between clusters of schools in Cumbria (UK) and Hualien (Taiwan), and thus offer language learners an opportunity to communicate and work directly with their international peers. I have been involved in this intercultural exchange project since the beginning of the partnership between Hualien and Cumbria, when I was a secondary school teacher in Hualien. I have also served as the main corresponding person in the programme, helping to organise both online and face-to-face conferencing between teachers, students and Council officers from the two countries. Given my critical role in this partnership, I was granted permission to conduct research on the project and collect language samples for further analysis. The thesis is based on this project, with the aim to explore the particular linguistic features of the discourse used by the adolescent British and Taiwanese participants in online and spoken communication. I also noticed the importance of researcher reflexivity and endeavored to act as an objective individual researcher throughout the data collection process so that the data production would not be affected by my engagement.

3.2.2 Participants

The participants recruited for the study were 35 EFL learners from Hualien (Taiwan) and 35 English secondary school students from Cumbria, between 13 and 14 years of age, all participating in the Connecting Classrooms Project. Hualien and Cumbria have a number of common characteristics, including their natural environment and economy. Most of the Taiwanese participants are English learners at early secondary schools (lower intermediate level), having learned English for on average five to six years, so they are equipped with some basic knowledge and skills of using English for communication. In addition, nearly all (97.5%) have never had the experience of interacting with students with a different linguistic and cultural background, and from English-speaking countries in particular; similarly, few of the English participants have a friend or an online pen pal from an Asian country.

Ethical consent to conduct research is another important issue in the current study. Since the participants recruited in the study were under 18 years old, consent was acquired from all of the participants' parents before the study commenced (see Appendix C). Included in this process was information about the purposes of the study, the use of the participants' online and spoken data and their consent for the participants to complete a subsequent questionnaire or an interview for the investigation of the intercultural experiences in both online and face-to-face interaction. Participants were guaranteed anonymity in all communication and correspondence relating to the study.

Pseudonyms were therefore used in all reporting and write-ups of the thesis. In order to obtain the consent from the participants' parents, I made a brief

presentation about the study for all of the participants and their parents in an International Exposition in Cumbria. After the presentation all of the consent forms were returned to me and all of the British participants and their parents were happy with my research. The same consent forms were collected from the Taiwanese participants' parents as well. Permission to collect data from the participants for the research was kindly provided by all parents.

3.2.3 Procedure

The project began with asynchronous CMC, interacting on an electronic discussion board. Prior to the outset of the programme, Ms Alison Phillips from Stainburn School, who is the main corresponding person in Cumbria and I developed a Moodle website (http://vle.connectingclassrooms.cleo.net.uk/) particularly for this project. Moodle was chosen as the main exchange platform because it is an open resource and very adaptable to the necessities of teachers and learners, as well as being both free to install and easy to use (Markey, 2007). To ensure that the online learning community remained safe, only the participants enrolled on the Moodle website were able to log onto it. Therefore, the children and their identities were not available to outsiders. At the same time, the online environment was monitored all through the programme by the researcher and Ms Alison Phillips. Figure 3.1 below reproduces the website's homepage.

Figure 3.1 *The Moodle Homepage of the Project*

To be certain that both British and Taiwanese participants were prepared to use the Moodle website, an introductory session was offered to familiarise them with the functions and operations of the website. Both British and Taiwanese participants were then encouraged to log onto the website and either write something or respond to others' messages in the online forum every week. Such regular correspondence between participants is fairly important since personal relationship building in online communication settings mainly relies on consistent interaction (Belz, 2007). In this setting, English was used as the major communication language, and the students had the choice to create any topics that they felt interested in or respond to any topics created by other participants on the website. They could also read articles about their respective cultures that other participants uploaded onto the website, and then write their opinions about the articles. Every time a message was posted to the forum, an e-mail alert was sent to all of the participants so that they and the teachers would know which topics were currently being discussed.

At the outset of the research, it was my intention to further carry out regular

synchronous CMC. However, deciding when participants from both countries could get online at the same time was not an easy task, a challenge that has also been noted by other researchers (Liaw & Master, 2010; Hauck, 2007; O'Dowd & Ritter, 2006). Due to the eight-hour time difference between Taiwan and the UK, when British participants arrived at school at 9 a.m., it was already 5 p.m. in Taiwan, the time at which most of the Taiwanese students left school. Therefore the time that participants from the two countries could chat online was rather limited. Although some synchronous CMC data was collected, it was not enough to analyse for the purposes of this research and was therefore excluded from the thesis.

The asynchronous CMC in this project lasted one year, during which time participants interacted weekly with one another on the Moodle electronic discussion board, followed by a one-week face-to-face meeting in Taiwan. During the last three months of online interaction, it can be observed from the messages on the board that the participants from both countries were getting intense and animated, as demonstrated by the use of more emotional words, lengthier communication being produced, etc., as they were eventually going to meet their Internet international peers in person.

After a one-year period of online communication, the British participants travelled to Taiwan and stayed with the Taiwanese participants for one week. They were divided into five groups with equal numbers of English natives and nonnatives. Oral group discussions in the face-to-face meetings were audio-recorded and then co-transcribed by the researcher and the Taiwanese participants' English teachers. The transcription was then checked with the original recording for accuracy. The

spoken data collection resulted in approximately 3 hours of recorded conversations. The data collection and the process of corpus compilation will be further discussed in detail in the following section.

3.3 Corpus design and construction

A corpus can be broadly defined as a body of naturally occurring texts, which may be written or spoken in origin. According to Sinclair (2005), a corpus is "a collection of pieces of language text in electronic form, selected according to external criteria to represent, as far as possible, a language or language variety as a source of data for linguistic research" (p. 16). As such, a corpus is not any collection of texts; a number of design criteria need to be met. The following sections will first introduce some criteria that a corpus needs to satisfy and how the data collection in this study meets these criteria. I will then present a more detailed discussion of the composition of my own corpora British and Taiwanese Teenage Intercultural Communication Corpus (BATTICC) and Taiwanese EFL textbook corpus of conversation (TETCOC).

3.3.1 British and Taiwanese Teenage Intercultural Communication Corpus (BATTICC)

McEnery, Xiao and Tono (2006) note that a corpus is a collection of "(1) *machine-readable* (2) *authentic* texts (including transcripts of spoken data) which is (3) *sampled* to be (4) *representative* of a particular language or language variety" (p. 5). In this regard, any collection of naturally occurring language data based on these four essential criteria can be labeled a corpus. These criteria have also been mentioned by corpus linguists (e.g., Adolphs & Lin, 2010; Biber, 1993;

Cheng, 2012b; Hunston, 2002; Reppen, 2010; Sinclair, 2004, 2005; O'Keeffe et al., 2006, 2011). In the development of BATTICC, the first criterion, machine-readability, was achieved by transferring all of the data into Notepad files (plain text files), which are suitable for corpus analysis using programmes such as *WordSmith Tools* and *WMatrix*. Computer-readable data is easy for researchers to manipulate and exploit by, for example, searching, selecting, comparing, sorting and formatting, and it can also help avoid human bias in an analysis, thus making analytical findings more reliable (Scott, 2010).

Another criterion, authenticity, can be met due to the naturally occurring and unplanned nature of interaction in BATTICC, which was collected from messages posted to an electronic discussion board and from natural face-to-face interactions between teenage participants from Taiwan and the UK. Throughout the process of data collection, great importance was placed on the naturally occurring nature of the discourse; that is, language data was produced based on the participants' volition, not through the instigation of the researcher. By emphasising this, it was believed that the credibility of the research results would be increased. In the process of their communication, therefore, the participants were not given any specific guidelines about how to start the conversation, what to talk about or how to structure their chats, so they had their own choices to create a conversational topic and sustain the communication. The enormous potential that authentic texts offer in yielding reliable quantitative and qualitative information that intuition alone cannot perceive has been described (Adolphs, 2006; Carter et al., 2010; McEnery et al., 2006), and this authentic data can be further applied to language teaching and to inform materials development (e.g., Gilmore, 2004, 2007; O'Keeffe et al., 2007; Römer, 2009; Meunier & Gouverneur, 2009).

The other two essential features of a corpus are *representativeness* and *sampling*. McEnery et al. (2006) note that these are the features typically used to distinguish a corpus from an archive (p. 13). That is, an archive is simply a random collection of texts whereas a corpus is designed to provide insight into a particular genre. In this study, BATTICC was constructed to represent the informal nature of intercultural communication by young learners in online and face-to-face settings. As such, all the samples collected represent that genre. Biber (1993) defines representativeness as "the extent to which a sample includes the full range of variability in a population" (p. 243). This suggests that one should strive to collect samples from all the possible situations within a certain genre to completely present the language being studied. As regards the data of online discourse for BATTICC, all the concordances on the electronic discussion board were collected to achieve a complete representation. However, for collecting spoken discourse, it does not seem to be possible to record all of the spoken interactions in the participants' daily lives. Koester (2010) suggests that "what is important is to ensure that the samples are collected from a range of fairly typical situations" (p. 69). In this regard, as the aims of the intercultural exchange project were to build relationships between participants in casual settings, spoken data for BATTICC was collected from a range of informal chats between Taiwanese and British participants during the intercultural exchange programme, in a wide variety of locations such as schools, homes, restaurants, tourist spots, public parks and social gatherings, wherever possible consisting of the entire speech event. Nevertheless, it needs to be noted that this thesis has attempted to demonstrate the particular linguistic patterns via a case study of the intercultural communication project, and consequently the sample might not lead to any generalisable observations for intercultural interaction in general.

Apart from the process of collecting and recording the actual interaction among participants, any further information about the event itself, namely metadata, was also documented, so the source information can be easily retrieved. As suggested by Adolphs and Knight (2010) and Sinclair (2005), metadata is critical to a corpus to help achieve the standards for representativeness. Burnard (2005) also notes that "without metadata, the investigator has nothing but disconnected words of unknowable provenance or authenticity" (p. 31). In BATTICC, metadata is kept in a separate document, including documentary data, information about participants, contexts, locations and the relationship between corpus components and their original source. By keeping this detailed information, the corpus can be shared or reused by other researchers in the future.

Table 3.1
The Composition of BATTICC

	BATTICC-O	BATTICC-O	BATTICC-F	BATTICC-F
Messages/turn	624	683	750	1,073
Words	16,998	15,450	7,624	12,475
Types	2,199	1,993	919	1,509
TTR	12.94	12.90	13.45	13.99
Sentence	21.34	14.01	6.29	7.70

Table 3.1 shows the composition of BATTICC produced by *WordSmith Tools 5.0*, details of which will be discussed in the following subsections. Data for BATTICC was divided into two sub-datasets: BATTICC-O, including discourse of online communication and BATTICC-F, representing spoken discourse in face-to-face interaction, with 32,442 and 20,099 words respectively (see Appendix B). In order to reveal more information about cultural differences as well as differences in language use by different groups of participants, both

BATTICC-O and BATTICC-F are also divided according to users of different countries. Although this collection of data is relatively small compared to some existing, ready-made corpora, many corpus linguists note that the required size of a corpus depends upon the purposes for the study and the language parts to be analysed (Adolphs, 2006; Biber, 1993; Hunston, 2002; McEnery et al., 2006; Koester, 2010; Sinclair, 2001). The comparative nature of this study, which focuses on different patterns of language use in different communication modes by different groups of people, may in fact make this smaller size of corpus an advantage. As Sinclair (2001) claims, "comparison uncovers differences almost regardless of size" (p. xii). In this regard, small specialised corpora are adequate to provide sufficient examples of frequent linguistic patterns for illuminating differences between registers. Previous studies based on small specialised corpus include Cutting's (2000) exploration of language use contributing to in-group identity using a 26,000-word corpus of conversation, Koester's (2006) investigation of workplace discourse which uses a corpus of approximately 30,000 words, and O'Keeffe's (2006) study of radio discourse, based on a 55,000 word corpus of calls to phone-ins.

In addition, although it is accepted that statistical analysis over many millions of words and broad contexts in large corpora may demonstrate reliable evidence, this approach tends to mainly be used for quantitative analysis and thus is less easily applied to qualitative analysis of language use in a specific situation. On the other hand, the size and composition of a specialised corpus make it more manageable for qualitative studies as it is more possible to examine most of the concordance lines (not just a random sample) of particular linguistic features in contexts, which can provide a rich source of data to complement more quantitative studies

(Hunston, 2002; Flowerdew, 2004; Harvey, 2008; Koester, 2006, 2010). Furthermore, in working with very large corpora, it is always difficult to describe the original context of use of the utterances since the samples come from many vastly different contexts (Flowerdew, 2004; Koester, 2010). As a result, in this study the specialised corpus BATTICC allows examination using both quantitative and qualitative approaches, with a much closer link between the corpus and the context in order to understand the key linguistic features in online communication (BATTICC-O) and face-to-face interaction (BATTICC-F) between Taiwanese and British young learners. It also allows me to compare the language presented in the Taiwanese EFL textbook conversation. Additionally, for the pedagogical implications of corpus data, large corpora are sometimes not suitable for use in teaching and learning. Tribble (2002) argues that they provide "either too much data across too large a spectrum, or too little focused data, to be directly helpful to learners with specific learning purposes" (p. 132) when learners interact with a large corpus directly. In this regard, smaller and more focused corpora are of particular help for pedagogical purposes, and they can further be used to inform materials development for this present study.

3.3.1.1 Online discourse: BATTICC-O

The data that forms the basis of BATTICC-O was collected from messages posted to an electronic discussion board by participants from Taiwan and the UK between September 2010 and August 2011, amounting to a total of 1035 messages, comprising 31,910 words, as shown in Table 3.1. In the corpus construction, some of the data were excluded from the corpus otherwise the data may be skewed and affect the accuracy of subsequent analysis. As suggested by previous research on

the data cleaning process of a CMC corpus (e.g., Harvey, 2008; Riordan & Kreuz, 2010), the removals in BATTICC-O include website and e-mail addresses, usernames of writers, any non-English language words, and all detectable duplicate entries, such as those present in forwarded messages. Moreover, the information that was not part of the main message, such as time of posting, which was not generated by the user was removed, and the titles of the messages under the same topic in the discussion board were left only once to indicate the first occurrence.

Another issue involved in the development of BATTICC-O is the spelling as the conventions that Taiwanese learners adopt is typically more American style, in which inconsistency of spelling with the British participants arises. Due to the comparative nature of the study, a consistent spelling system in the discourse by the two groups of participants would be more appropriate to achieve a higher level of quantitative validity. As Harvey (2008) notes, typos, nonconventional spelling and other linguistic irregularities are likely to skew statistical measures, such as frequency and keyword counts, which provide important quantitative insights into characteristics of the corpus (p. 111). As such, the words used by Taiwanese participants that had American spelling were amended to be in agreement with the British style, including *favorite* (52 instances), *color* (9 instances), *neighbor* (4 instances) and *program* (4 instances). Nevertheless, it was found that some Taiwanese participants adopted the British spelling in their messages, which was found in 6 instances of *favourite* and 1 instance of *programme*.

3.3.1.2 FTF discourse: BATTICC-F

BATTICC-F includes approximately 3 hours of discussions made up of 21

separate conversations recorded by digital recorders, and the length of each conversation varied, with the longest conversation lasting 35 minutes and the shortest 1 minute 57 seconds. I strove to ensure that the data collected was accurate and exhaustive, capturing as much information from the content and context of the discursive environment as possible, as suggested by Adolphs and Knight (2010). The spoken material is not used in its original audio format but has been transcribed into the electronic written transcripts based on standard orthographic practices in order to be analysed using currently available corpus tools and to make it re-usable by the research community. The layout of the transcripts is the most commonly used format, namely a linear representation of turns with varying degrees of detail of linguistic and extra-linguistic information. The transcription convention for this study was adapted from the VOICE Transcription Conventions [2.1] (VOICE Project, 2007) and the work done by Carter (2004), but only items that are related to the purpose of this study were selected (see Appendix A). As can be seen in the guidelines, the speaker codes (e.g., <TW01>, <TW02>, <BT01>, <BT02>,...) represent speakers from two different countries, namely TW referring to Taiwanese learners and BT for British participants. Speakers are generally numbered, and the code is given at the beginning of each turn. Extralinguistic information is also included, which is described in square brackets [ ], such as [laughter], [coughing] and [inaudible speech]. Additionally, interrupted sentences (e.g., <TW01>: I think +) are marked with a plus "+", unfinished words with an equals sign "=", lengthened sounds with a colon ":" and exceptionally long sounds (i.e., approximating 2 seconds or more) are marked with a double colon "::". Moreover, a sequence of two dots ".." indicates a brief break in speech rhythm, and a longer pause is marked as a sequence of three dots "...". All repetitions of words and phrases (including

self-interruptions and false starts) were also transcribed. However, the representation of prosodic and kinesic elements of spoken interaction is not included in the transcription as these features are not closely related to the research questions of the thesis. More details of the transcription convention are shown in Appendix A.

As can be seen in Table 3.1, BATTICC-F contains 20,099 words in total. This collection is further divided into British and Taiwanese sub-datasets, with 12,475 words and 7,624 words respectively. The number of words produced by the two groups of participants seems to be relatively unbalanced in that the Taiwanese learners speak less in group discussions. The uneven distribution may be due to the Taiwanese participants' lower language proficiency than the native-speaking participants. In light of the potential significance of the research to EFL teaching and learning, I also analysed the EFL textbooks used in Taiwanese junior high schools as they constitute the main and perhaps only source of language input that the Taiwanese learners receive. It is to the assembly of the Taiwanese EFL Textbook Corpus of Conversation (TETCOC) that we now turn.

3.3.2 Taiwanese EFL textbook corpus of conversation: TETCOC

The TETCOC is a newly developed corpus since there are no existing computerised collections of EFL textbooks for use in junior high schools in Taiwan available. Developing an electronic corpus of Taiwanese EFL textbooks therefore allows specific language patterns in several volumes of textbooks to be examined easily, and it also makes it possible to analyse the data with computer applications (e.g., *WordSmith Tools*) to generate useful information (e.g., word frequencies, recurrent sequences etc.) rapidly for this study.

Table 3.2

The Composition of TETCOC

	Nan-I	*Kang-Xuan*	*Han-Lin*	Whole
Words	9,316	7,554	7,670	24,540
Types	1,422	1,256	1,264	2,219
TTR	15.26	16.63	16.48	9.24
Sentences	860	624	719	2,203
Sentence length	10.14	10.72	9.52	10.12

As presented in Table 3.2, TETCOC consists of spoken texts taken from the three most commonly-used EFL textbook series, *Kang-Xuan English* (Chou, 2010), *Nan-I English* (Liu, 2010) and *Han-Lin English* (Huang, 2010), each containing 6 volumes (low-intermediate to intermediate level) and all published in 2010. Since this corpus contains the textbooks used in more than 95% of the junior high schools in Taiwan, it can be considered appropriately representative of this particular discourse genre. In order to support comparison with authentic face-to-face intercultural communication, only instances of conversational transcripts in the textbooks were included in this corpus. Exclusively written materials, such as narratives, reading comprehension texts, grammar exercises or excerpts from novels, were not included. The corpus includes approximately 27,960 words in total, which is of a similar size to BATTICC and therefore enables the two corpora to be more comparable.

3.4 Data analysis

This section begins with a discussion of the usefulness and limitations of the corpus analytical tools and frequency-driven approach that are employed throughout the thesis. Three major perspectives of data analysis – a keyness

approach, a discourse analytical approach and a multi-word sequence perspective – will then be discussed by considering the insights and their potential limitations.

3.4.1 Corpus-based automatic extraction tools

Computers have affected the methodological scheme of language research. As noted in the previous section, one of the important criteria for a corpus is machine-readability. Such data is easy to manipulate and exploit at minimal cost, for example through searching, selecting, comparing, sorting and formatting. It can also avoid human bias in an analysis as these processes are fully automated with the least amount of human intervention, thus making analytical findings more reliable (McEnery et al., 2006; see also Dahlmann, 2009; Scott, 2010). By using computer programmes for corpus analysis, the speed of processing that computers afford is perhaps the central advantage (ibid.). What is needed is simply to upload the data to the programme and operate the functions, then the results are displayed on the screen.

WordSmith Tools 5.0 (Scott, 2008) was mainly employed throughout the study. It is an integrated suite of computer programmes for lexical analysis, looking at how words behave in texts (Scott, 2010). In this software three central functions were frequently used in the thesis. The *WordList* tool generates a list of all the words or word-clusters in a text in alphabetical or frequency order. I concentrated particularly on the frequency list as it provided important information on how lexical items were used in discourse. Section 3.4.2 will look at this approach in more detail. Another useful function is the concordancer, *Concord*, which helps to illustrate any particular word or phrase that I would like to look at in context. With

KeyWords I can find the keywords or key sequences in a text by automatically working on the basis of statistical comparisons of frequency lists. This will be further discussed in 3.4.3.

WMatrix is another software tool used for corpus analysis and comparison in this study. It was developed by Paul Rayson in the REVERE[4] project at Lancaster University. It provides a web interface to CLAWS (Constituent Likelihood Automatic Word-tagging System) developed by Garside and Smith (1997) for part-of-speech (POS) annotation and the USAS (UCREL Semantic Analysis System) tagger (Rayson et al., 2004) for semantic annotation, which automatically assign parts-of-speech and semantic domains respectively to each word or multi-word expression in a corpus (Rayson, 2008).[5] It also contains the standard corpus linguistic functionality such as frequency lists, keywords and concordances; however, it extends the keywords method to key grammatical categories and key semantic domains, which *WordSmith Tools* cannot do. Section 3.4.3 will provide a more detailed discussion of this issue.

3.4.2 Frequency-driven approach

Frequency, referring to "the arithmetic count of the number of linguistic elements (i.e., tokens)" is "the most quantitative data a corpus can provide (McEnery et al., 2006, p. 52). It offers an "immediate snapshot of the characteristics of a particular language variety" (Harvey, 2008, p. 117) and it plays "a prominent role in how lexical items are employed in discourse" (Schmitt, 2010, p. 63). It also has the

[4] Further details of the project can be found at http://www.comp.lancs.ac.uk/computing/research/cseg/projects/revere/

[5] Further details of the software and tagsets employed can be found at http://ucrel.lancs.ac.uk/claws/ and http://ucrel.lancs.ac.uk/usas/.

benefit of being more systematic in the identification of linguistic elements and somewhat less subjective than other approaches (Adolphs & Durow, 2004; Teubert, 2005). Sinclair (1991) noted that "anyone studying a text is likely to need to know how often each different word form occurs in it" (p. 30). A frequency-driven approach therefore is employed throughout this research to obtain an insight into how language is used in intercultural communication.

Table 3.3
Distinctive List Contrasting Speech and Writing

Word	Spoken	Written	Word	Spoken	Written
er	8,542	11	***the***	39,605	64,420
you	25,957	4,755	***by***	1,663	5,493
I	29,448	6,494	***of***	14,550	31,109
yeah	7,890	17	***however***	90	664
know	5,550	734	***thus***	8	228
think	3,977	562	***as***	1,558	3,174
mean	2,250	198	***also***	556	1,328
just	3,820	982	***while***	156	543
okay	950	7	***most***	199	607

Research has shown that frequency facilitates enquiry across different corpora, different language varieties and different contexts of use (Adolphs, 2006, 2010; Baker, 2006; Leech, Rayson & Wilson, 2001; O'Keeffe et al., 2007, 2011). For example, Leech et al. (2001) contrasted the word frequencies between spoken and written texts based on the British National Corpus (BNC). A number of distinctive differences between them adapted from their list are shown in Table 3.3, with the rounded frequencies (per million word tokens). It can be easily noticed that some items, such as *er, you, I, yeah, know* and *okay,* are more common in spoken data, while some words, such as *the, however, by, of* and *also*, are used much more frequently in written discourse. This therefore shows people's preferences of lexical choices in different genres of discourse. In this respect, frequency of

vocabulary use provides valuable evidence for understanding the human processing of language (Leech et al., 2001).

It is also interesting to note that the words *know, think* and *mean* are used considerably more frequently in speaking. This is probably the result of the high-frequency use of spoken interpersonal markers, such as *you know, I think* and *I mean*. Table 3.4 shows the 10 most frequent two-word, three-word and four-word recurrent sequences generated by Adolphs (2006, p. 42) from CANCODE, a five-million-word corpus of spoken discourse. It is clear that many of the top 10 sequences in both two-word and three-word units include the words *know, think* and *mean*. It seems therefore that many high-frequency words often appear in recurrent sequences of words. This thus appears to present one of the weaknesses of such a word count. That is, it is usually based on orthographic words and seldom captures the frequencies of language patterns that co-occur (Schmitt, 2010, p. 66). As a result, this present study focuses not only on individual words but also on recurrent continuous sequences, namely multi-word sequences. This will be discussed in more detail in section 3.5.3.

Table 3.4
Ten Most Frequent Two-word, Three-word and Four-word Sequences in CANCODE

rank	two-word units	three-word units	four-word units
1	*you know*	*I don't know*	*you know what I*
2	*I mean*	*a lot of*	*know what I mean*
3	*I think*	*I mean I*	*innit isn't it*
4	*in the*	*I don't think*	*I don't know what*
5	*it was*	*do you think*	*the end of the*
6	*I don't*	*do you want*	*at the end of*
7	*of the*	*one of the*	*do you want to*
8	*and I*	*you have to*	*a bit of a*
9	*sort of*	*it was a*	*d'you do you*
10	*do you*	*you know I*	*do you know what*

Although frequency information is one important insight into language use, it is suggested that this should be interpreted with caution. One problem is that using computer programmes to generate wordlists does not consider the sense of a word (Leech et al., 2001; Adolphs, 2006; McEnery et al., 2006). That is, the programme is probably not able to distinguish between different meanings of the same word. We cannot, for example, exactly know if the word *park* means a public area maintained for recreational and ornamental purposes, whether it refers to cars, or whether it has other different meanings based only on frequency data of isolated lexical items excluding their context of use. As a result, researchers have been advised to further examine the concordance lines, which offer "a chance to see any word or phrase in context – so that you can see what sort of company it keeps" (Scott, 2007, p. 2). In this present study, for example, the word *Chinese* can be found with a high frequency in the Taiwanese participants' discourse. It may not be possible to know whether it means the Chinese language, people, food, a school subject or is used as an adjective to describe something that resembles a Chinese style. The following output shows a random selection of 10 lines for *Chinese* in concordance lines.

haha I can only speak	*Chinese*	My English is not good, so I don't
I'll cheer. (BT01), do you learn	*Chinese*	? I like to learn other languages
I love PE, art , music , English ,	*Chinese*	.... don't like math and science.
I go home at 4:55 p.m. I have	*Chinese*	English, math, science and PE lessons
That is about Moon festival. I like	*Chinese*	New Year very much. My family come
My favourite festival is	*Chinese*	New Year, we have six days holiday in
from my mother. My favorite festival is	*Chinese*	New Year. All my relatives will get
red envelopes. I hope that every day is	*Chinese*	New Year!! I want to introduce all of
Because we do many things during	*Chinese*	New Year. For example , everybody
It is my grandpa 's house. I like	*Chinese*	New year!! I like New Year , because

From this concordance output, some of the instances of *Chinese* are referred to as a school subject or a language, while many of them indicate the Taiwanese

students' tendency to use *Chinese* with *New Year* to describe one of the popular holidays in Taiwan. Therefore, it appears that the concordance can provide information about the context of use for particular items. It also has applications to language teaching, providing learners with information about word use and how the same word can sometimes have more than one meaning (O'Keeffe et al., 2007; Reppen, 2010). As a result, the data analysis for this study is not only based on quantitative data that frequencies provide, but also extends this information to further qualitative analysis using concordance lines.

3.4.3 Keyness Approach

Although the frequency lists provide a foundation for understanding the characterisation of different text genres, keyness analysis, which works on the basis of statistical comparisons of frequency lists, constitutes a more sensitive quantitative measure of linguistic features (McCarthy & Handford, 2004, p. 174). It is thus better suited to highlighting the main elements that are characteristic of a specific collection of texts. The procedure of keyness analysis works by comparing the actual observed frequency of each item in the target corpus with its equivalent in the reference corpus (Adolphs & Lin, 2010; Baker, 2006; Scott, 2010). As a result, this analysis gives a measure of saliency and it therefore serves as a useful tool for directing researchers to significant lexical differences between texts (Baker, 2006, p. 125).

A traditional way to explore particular linguistic features across texts is often attributed to the works done by Biber (1988, 1995). He employed a multi-dimensional methodology, especially factor analysis, to analyse the distribution of linguistic features across texts and the systematic co-occurrence of

patterns among linguistic features. Although it offers a reasonably robust basis for the linguistic specification of a genre, Lee (2001) criticised the approach as there are many questions surrounding the statistical validity, empirical stability and linguistic usefulness of the linguistic dimensions from which Biber derives these text types, or clusters of texts sharing internal linguistic characteristics (p. 40). Tribble (2000, p. 78) also identifies the practical difficulties of Biber's method for language teaching, claiming that the approach will not give students a direct insight into the ways in which expert writers draw on (often highly patterned and conventional) lexical resources, and it requires a POS marked up version of the research corpus prior to analysis, which is not generally available to classroom teachers. However, these problems can be solved by using corpus analytical tools without extensive training or effort. *WMatrix*, for example, provides users with a web-based tool to create and process the tagging rapidly and automatically, and concordance lines present how particular lexical items are used in context. This means that texts or groups of texts can be easily compared to elucidate salient lexical and POS features in each group.

Xiao and McEnery (2005) compared keyness analysis and Biber's multi-dimensional approaches via a case study of casual conversation, speech and academic prose in modern American English. They found that the keyness technique could capture important genre features revealed by multi-dimensional analysis, but that it also constituted a better representation of a text and was less demanding since Biber's approach requires sophisticated data extraction and statistical analysis (p. 77). As a result, keyness analysis has been widely used in many areas of applied linguistics research, particularly with regard to the identification of language variation, styles and genre. For example, Baker (2006)

compared the FLOB corpus of British English with the FROWN corpus of American English and found a number of underlying cultural differences between the two language varieties. Harvey (2008) compared a corpus of online adolescent health communication with a corpus of general spoken and written English and identified a set of keywords. Harvey then examined these key items in their original discourse contexts to gain a deeper understanding of both teenagers' experiences of physical and mental health problems and the linguistic particularities of online health communication. Culpeper (2009) analysed the speech of six characters in Shakespeare's Romeo and Juliet by comparing each character's speech against the speech of all the other characters. Such analysis helps to produce key items that reflect the distinctive styles of each character compared with the other characters in the same play. In addition, Ooi, Tan and Chiang (2007) applied keyword analysis to the examination of Singaporean English on personal weblogs and compared it with large corpora of spoken and written English. A deeper understanding of the various cultural identities, gender differences and other sociolinguistic variables was obtained.

Although previous studies have extensively employed keyness analysis in the investigation of different contexts of language use, few, if any, have applied this method for understanding intercultural discourse and informing English teaching and learning. This research thus intends to demonstrate the pedagogical merit of keyness analysis as a useful tool for learners to obtain a direct insight into the ways in which they and expert writers draw on lexical resources, and will also act as a means to inform teachers and materials developers in designing courses for adolescent intercultural online interaction.

There are typically two strategies used to conduct keyness analysis. One approach is used to compare two corpora of similar size. This comparison by using corpus analytical programmes such as *WordSmith Tools* provides two lists of key items that are salient (with a particularly high frequency) in each corpus as compared with the other as the norm. The other approach to keyness analysis can be done in the comparison of a small corpus with a much larger reference corpus, which generates a list of key items that are unusually frequent and a shorter list of items that are unusually infrequent in the smaller corpus (Bachmann, 2011; Scott, 2010).

The current study applies the keyness method to three levels of analysis: keywords, semantic domains and parts-of-speech. Key semantic analysis is used to compare and contrast the small specialised BATTIC with large reference corpora of online and spoken discourse in order to identify the themes that young people are particularly concerned with in intercultural online communication. In addition, the key grammatical categories that the Taiwanese participants used significantly differently from the British participants are identified by applying key POS analysis to the comparison of the discourse by the two groups of participants.

3.4.3.1 Keywords analysis

Keywords refer to those items that occur unusually frequently (positive keywords) or unusually infrequently (negative keywords) in comparison with some kind of reference corpus (Scott, 2010). They are identified on the basis of statistical comparisons of words in a text with a reference set of words, and consequently any word that is found to be outstanding in its frequency in the comparison is considered key (ibid.). In the present study, concerned with identifying the key

lexical items of adolescent online and face-to-face intercultural communication, the frequency lists of BATTICC-O and BATTICC-F were first generated and then compared with the frequency lists of two reference corpora: CANELC[6] [Cambridge and Nottingham e-Language Corpus] and the BNC Sampler Spoken respectively. CANELC contains approximately 500,000 words of asynchronous CMC discourse, collected from a variety of online communication sources including online discussion boards, blogs, Tweets and e-mails; the BNC Sampler Spoken is a subset of the British National Corpus (BNC), including 982,712 words of spoken conversation transcription. The online and spoken nature of these corpora resembled the computer-mediated and face-to-face interaction respectively in this project and thus made them suitable resources for a comparative study. This comparison helps determine the items that occurred with a significantly higher or lower frequency in the target corpus than in the reference corpus.

3.4.3.2 Extending keywords to key domains

Although the keyword analysis highlights those lexical items that are salient in the target corpus when compared with a reference corpus, there are a number of practical limitations. One is identified by Baker (2004), who claims that keywords only focus on lexical differences, rather than semantic, grammatical or functional differences (p. 354). Baker (2006) further points out that it is sometimes the case that some words do not occur often enough to make a sufficient impact, and they are therefore not included in the keyword list and tend to be overlooked. He

[6] CANELC is an ongoing corpus development project with the aim of collecting one million words of e-Language data. The corpus was developed as a joint project between the University of Nottingham and Cambridge University Press with whom sole copyright resides.

exemplifies this with the notion of "largeness" in a text which is instantiated through various synonyms used by writers such as *big, large, huge, great, giant* and *massive*. However, since none of these occurs individually with high frequency, these would actually be excluded from the keyword list since the semantic similarities between words are not explicitly taken into account in a keyword analysis. In addition, keyword analysis usually generates far more keywords than it is possible for the researcher to analyse (Berber-Sardinha, 1999, p. 5; see also Rayson, 2008). Key domain analysis, therefore, would reduce the number of key items that the researcher should examine and highlight the most significant categories. As a result, conducting the keyness method, usually applied at the word level, to two additional levels, namely parts-of-speech and semantic domains, could provide valuable information on the understanding of particular discourse types, and thereby avoid the limitations of wordlists and keyword lists.

To this end, Rayson's web-based suite of tools constituting *WMatrix* (Rayson, 2008; see also http://ucrel.lancs.ac.uk/wmatrix) was used to easily and rapidly annotate the datasets for both grammatical and semantic categories, and then to identify which categories were key. *WMatrix* employs CLAWS (Constituent Likelihood Automatic Word-tagging System) and the USAS (UCREL Semantic Analysis System) tagger for POS and semantic annotation, which automatically assign POS and semantic fields respectively to each word in a corpus (Rayson, 2008). Frequency data of each tagged category were then analysed to identify the most significant features of texts. Such a procedure of analysis is labeled by Rayson as data-driven since it starts from frequencies in the language data rather than the researchers' assumptions about language features. However, this may not be purely data-driven as the analytical framework for text annotation is

pre-constructed; as noted by Sinclair (2004), one of the problems of tagging is the perceived necessity for human intervention (p. 191). Sinclair further reminds us that automated tagging sometimes pays no attention to the clarity of the categories in the data, and when using marked-up texts, the data are simply processed through the tags; that is, anything the tags are not sensitive to will be missed (p. 191). However, Rayson (2008) reports that the accuracy rates of the POS and semantic tagging are approximate 96-97% and 91% respectively. As a result, while users can be largely confident in the tagging accuracy, the interpretation of the results should take into account that possible tagging errors and manual checking of concordances are therefore necessary (Culpeper, 2009; Rayson, 2008).

3.4.3.3 Tests of statistical significance: Log-likelihood (LL) test

Although there are a number of ways of calculating statistical significance for keyness, the Chi-square test and Log-likelihood (LL) analysis (Dunning, 1993; Oakes, 1998) are probably the most common ones, which can be chosen from *WordSmith Tools*. They both compare the differences between the observed values and the expected values. The greater the difference between the two values, the more likely it is that the relationship between the two items is not due to chance, but that other factors influence their relationship (Adolphs, 2006; McEnery et al., 2006). In this way, the items that are characteristic in a target corpus can be generated by chi-square and LL tests. However, the LL test is preferred in this study as it "does not assume that data are normally distributed" (McEnery et al., 2006, p. 55). Dunning (1993, p. 65) notes that "using the normal distribution overestimates the significance" and "the use of likelihood ratios leads to very

much improved statistical results", particularly with very small volumes of text. Evert (2008) also compared the two measures for the collocation studies and found that the LL-ratio performed at a much higher precision rate than the Chi-square test. Adolphs (2006) points out another problem with the Chi-square calculation. She suggests that:

> It can produce distorted statistics if the expected frequencies of individual items are low, particularly when words in a small corpus are compared with a large reference corpus. It is therefore not advisable to the chi-square calculation under those circumstances. (p. 50)

It appears that the LL test is better suited for the present study as the corpus data is relatively small. However, one might argue that in any comparison involving corpora of markedly different sizes, frequencies often need to be normalised to a common base. Rayson (2008, p. 527) claims that there is no need to normalise the figures before doing the LL analysis as the calculation for the expected values takes account of the size of the two corpora. This is very important since claiming statistical significance is based on "not only the *magnitude* of the result but also the *size of the sample* investigated (Dörnyei, 2007, p. 210). Rayson (2008) further presents the formula for calculating the LL statistic, based on the contingency table as shown in Table 3.5.

Table 3.5
Contingency Table for Log-likelihood Calculation

	Corpus one	Corpus two	Total
Frequency of a word	a	b	a+b
Frequency of other words	c – a	d – b	c + d – a – b
TOTAL	c	d	c+d

In Table 3.5, the values *a* and *b* are the observed values (O), namely raw frequencies of a word in two corpora; the values *c* and *d* correspond to the total numbers of words in the two corpora respectively. The expected values (E) are first calculated based on the following formula:

$$E_i = \frac{N_i \sum_i O_i}{\sum_i N_i}$$

At this point, according to Table 1, N1 = c, and N2 = d. So, for this item, the expected values (E) would be:

$$E_1 = \frac{c(a+b)}{(c+d)}, \text{ and } E_2 = \frac{d(a+b)}{(c+d)}$$

The log-likelihood value is then generated according to the following formula:

$$LL = 2\sum_i O_i \ln\left(\frac{O_i}{E_i}\right)$$

More precisely, in this case, the formula can be expanded as follows:

$$LL = 2\times\left(\left(a\times\ln\left(\frac{a}{E_1}\right)+b\times\ln\left(\frac{b}{E_2}\right)\right)\right)$$

$$\Rightarrow LL = 2\times\left((a\times(\ln(a)-\ln(E_1))+b\times(\ln(b)-\ln(E_2)))\right)$$

To exemplify these formulae, for example, assuming the frequencies of a word in Corpus one and Corpus two are 10 and 20 respectively, and the size of the two corpora is 400 and 600 words respectively, this case can be illustrated with the following numerical values, where a=10, b=20, c=400 and d=600. Firstly, the expected values (E) can be calculated as follows:

$$E_1 = \frac{400\,(10+20)}{(400+600)} = 12, \text{ and } \; E_2 = \frac{600\,(10+20)}{(400+600)} = 18$$

Then the LL value can be generated as follows:

$$LL = 2\times\left((10\times(\ln(10)-\ln(12))+20\times(\ln(20)-\ln(18))\right)$$
$$\Rightarrow LL = 2\times\left((10\times(2.3025-2.4849)+20\times(2.9957-2.8903)\right)$$
$$\Rightarrow LL = 5.68$$

In such a case, the LL value in this keyness analysis is 5.68. In the computer programmes used in the present study, the procedure is automatically applied to each single item in the two frequency lists retrieved from two corpora, while in POS or semantic domain analysis, the comparison is done with tag frequencies rather than word frequencies (Rayson, 2008). The most significant items or tags in corpus one as compared to corpus two would then emerge, and they are sorted by the resulting LL values. In this case, the larger the LL value, the more significant the relative frequency difference between the two corpora is. Nevertheless, the question remains how large the value needs to be in order to be considered statistically significant and how confident we can be to make this claim. The latter is the so-called *p* value, namely the probability coefficient, which normally ranges from 0 to 1. It has been noted that in social science studies a hypothesis can be accepted only when the level of significance is less than 0.05 (i.e., $p < .05$), which means that one must be more than 95% confident that the differences observed are not due to chance (Dörnyei, 2007; McEnery et al., 2006). In the LL test, the critical values with 1 degree of freedom (or d.f.) are 3.83, 6.64, 10.83 and 15.13 for the significance levels of 0.05, 0.01, 0.001 and 0.0001 respectively (McEnery et al., 2006; Rayson, 2008). In this regard, the *p* value close to 0 indicates a statistical significance.

This section has focused on the keyness approach by considering three levels of analysis: keywords, semantic domains and parts-of-speech. While extending the keyness approach to grammatical and semantic domains working with tagged categories can contribute to a greater understanding of learner grammar and lexis, the discourse analytical approach working with authentic texts rather than tagged data, an approach strongly endorsed by Sinclair (2004), can add greater detail and depth of description of language used in an intercultural setting. The next section will demonstrate the second approach of the thesis from a discourse analytical point of view.

3.4.4 Discourse analytical approach

The discourse analytical approach looks at language use in its social context, drawing specifically on studies of the relationship between texts and contexts in which they arise and operate (McCarthy, Matthiessen, & Slade, 2010). In this thesis, initial quantitative analysis is employed to inform further qualitative analysis, an approach that is particularly appropriate to smaller corpora, as noted by Evison (2008). This approach will further demonstrate a describable patterning of language used in different modes of intercultural communication and examine the pragmatic and discourse functions of linguistic items, which may not be easy to uncover simply by using a keyness approach.

While traditionally the primary focus in previous studies of learner grammar and lexis was on written language, CMC and spoken discourse are often considered as formless and ungrammatical. However, a number of studies on the analysis of CMC (e.g., Crystal, 2006, 2011; Herring, 2013) and spoken discourse (e.g., Carter & McCarthy, 2006; Cutting, 2011) have shown that electronic and spoken

discourse do have a consistent and describable language patterning and exhibit a highly elaborate organisation that is grammatically intricate. In doing so, two analytical frameworks are employed for analysing CMC and spoken data, which will be discussed respectively in the following subsections.

3.4.4.1 Analysing CMC discourse

As discussed in Chapter 2, the grammar in text-based computer-mediated communication (CMC) is somewhat different from its usual sense as in speech or written discourse. Crystal (2006) proposed the vision of his term *Netspeak* as "speech + writing + electronically mediated properties" (p. 51) in that online discourse is not only an aggregate of spoken and written features, but the adaption to the specific medium and the available technology. As such, a number of features of CMC discourse are often linked to the concept of linguistic economy. Werry (1996) identified various strategies of economical language use in Internet Relay Chat (IRC), a form of real-time multi-participant Internet text communication. These economical features include abbreviations, ellipsis and orthographic reduction (e.g., *bb ppls* for *bye bye peoples*). Similarly, Cho (2010) examined the features of CMC discourse that were viewed as being indicators of linguistic economy, including abbreviation/clipping, use of lower case in place of upper case and omission of pronouns, articles, the verb *be*, essential punctuation, existential *there,* etc.

In addition, the language of CMC has been shown to exhibit a wealth of non-verbal cues, providing information and expressing emotional intimacy, which are particularly unique to face-to-face communication, in order to compensate for missing visual and aural cues. Crystal (2006, 2011) notes the existence of several

types of paralinguistic and prosodic cues available in online discourse, such as asterisks, capitalised words, repeating letters and exclamation points. Frehner (2008) also claims that "computer-mediated communication can be considered incomplete for its lack of paralinguistic cues" (p. 76). Recent evidence on CMC studies has demonstrated different types of cues in online communication and their unique linguistic features. Kalman and Gergle (2010), for example, examined the use and role of prosodic cues, i.e., vocal spelling and repeating punctuation marks. It was found that CMC users apply these creatively to achieve a host of effects which are often analogous to those achieved through paralinguistic cues in spoken conversation (p. 2). Riordan and Kreuz (2010) analysed nine types of nonverbal cues commonly used in CMC, including capitalised words, vocal spelling, repeating punctuation, emoticons, angled brackets, underscores, tildes and curly brackets. In addition, Garrison et al. (2011) examined emoticons in their own right as conventions of instant messaging discourse, including frequency, type and placement. While the variation of linguistic and paralinguistic features of CMC has been explored in a range of previous studies, their pragmatic meanings and functions do not seem to be an area upon which focus has concentrated, thereby limiting the understanding of different types of CMC features in context. Also, the cultural differences with regard to what kinds of features are used and how frequently they are employed by native and non-native young learners of English has not been fully explored.

Although there is not a single grammar for all varieties of CMC language, in this study I select the four most distinctive features of CMC that rarely occur in traditional written grammar or major dictionaries. As shown in Table 3.6, the four categories are: (1) use of the upper and lower cases, (2) nonconventional spelling,

(3) emoticons and (4) punctuation use. These features are also included in the most up-to-date research on the analysis of CMC discourse (e.g., Herring, 2013; Kalman & Gergle, 2010; Riordan & Kreuz, 2010).

The first feature that I look at is the use of the upper and lower cases, in which two phenomena can be widely found in online discourse: (a) the use of lower case instead of upper case and (b) nonstandard capitalisation. The use of lower case instead of upper case allows exploring linguistic economy in CMC. On the other hand, examples of nonstandard capitalisation often require an additional effort from CMC users, although economical language use by reducing the use of capital letters to a minimum is pervasive in CMC. These analyses will illustrate how capitalisation and minusculisation are employed in the context and to what extent cultural differences exist.

Another distinct feature of CMC is nonconventional spelling, since in online communication spelling practices often suggest "loosened orthographic norms" (Herring, 2013). The term nonconventional spelling in this study is referred to as spelling that does not correspond to the orthography of major dictionaries. It mainly includes two aspects: (1) abbreviation, acronyms and substitution and (2) vocal spelling. Words that are shortened by removing one or more phonemes or morphemes are classified as abbreviation (e.g, *pls* for *please*); acronyms are words formed from the initial letters in a set phrase or series of words (e.g., *BTW* for *by the way*); substitution can be a word or part of a word with an alphabetic name (e.g., *u* for *you*) or a number (e.g., *2morrow* for *tomorrow*). As noted in Chapter 2, they can be commonly seen in different modes of CMC, and they are likely made by users to economise on typing effort, mimic spoken language features or

express themselves creatively (Herring, 2003, 2013). I will also investigate the use of vocal spellings, focusing on emulated prosody and onomatopoeic exclamatory spelling. The former represents prosody or nonlinguistic sounds (e.g., *'calling voice' helloooo)*, and the latter includes onomatopoeia such as *hehe, haha* or *ahhh*, indicating an entry that reproduces a sound.

Furthermore, the third category is the use of sequences of keyboard characters, namely emoticons (e.g., *;), :D, XD*), "a visual representation constructed through the use of a series of typographic symbols" (Garrison et al., 2011, p. 112), which are common in CMC to show a textual face demonstrating a writer's mood. Crystal (2006) describes emoticons as "combinations of keyboard characters designed to show an emotional facial expression" (p. 39). It appears that emoticons generally serve as emotion indicators in CMC. However, Dresner and Herring (2010) claim that not all of the facial emoticons are used to express emotion. For example, *;-P* representing a face with the tongue sticking out does not seem to represent a sign of a specific emotion; it does, however, indicate that the user is joking, teasing or otherwise not serious about the content of the message. Such use, therefore, may not contribute to the propositional content (the locution) of the language used, nor indicate emotion. Rather, emoticons help convey the speech act, or they clarify what the user intends. As such, they can serve as indicators of non-emotional meanings or illocutionary force. In this analysis, the emoticon is examined in its context with a view to identifying the functional use and the pragmatic meaning in the given context.

In addition, users of CMC usually hold a rather lax attitude towards the use of punctuation (Frehner, 2008). It is often considered "apparently carefree" (Baron,

2009, p. 914) and "creative" (Thurlow, 2001, p. 288). Repeating punctuation marks is one of the significant features. Previous studies have shown that the repetition of exclamation points and question marks is the most prevalent in CMC (e.g., Cho, 2010; Frehner, 2008; Kalman & Gergle, 2009, 2010; Riordan & Kreuz, 2010). I also examine the use of apostrophes, which mark the ellipsis of one or more letters, as in the contraction of *do not* to *don't*. They can also be the marking of possessive case, as in *my mother's hair*. In CMC, however, there is a remarkable tendency to omit them to save typing effort. Thurlow (2001) points out that CMC discourse is often "blamed for the 'death' of the apostrophe" (p. 289). Previous studies have shown that 18-43% of the apostrophes are omitted in the context of e-mail and text messaging (Frehner, 2008).

Table 3.6
Discourse Analytical Framework: E-grammar

Features of e-grammar	**Examples**
Use of the upper and lower cases	Nonstandard capitalisation (e.g., *I LOVE* the picture!) Inconsistent use of capitalisation/minusculisation
Nonconventional spelling	Abbreviation (e.g., *pls* for *please*) Substitution (e.g., *4* for *for, 2day* for *today*) Vocal spellings, representing prosody or nonlinguistic sounds (e.g., calling voice *hellooo*) Onomatopoeic spelling (e.g., *haha, hehe, ohhh*)
Emoticons	Emoticons, or sequences of keyboard characters (e.g.,*:D,:P, XD*)
Punctuation	Repeating punctuation (e.g., *!!!, ???*) Apostrophe omission (e.g., *im, dont*)

3.4.4.2 Analysing spoken discourse

With the intention of examining the linguistic features of spoken discourse in face-to-face interaction between Taiwanese and British participants, the

framework shown in Table 3.7 was employed. As there is not one standard terminology for describing spoken grammar, the framework used for the current study is mainly adapted from Carter and McCarthy's (2006) analytical categories of spoken language. Some of the features in their work are also described by Biber et al. (1999), Leech (2000), Thornbury and Slade (2006), Cullen and Kuo (2007), Mumford (2009) and Cutting (2011), as discussed in Chapter 2. Since the grammar of informal and conversational English, rather than that of spoken discourse characteristics of more formal settings such as debates or speeches, is one of the main focus points for this study, Carter and McCarthy's (2006) approach is suited to highlighting the significant lexical and grammatical features in face-to-face communication discourse. In addition, their framework was based on analysis of the Cambridge International Corpus (CIC), which is mainly composed of British English. This is in concordance with the BATTICC for the purposes of this research.

In this study five main dimensions of spoken language are examined: (1) vague expressions, approximations and hedging, (2) situational ellipsis, (3) headers and tails, (4) pauses, repeating and recasting and (5) discourse marking, which constitute lexical, syntactic and discourse features of spoken discourse. These are chosen because they typically feature very rarely in written discourse but are significant features of informal spoken grammar. The analytical framework is also applied to the analysis for online communication, BATTICC-O, in order to examine the extent to which online discourse presents the features of spoken grammar. As shown in Table 3.7, I firstly examine the lexical features of BATTICC-F, including vague expressions, approximations and hedging. Vague expressions for this study involve the use of words and phrases such as *sort of,*

kind of, or anything, and stuff, and so on and *and things like that*, which deliberately refer to people and things in an imprecise way (Carter & McCarthy, 2006; O'Keeffe et al., 2007). Approximations, which are often described as vague language used with numbers and quantities, are included in this category, as in "*around* six", "*a couple of* days ago". Carter and McCarthy (2006) claim that vague expressions and approximations are motivated and purposeful and they are often marks of sensitivity and the skills of a speaker (p. 202).

Syntactic features include situational ellipsis, headers and tails. Situational ellipsis involves the deliberate omission of items such as subject pronouns and verb complements, which might not be necessary in utterances as these contain enough information for the purpose of the conversation (Carter & McCarthy, 2006; Thornbury & Slade, 2006). That is, the omission of items is usually "retrievable from the immediate situation" (Cullen & Kuo, 2007), as the following examples from Carter and McCarthy (2006, p. 181) illustrate:

A: Don't know what's gone wrong here.
B: Oh. Need any help?

In this case, speaker A's utterance is understood as *I don't know what's gone wrong here*, with the ellipsis of the subject pronoun *I*; speaker B omitted *Do you* prior to *Need any help?*. Although the sentences appear to not be grammatically correct, the listeners can still retrieve the meanings from the contexts. In this analysis, fixed expressions are not included in the discussion, such as *Good job* for *It's a good job* and *See you soon* for *I'll see you soon*. Since such expressions are widely used in daily life, I excluded them from the category of situational ellipsis.

In addition, as discussed in Chapter 2, word order is more flexible in speech than written discourse since "it is constructed in real time and follows the order of ideas emerging from a speaker, which may override grammar rules" (Carter & McCarthy, 2006, p. 172). Headers and tails as a result may alter the word order of traditional written grammar. Headers (e.g., *the teacher*, he is very nice) are often used to emphasise what the speaker thinks to be particularly important (Carter & McCarthy, 2006). Tails (e.g., They're really nice, *my teachers*), on the other hand, normally occur after clauses and are commonly used to "clarify or make explicit something in the main clause" (ibid., p. 194). In addition, Carter et al. (2011) describe them as "interpersonal grammar", which serves an interpersonal function that is listener-sensitive (p. 82). In this regard, the speaker attempts to involve the listener by expressing his or her own personal feelings and attitudes.

Table 3.7
Discourse Analytical Framework: Spoken Grammar

Lexical features	**Syntactic features**	**Discourse features**
Vague expressions *sort of thing, and stuff, or anything, something like that* **Approximations** *about, a couple of, loads of* **Hedging** *sort of, a bit, a little bit*	**Situational ellipsis** *(Do you) Need any help?* **Headers** *The teacher, he is very nice.* **Tails** *They're really nice, my teachers.*	**Pauses, repeating and recasting** *Erm.. I'm I'm not sure.* **Discourse marking** *interpersonal, referential, structural and cognitive discourse markers*

Concerning discourse features, I first examine pausing, repeating and recasting. Pausing can be unfilled, which is simply a silence, or filled, which is identified by

a vocalisation such as *er* and *erm* (Carter & McCarthy, 2006). These items can mark a hesitation on the part of the speaker and are typically used to fill pauses between the elements of utterances. Repeating can be one word (e.g., *I'm I'm not sure*) or phrases/sentences (e.g., *We're meant to be talking ..er.. we're meant to be talking about the walk.*). Recasting is identified as instances of reformulating words, phrases, clauses or sentences (e.g., Before we start ... *before we go into that level of detail*, I'm going to write it on the OHP).

Moreover, I look at the use of discourse markers (DMs), which play a predominant role in spoken real-time interaction as discussed in Chapter 2. I employ the multi-category framework proposed by Fung and Carter (2007), which embraces a functionally-based account for the categorisation of DMs in spoken language. Their work contains four main functional domains: interpersonal, referential, structural and cognitive categories. Interpersonal DMs mark shared knowledge (e.g., *you see, you know*, etc.), indicate the attitudes of the speaker (e.g., *well, I think, you know, sort of, like*, etc.), show responses (e.g., *okay, oh, right/alright, yeah, I see*, etc.) and express a stance towards propositional meanings (e.g., *basically, actually, absolutely, exactly*, etc.). Referential DMs "work on a textual level and mark relationships between verbal activities preceding and following a DM" (p. 415). One common example of this type is the use of conjunctions, which indicate cause (e.g., *because, cos*), contrast (e.g., *but, however*), coordination (e.g., *and*), consequence (e.g., *so*), disjunction (e.g., *or*), digression (e.g., *anyway*) and comparison (e.g., *likewise, similarly*).

Structural DMs serve to indicate sequential relationships (e.g., *first, second, next, then, finally*, etc.), topic shifting (e.g., *so, now, and what about*, etc.), and

signposting of opening and closing of topics (*now, OK, right, by the way, let's start*, etc.). Some other DMs mark a cognitive process of discourse. For example, *well* often indicates the thinking process when the speaker does not have an immediate response (Fung & Carter, 2007). Aijmer (2011) describes it as "primarily a "mental state" interjection" which can be associated with the speaker's deliberation (p. 235). Biber et al. (1999) also state that *well* "appears to have the general function of a 'deliberation signal', indicating the speaker's need to give (brief) thought or consideration to the point at issue" (p. 1086). As such, the use of *well* can allow speakers to buy time for planning and processing. Similarly, functions involving cognitive processing also include reformulation (e.g., *I mean, that is, in other words*), elaboration (e.g., *like, I mean*) and thinking process (e.g., *well, I think, I see*). It should also be noted that each DM may perform more than one of these functions, as can be seen in the examples above. In this analysis, the use of each DM in its discourse contexts is examined to identify the primary function in my datasets.

3.4.5 Analysis of recurrent multi-word sequences

The third approach employed in this thesis is based on a multi-word sequence perspective. Kjellmer (1994) suggests that "[t]here is no doubt that natural language has a certain block-like character. Words tend to occur in the same clusters again and again" (p. ix). Previous research has indeed highlighted the fact that both written and spoken discourse contains a large proportion of highly recurrent sequences of words, reflecting the phrasal nature of the English language (Adolphs, 2006; Biber, Johansson, Leech, Conrad, & Finegan, 1999; Greaves & Warren, 2010; Nation & Webb, 2011; Schmitt, 2010, 2013; Wood, 2010; Wray, 2002, 2013). Biber et al. (1999), for example, illustrate that two-word

(e.g., *I think*), three-word (e.g., *a lot of*) and four-word (e.g., *what do you think*) recurrent sequences made up nearly 45% of the spoken conversation and approximately 21% of the academic written discourse they studied (the cut-off was set at a frequency of 20 occurrences per million words). Erman and Warren (2000) also calculated that recurrent multi-word sequences constituted 58.6% of the spoken corpus and 52.3% of the written discourse analysed in their study. In addition, Foster (2001) analysed the transcripts of unplanned speech of English native speakers and found that 32.3% consisted of recurrent sequences, while in Hill's (2001) study up to 70% of language (spoken and written discourse) comprised fixed expressions. Adolphs and Durow (2004) examined the three-word recurrent sequences in EFL learners' interview transcripts and the percentage ranged from 9.55% to 20.98%. Despite the variation in the reported percentage of recurrent multi-word sequences encountered in these studies, they all indicate an observable tendency for particular items to co-occur in the written and spoken discourse of both native and non-native speakers of English, and for these co-occurrences to make up an appreciable proportion of authentic language use.

There has been a burgeoning field of research looking at multi-word sequences in different registers and settings, identifying different kinds of sequences and describing how they are employed in particular contexts. These contexts include academic writing (Chen & Baker, 2010; Simpson-Vlach & Ellis, 2010), university classroom teaching (Biber, Conrad, & Cortes, 2004), small group teaching contexts such as tutorials and seminars (Walsh, Morton & O'Keeffe, 2011), textbook discourse (Chen, 2010; Wood, 2010), and spoken interview discourse (Adolphs & Durow, 2004). Although multi-word sequences used in various

contexts have been extensively studied, relatively little research has focused on recurrent sequences in an intercultural setting and further compared their use in two important registers, namely computer-mediated communication (CMC) and face-to-face (FTF) interaction. Research in intercultural discourse increasingly represents a particularly important endeavour as it offers insights into language variety which reflects the social and cultural differences of the writers and speakers (Hanna & de Nooy, 2003; Liaw & Master, 2010).

3.4.5.1 What are multi-word sequences?

Research demonstrates that natural language use contains a considerable number of recurrent patterns (Biber et al., 1999; Carter & McCarthy, 2006; Ellis et al., 2008; Sinclair, 2001). This suggests that vocabulary tends to occur not simply as single words but rather "has a strong tendency to occur in multiple word phraseological units" (Schmitt, 2010, p. 117). There is a wide range of technical terms used to describe the phrasal nature of language, including *prefabricated patterns* (Hakuta, 1974), *routine formulae* (Coulmas, 1979), *lexical phrases* (Nattinger & DeCarrico, 1992), *lexical bundles* (Biber et al., 2004; Biber, 2009), *lexical clusters* (Wood, 2010), *recurrent continuous sequences* (Adolphs, 2006), *chunks* (De Cock, 2004; O'Keeffe, McCarthy, & Carter, 2007), *clusters* (Scott, 2010), *multi-word units* (Greaves & Warren, 2010; Nation & Webb, 2011) and *formulaic sequences* (Schmitt, 2004, 2010; Wray, 2002, 2013). This variety notwithstanding, the various terms all indicate an observable tendency for particular lexical items to co-occur in the written and spoken discourse of both native and non-native speakers of English, and for these co-occurrences to make up an appreciable proportion of authentic language use.

Wray (2002) defines a formulaic sequence as follows:

> A sequence, continuous or discontinuous, of words or other meaning elements, which is, or appears to be, prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar. (p. 9)

This definition shows that in many cases, sequences of words are stored in the mind like single words and processed as a chunk that can be retrieved holistically at the time of use. For example, in the case of a self-introduction, learners just need to retrieve the chunk *my name is ..., I am from ...* etc. instead of building the sentence word by word. This concept of holistic storage and retrieval of formulaic language has been demonstrated in a range of studies (e.g., Conklin & Schmitt, 2007; Tremblay & Baayen, 2010). Nevertheless, whether or not the recurrent elements are prefabricated in speakers' or writers' minds is still debatable. In this study, since the concept of holistic storage and retrieval proposed by psycholinguists is not the main focus of this current study, I use *multi-word sequence* as an umbrella term to cover all types of recurrent sequences of words, that is, "frequently occurring contiguous words that constitute a phrase or a pattern of use" (Greaves & Warren, 2010, p. 213). In this regard, a multi-word sequence does not necessarily have to be a complete grammatical structure or idiom. This definition is similar to Biber et al.'s (1999) definition of lexical bundles, which are defined as "recurrent expressions, regardless of the idiomaticity, and regardless of their structural status. That is, lexical bundles are simply sequences of word forms that commonly go together in discourse" (p. 990).

3.4.5.2 Identifying multi-word sequences

The present study investigates the recurrent multi-word sequences in TETCOC and BATTICC. A form-based, frequency-driven approach was employed for the identification of recurrent sequences using the corpus analytical programme *WordSmith Tools 5.0* (Scott, 2008). As has been discussed in 3.3.2, research has shown that frequency data facilitates enquiry across different corpora, different language varieties and different contexts of use (Baker, 2006; Leech, Rayson, & Wilson, 2001; O'Keeffe et al., 2007, 2011), so it is "an important parameter for detecting recurrent patterns" (Teubert, 2005, p. 5). In addition, I focus particularly on three-word sequences (e.g., *I don't know, I would like*). A unit size of three words per sequence was chosen because this includes sufficient contextual information for the assessment of units' discourse functions and is also analytically more manageable. Analysing two-word sequences (e.g., *of the, to be*) would include too many phrasal verbs and grammatical colligations that are not the main focus of the present study, although they offer access to both paradigmatic and syntagmatic features of language (Crossley & Salsbury, 2011). On the other hand, considering larger units, such as four or more words in the sequence (e.g., *at the end of the*), would reveal too few examples, although they might well offer more clues to the context of the sequences used than two and three-word units as shown in Biber et al.'s study. As a result, the use of three-word sequences is mainly examined to reveal the degree of preference of certain sequences in the British and Taiwanese participants' discourse. I compared these with the texts of the British participants in this project, as well as CANELC, a 500,000 word corpus of online discourse.

In order to obtain a deeper insight into the use of recurrent sequences over time in CMC and FTF interaction, the electronic messages are divided into three data

subsets according to the time of posting, resulting in three four-month phases. The highly recurrent three-word sequences retrieved from different phases of the program are then examined. I further compare the sequences in BATTICC-O and BATTICC-F and, as a reference for comparison, also list the high-frequency sequences in a general large corpus of online discourse (CANELC) and spoken discourse (CANCODE[7]). The online and informal spoken nature of these respective corpora resemble the computer-mediated and face-to-face interaction in this project and thus makes them suitable resources as reference corpora. The 50 most common three-word sequences retrieved from the four datasets are then inductively grouped into three central categories with regard to the discourse function that they serve in the context.

3.4.5.3 Multi-word sequences and functional language use

A range of studies have now demonstrated that multi-word sequences serve various types of discourse functions in language use (Biber, 2009; Biber et al., 2004; Nattinger & DeCarrico, 1992; Wood, 2010; Wray, 2002; Wray & Perkins, 2000). One overriding function of multi-word units may also be to facilitate efficient and effective communication (Wood, 2010). Nattinger and DeCarrico (1992) developed a taxonomy that captures three central functions served by what they called *lexical phrases*: (1) social interaction, (2) necessary topics and (3) discourse devices. In their framework, social interaction sequences are associated with social relationships, consisting of conversational maintenance (e.g., *excuse*

[7] CANCODE stands for Cambridge and Nottingham Corpus of Discourse in English, a five million word corpus of transcribed conversations in mainly informal spoken situations. It was established at the School of English, University of Nottingham, and is funded by Cambridge University Press, with whom the sole copyright resides. The corpus recordings were made in a variety of informal settings including shops, private homes, public places and educational institutions across Britain and Ireland. For further details of the corpus and its construction see McCarthy (1998).

me; how are you?), and functional meaning relating to conversational purpose, such as expressing politeness (e.g., *thanks very much*), questioning (e.g., *do you like X?*), requesting (e.g., *may I X?*), offering (e.g., *would you like X?*), complying (e.g., *of course*), responding (e.g., *oh, I see*) and asserting (e.g., *I think that X; there is/are/was/were X*). Necessary topics are phrases marking domain-specific topics that often feature in daily conversation, such as autobiography, shopping, food, school, time and location. For example, in autobiography, formulaic expressions such as *my name is*, *I am from* and *I'm X years old* would be quite helpful. With regard to shopping, expressions such as *how much is X?*, *I want to buy X*, *too expensive* or *costs X dollars* may be highly recurrent sequences for use in daily conversation. Discourse devices are lexical phrases that connect the meaning and structure of the discourse, such as logical connectors (e.g., *as a result*), temporal connectors (e.g., *and then*), fluency devices (e.g., *you know; it seems to me that*), exemplifiers (e.g., *for example; it's like*), evaluators (e.g., *as far as I know*) and so on. Each of these three main categories has a number of sub-categories associated with more specific functions and meanings (Nattinger & DeCarrico, 1992).

Biber et al. (2004, p. 384) identify three primary discourse functions for multi-word sequences in English (they use the term *lexical bundles*): (1) stance expressions, (2) discourse organisers, and (3) referential expressions. According to them, stance bundles "express attitudes or assessments of certainty that frame some other proposition", such as *I want you to* and *I don't think so*. These are usually used to convey personal attitudes, intention, prediction and so on. Discourse organisers "reflect relationships between prior and coming discourse" serving two major functions: topic introduction/focus (e.g., *what I want to do is*; *if*

you look at) and topic elaboration/clarification (e.g., *on the other hand; know what I mean*). Referential bundles "make direct reference to physical or abstract entities, or to the textual context itself". Examples of this include identification bundles (e.g., *those of you who*), imprecision bundles (e.g., *and things like that*), bundles specifying an attribute (e.g., *have a lot of*) and time/place/text-deixis bundles (e.g., *in the United States; the end of the*).

In addition, Carter and McCarthy (2006) illustrate the functions of multi-word expressions (they use the term *clusters*): relations of time and space (e.g., *in the; on the; the bottom of the*), other prepositional relations (e.g., *with a; for the*), interpersonal functions (e.g., *I don't know what; you know what I mean*), vague language (e.g., *sort of; and stuff; something like that*), linking functions (e.g., *and it was; but I mean*) and turn-taking (e.g., *what do you; do you think*) (pp. 834-837). These studies all demonstrate that multi-word expressions in English have systematic discourse functions although most of them are not semantically or grammatically complete patterns. As claimed by Biber (2009),

> Although they are neither idiomatic nor structurally complete, lexical bundles are important building blocks in discourse. Lexical bundles provide a kind of pragmatic 'head' for larger phrases and clauses, where they function as discourse frames for the expression of new information. (pp. 284-285)

Since it is accepted that multi-word expressions develop to serve the most important communicative needs of speakers or writers (Biber, 2009; Carter & McCarthy, 2006; Nattinger & DeCarrico, 1992; Schmitt, 2010; Wood, 2010; Wray & Perkins, 2000), in this study, the concordance listings of the recurrent sequences

were examined to analyse the functions of recurrent three-word sequences in their extended discourse context. The framework I employed for this analysis was drawn mainly from work done by Nattinger and DeCarrico (1992), and partly adapted from taxonomy works done by Biber et al. (2004) and Carter and McCarthy (2006). As discussed above, Nattinger and DeCarrico's (1992) function-based description of multi-word sequences is detailed and sufficient, and it is particularly developed for learners of English as a second or foreign language, thus making this taxonomy useful for the present research purpose. Biber et al.'s (2004) framework, on the other hand, is indeed comprehensive. However, it was developed based on classroom teaching and textbooks used at a university level, which are not the main concern of this present study. Additionally, although classroom teaching is a spoken register, it usually focuses on specific topics and most of the content might be pre-planned by the instructors, which contradicts the informal and unplanned nature of the BATTICC for this study.

As a result, Nattinger and DeCarrico's (1992) framework is mainly used for the analysis of multi-word expressions in this study. Nonetheless, in their work, only structurally and semantically complete sequences (e.g., *as a result*; *and then, I think*) are included, while sequences such as *and I think that, and it was* and *but I mean*, which are found with a high frequency, are ignored in their work. Therefore, Biber et al.'s (2004) and Carter and McCarthy's (2006) taxonomies provide useful supplements.

For this analysis, the top 50 recurrent three-word sequences were firstly retrieved from Taiwanese learners' discourse and British pupils' discourse in BATTICC, as well as the reference corpora CANELC and CANCODE. Sequences which serve

similar functions were then grouped into the same domain. However, it should be emphasised that it is sometimes difficult to assign a multi-word sequence to a particular category since, in some cases, a sequence serves multiple functions and is functionally ambiguous. For example, the sequence *would you like* in an interrogative clause might function as an offer, an invitation, a request or simply a question, and can sometimes be used to perform two or more speech acts at the same time. As Tsui (1994) argues, the source of multiple functions often lies in the sequential environment of the conversation in which the utterance occurs (p. 45). As a result, the use of each multi-word sequence in its discourse context is examined to identify the primary function of each sequence in its own context. The findings of this analysis will be presented in Chapter 6.

3.5 Summary

This chapter has sought to detail the project background, participants, data collection and data analysis, paying particular attention to the development of my own corpus BATTICC, which includes online (BATTICC-O) and spoken (BATTICC-F) datasets, and the development of the Taiwanese EFL textbook corpus TETCOC. This chapter has also highlighted the methodological issues and justified the general considerations taken in this thesis. Some of the key practical, technological and ethical questions that are faced have also been outlined and discussed. Some of the issues raised here will be revisited in the following chapters when analysing the data. Chapters 4, 5 and 6 will present in turn the analysis of online and spoken intercultural discourse from three points of view: keyness approach, discourse analytical approach and multi-word sequence perspective.

CHAPTER 4

Online and Spoken Discourse: A Keyness Approach

4.1 Introduction

Previous studies looking at intercultural discourse have typically selected particular linguistic features to study prior to the start of research (e.g., Davis & Thiede, 2000; Hanna & de Nooy, 2003; Liaw & Master, 2010; Montero et al., 2007). In this study, however, decisions on which features to investigate are not made on the basis of the researcher's intuitions or previous research; rather, they are derived from frequency information extracted from the sample of corpus data I collected. This approach is referred to as keyness analysis, derived from a corpus linguistics approach, which allows:

> macroscopic analysis (the study of the characteristics of whole texts or varieties of language) to inform the microscopic level (focusing on the use of a particular linguistic feature) and thereby suggesting those linguistic features which should be investigated further. (Rayson, 2008, p. 39)

That is, the specific linguistic features (the microscopic level) highlighted for further investigation are informed by macroscopic analysis. This macroscopic analysis is based on identifying significant differences between the frequencies of lexical and grammatical features in two groups of texts, such as parts-of-speech used by Taiwanese learners and native English speakers used in the present research. Once identified, these features are then subjected to further (microscopic) analysis. For example, a keyness comparison of parts-of-speech might identify a

statistically significant difference in the use of grammatical categories by different groups of participants.

As discussed in 3.4.3, the keyness approach works by comparing the actual observed frequency of each item in the target corpus with its equivalent in the reference corpus, and it therefore serves as a useful tool for directing researchers to significant lexical or grammatical differences between texts (Adolphs & Lin, 2010; Bachmann, 2011; Baker, 2006; Scott, 2010). This study employs a keyness approach to find the distinctive patterns of language use by a group of adolescent Taiwanese learners of English interacting with adolescents based in the UK on electronic discussion boards, i.e., computer-mediated communication (CMC) and in an informal face-to-face (FTF) meeting. Specifically, the following questions are addressed:

(a) What topics are the young people mainly concerned with in CMC and FTF intercultural communication?

(b) What are the statistically significant differences of lexical and grammatical features between the Taiwanese and British participants' discourse?

This study examines three levels of keyness analysis, namely keywords, semantic domains and part-of-speech levels. Keywords and semantic analysis are used to compare and contrast the BATTIC and the CANELC in order to identify the themes that young people focus on in intercultural online communication. In addition, the key grammatical categories that the Taiwanese participants used significantly differently from the British participants are identified by applying

key part-of-speech analysis to the comparison of the discourse by the two groups of participants.

This chapter will begin with an analysis of frequency, followed by keywords and key semantic domain analysis, comparing and contrasting BATTIC and reference corpora in order to identify the themes that young people focus on in online and spoken communication. The key grammatical categories that the Taiwanese participants used significantly differently from the British participants are then identified by applying key part-of-speech analysis to the comparison of the discourse by the two groups of participants.

4.2 Frequency

As discussed in 3.4.2, the frequency of a word or a phrase in different text types is an important part of its description in the context of use. It is therefore a good starting point for subsequent analysis. The first step in the analysis was to produce word frequency lists. Using the *WordList* function in *WordSmith Tools 5.0* (Scott, 2008), I generated frequency lists for the BATTICC, both online communication (BATTICC-O) and spoken interaction (BATTICC-F), as well as the reference corpus of online discourse (CANELC: General E-language Corpus) and spoken discourse (CANCODE: Cambridge and Nottingham Corpus of Discourse in English), as shown in Table 4.1.

The most striking feature in the comparison of the datasets of online discourse, BATTICC-O and CANELC, is the use of first person singular variants. In BATTICC-O subjective *I* and its possessive form *my* rank first (5.64%) and fourth

(2.82%) respectively, while in CANELC these two items only occupy the sixth (1.57%) and the 21st (0.46%) positions respectively. This is partially explained by the fact that the participants in this project were writing about themselves, their interests and their personal experiences, showing a high degree of intimacy, informality and in-group membership. In contrast, CANELC contains texts collected from a wide range of digital communication, some of which (e.g., monologues on blogs) may not be as interpersonal and interactional as BATTICC-O. As such, it seems that the overwhelming use of the pronoun *I* is one of the most distinct linguistic features of BATTICC-O from this observation.

Table 4.1

Most Frequent Items in BATTICC-O, BATTICC-F, CANELC and CANCODE

	BATTICC-O	Freq.	%	CANELC	Freq.	%	BATTICC-F	Freq.	%	CANCODE	Freq.	%
1	I	1,829	5.64	THE	20,374	3.96	YOU	443	3.17	THE	155,36	3.31
2	AND	982	3.03	TO	12,300	2.39	I	409	2.93	I	142,06	3.03
3	IS	980	3.02	A	11,647	2.26	AND	366	2.62	AND	131,10	2.80
4	MY	915	2.82	OF	9,934	1.93	LIKE	354	2.53	YOU	127,99	2.73
5	TO	853	2.63	AND	9,749	1.89	THE	354	2.53	IT	99,029	2.11
6	THE	750	2.31	I	8,071	1.57	TO	285	2.04	TO	97,955	2.09
7	A	552	1.70	IN	7,331	1.42	YEAH	285	2.04	A	95,436	2.03
8	IN	516	1.59	IT	5,423	1.05	WE	240	1.72	YEAH	85,472	1.82
9	YOU	438	1.35	IS	5,272	1.02	IT	217	1.55	THAT	77,996	1.66
10	LIKE	415	1.28	FOR	5,259	1.02	IN	212	1.52	OF	70,028	1.49
11	HAV	388	1.2	YOU	5,089	0.99	A	209	1.49	IN	56,598	1.21
12	IT	378	1.17	THAT	4,786	0.93	DO	205	1.47	WAS	47,782	1.02
13	WE	343	1.06	ON	4,437	0.86	HAVE	188	1.34	IT'S	44,095	0.94
14	SCHOOL	311	0.96	WITH	3,046	0.59	IS	179	1.28	KNOW	43,702	0.93
15	VERY	284	0.88	BE	2,915	0.57	SO	164	1.17	MM	41,617	0.89
16	OF	283	0.87	THIS	2,796	0.54	ERM	152	1.09	IS	40,768	0.87
17	AM	278	0.86	BUT	2,710	0.53	IT'S	147	1.05	ER	40,745	0.87
18	BUT	253	0.78	HAVE	2,629	0.51	OF	146	1.04	BUT	38,422	0.82
19	ARE	251	0.77	AT	2,539	0.49	THAT	139	0.99	THEY	36,911	0.79
20	AT	234	0.72	AS	2,452	0.48	ER	133	0.95	SO	36,713	0.78

On the other hand, regarding the spoken discourse in BATTICC-F and CANCODE, the overwhelming use of the interactive personal pronouns *you* and *I*

is one of the distinct linguistic features of both of these datasets. In BATTICC-F, *you* and *I* are ranked as the first two most frequent items, and they are ranked second and fourth in CANCODE. In addition, some high-frequency words show markers of interactivity typical of the spoken nature of face-to-face communication, such as *yeah, mm* (response tokens) and *er, erm* (pauses). These serve to foreground choices that reflect the immediate social and interpersonal situation in spoken communication (Carter & McCarthy, 2006). Moreover, with regard to the comparison between BATTICC-O and BATTICC-F, the participants tended to use more tokens of *you* and *we* and less of *I* in face-to-face communication compared to online discourse. It seems, therefore, that the BATTICC-F is less self-oriented than the BATTICC-O and, in other words, is more interactive.

However, McCarthy and Handford (2004) argue that the frequency information sometimes fails to capture crucial differences between examples of discourse. This is particularly true when I look at the raw frequency lists derived from the four datasets, showing that the most frequent items are mainly grammatical ones, such as determiners (e.g., *the, my, what,* etc.), prepositions (e.g., *to, of, in, for,* etc.) and conjunctions (e.g., *and, but,* etc.), and that the four sets are all quite similar, although the order of frequency in which they occur is slightly different. This is hardly surprising, given that most languages are dominated by grammatical and functional items (Adolphs, 2006; Baker, 2006; Schmitt, 2010).

4.3 Keyword analysis

Although the high-frequency items indicate some features of online and spoken

texts, frequency analysis does not elucidate many significant discourse features of the BATTICC-O and BATTICC-F in terms of intercultural communication since the high-frequency words mainly consist of grammatical items. Consequently, as Baker (2006) claims, one way of finding what lexical items are interesting in a frequency list is to compare more than one list together (p. 124): to employ keyness analysis. This approach, as discussed in Chapter 3, is better suited to highlighting the main elements that are characteristic of a specific collection of texts.

Table 4.2

Keyword Lists: BATTICC-O vs. CANELC

	Keyword	Freq. (BATTICC-O)	%	Freq. (CANELC)	%	Positive/ Negative	Keyness
1	I	1,829	5.64	8,071	1.57	+	1,899.60
2	MY	915	2.82	2,385	0.46	+	1,583.00
3	SCHOOL	311	0.96	134	0.03	+	1,232.00
4	TAIWAN	170	0.52	1	-	+	949.40
5	IS	980	3.02	5,272	1.02	+	765
6	AM	278	0.86	507	0.1	+	614.3
7	LIKE	415	1.28	1,300	0.25	+	609.3
8	PLAY	164	0.51	115	0.02	+	563.4
9	HI	139	0.43	56	0.01	+	559
10	FAVOURITE	137	0.42	105	0.02	+	456.3
11	VERY	284	0.88	848	0.16	+	434.8
12	GO	221	0.68	604	0.12	+	364.8
13	FRIENDS	127	0.39	151	0.03	+	353.2
14	NAME	127	0.39	168	0.03	+	335.4
15	HAHA	118	0.36	209	0.04	+	265
16	FOOD	117	0.36	209	0.04	+	261.4
17	WE	343	1.06	1,869	0.36	+	259.7
18	THE	750	2.31	20,374	3.96	-	256.7
19	HELLO	80	0.25	83	0.02	+	236.5
20	OF	283	0.87	9,934	1.93	-	228.8

The procedure of the keyness analysis in this study works by comparing the actual observed frequency of each item in the target corpus (i.e., BATTICC-O and BATTICC-F) with its equivalent in the reference corpus (i.e., CANELC and CANCODE). Using *WordSmith Tools 5.0*, for 1 degree of freedom (d.f.), at 99%

confidence (p <.01), the cut-off of 6.63 illustrated 608 keywords. This reduces to 254 words at the 99.999% (p <.001) level, with the critical value of 15.13, as recommended by an evaluation reported by a number of researchers (e.g., Harvey, 2008; Culpeper, 2009).

Table 4.2 presents the top 20 words (with the largest LL values) that are key to the BATTICC-O as compared with the reference corpus CANELC. In the comparison, items occurring both unusually frequently (positive keywords: with a + sign) and unusually infrequently (negative keywords: with a - sign) compared to CANELC are illustrated. From the table it is clear that the personal pronoun *I* and its possessive form *my* are the most significant, with high LL ratios of 1,899.60 and 1,583.00 respectively, which means that they occurred appreciably more frequently in the BATTICC-O than in the CANELC.

Table 4.3

Keyword Lists: BATTICC-F vs. CANCODE

	Keyword	Freq. (BATTICC-F)	%	Freq. (CANCODE)	%	Positive/ Negative	Keyness
1	TAIWAN	63	0.45	5	0	+	697.55
2	LIKE	354	2.53	31,736	0.68	+	417.52
3	TAIWANESE	27	0.19	2	0	+	299.63
4	ENGLAND	49	0.35	310	0	+	285.94
5	FACEBOOK	15	0.11	0	0	+	174.53
6	VERY	129	0.92	11,004	0.23	+	160.45
7	FOOD	38	0.27	606	0.01	+	157.01
8	WE	241	1.72	32,458	0.69	+	151.85
9	SCHOOL	57	0.41	2,049	0.04	+	151.63
10	ENGLISH	42	0.3	903	0.02	+	150.49
11	UK	14	0.1	7	0	+	136.2
12	FAVOURITE	21	0.15	157	0	+	116.1
13	MOUNTAIN	17	0.12	80	0	+	108.24
14	DO	205	1.46	30,440	0.65	+	105.83
15	NAME	38	0.27	1,424	0.03	+	98.27
16	DRAGON	11	0.08	11	0	+	97.55
17	HOT	24	0.17	483	0.01	+	88.87
18	CHINESE	15	0.11	106	0	+	84.47
19	FESTIVAL	12	0.09	48	0	+	79.86
20	THEY	33	0.24	36,911	0.79	-	75.06

This is in line with the results presented in the frequency lists (see Table 4.1). Furthermore, the keywords in the list can be said to be indicative of "aboutness" in that they are identified to be unique to and particularly more frequent in BATTICC-O compared with the general reference corpus, so these words can then be described as defining characteristics of BATTICC-O. They also reveal the specific themes that the participants are mainly concerned with during their online discussion, including school life (e.g., *school*), what they like (e.g., *favourite, like, play*), food and so forth. Similarly, from Table 4.3, countries (e.g., *Taiwan, UK, England*), school life, food, festivals, friendship and so on could possibly be seen as the most popular topics when the participants meet face-to-face. In this regard, the keyword list appears to be a snapshot, revealing the predictable content domains that are frequently talked about by the participants. While this assumption is perhaps reasonable, undertaking an appropriate semantic annotation to explicitly categorise the semantic similarities between words would be more reliable (Culpeper, 2009; Rayson, 2008). As Wilson and Rayson (1993) state, semantic annotation is closely related to "content analysis", which is "concerned with the statistical analysis of primarily the semantic features of texts" (p. 2).

4.4 Key semantic domains in CMC: BATTICC-O

Using *WMatrix* to analyse the semantic domains of the discourse, for $p < .001$, the cut-off of log-likelihood = 15.13 generated 37 USAS tags that are either significantly overused or underused between the BATTICC-O and CANELC data. In this case, the analysis of key semantic domains makes the data more manageable than keyword analysis. The top 10 tags (with the largest LL values) in this set are shown in Table 4.4, including 8 overused (shown with +) and 2

underused (shown with -) domains.

Table 4.4

Ten Most Significant Differences between BATTICC-O and CANELC at Semantic Level

Rank	Semantic code	BATTICC-O		CANELC		Overuse or underuse	LL	Semantic domain
		Freq.	%	Freq.	%			
1	Z8	4676	16.08	18686	9.10	+	981.85	Pronouns
2	P1	554	1.97	470	0.23	+	979.30	Education in general
3	Z99	889	2.91	11258	5.48	-	586.83	Unmatched
4	F1	572	1.91	1025	0.50	+	455.08	Food
5	E2+	467	1.57	787	0.38	+	435.55	Like
6	K1	356	1.23	481	0.23	+	429.57	Entertainment generally
7	I1	25	0.09	1980	0.96	-	325.64	Money generally
8	S3.1	225	0.78	258	0.13	+	310.26	Personal relationships
9	K2	205	0.71	235	0.11	+	283.38	Music and related activities
10	A13.3	492	1.70	1566	0.76	+	197.27	Degree: boosters

As can be seen in the table, the most significant difference (LL value 981.85) in the semantic comparison is for the tag Z8 representing the semantic field Pronouns. This is consistent with the results generated in the keyword analysis (Table 4.2) and the frequency list (Table 4.1). This category is then followed by Education in general (P1), Food (F1), Like (E2+), Entertainment generally (K1), Personal relationships (S3.1), and Music and related activities (K2). These overused domains indicate the key semantic categories of the adolescent online intercultural communication identified in this study compared with general online communication; that is, these key themes are the ones that are most commonly discussed online by the participants. Yet it is noted that the topic of Money

generally (I1) is much less frequently discussed by the young people, with only 0.09% of the texts in the BATTICC-O data being concerned with money compared with 0.96% in the CANELC. Additionally, far fewer items were found in the domain Unmatched (Z99) in the BATTICC-O data (2.91%) than in the CANELC data (5.48%).

As far as the domain of Unmatched (Z99) is concerned, various types of acronyms (e.g., *OMG* – Oh my God!), ellipsis (e.g., *fav* – favorite, *thx* – thanks, *msg* – message, *hav* – have, *im* – I'm, *dont* – don't), repeated letters (e.g., *sooooo, waiitttt, huuugggee, doooooooooooo*), repeated punctuation marks (e.g., *!!!!!!, ????, :)))))))*), spellings that represent prosody or nonlinguistic sounds (e.g., *helloooo, woohoo, haha, hehe, ahhh*) and emoticons (e.g., *xxxxx, XD*) are included in this category because of their nonstandard spellings, as shown in the following examples:

- I absolutely cannot ***waiitttt !!!!!"!!!!!!!!!!!***
- ***OMG***!! really?! ***its*** that close?! ***ahhh***!! SO excited!! ***woohoo!***
- looks a bit strange but i ***looovveee*** oysters so i think it would be very nice!!***XD***
- nothing to ***doooooooooooo***! ***haha***...
- <BT28> is my bestfriend ***xxxxx***

(BATTICC-O)

Such distinctive language features are commonly seen in different modes of CMC, and they are likely used to economise on typing effort, to mimic spoken language features, or for creative expressions (Herring, 2003, 2011). According to Riordan and Kreuz's (2010) corpus-based study, discussed in Chapter 2, the role of these language features is mainly to disambiguate a message (36%), to regulate the interaction (24%), to express affect (15%) and to strengthen the message content (10%). In particular, these communication strategies demonstrate the ability of

users to adapt the computer medium to their expressive needs. It appears that the language of CMC has a wealth of cues, providing information and expressing emotion and intimacy to compensate for missing visual and aural cues (Crystal, 2006, 2011).

However, further examination of their use by English native speakers and Taiwanese learners shows that many more instances of paralinguistic features are found in the British adolescents' discourse, while only a small number are found in the Taiwanese texts in the last three months of the project. This indicates that young Taiwanese learners seldom employed these strategies at the beginning of the online exchange project but gradually learned from their British peers by observing the paralinguistic features in context. This is probably because they previously had limited opportunities for online contact with people using English in CMC and their textbooks generally do not include these features; as a result, they might not initially know how to employ these strategies appropriately. It may therefore be appropriate to introduce some of these distinctive features in the EFL classroom, which can further facilitate online communication.

Although many more instances of paralinguistic features were found in the British participants' discourse than the Taiwanese students' messages, they were not as pervasive as those in the CANELC. This is probably because the British participants realised that their Taiwanese peers may not understand some of the abbreviations or other paralinguistic features and therefore tended to use fewer nonstandard spellings so that the Taiwanese learners would understand them more easily. As was noted by Crystal (2006), although the members of an online community come from different backgrounds and write in different styles, they

tend to accommodate each other, and as such their contributions progressively develop a shared linguistic character.

Items in each semantic domain were examined and then divided into two groups, one comprising British and the other Taiwanese participants. Some of the examples are illustrated in Table 4.5. A number of differences in the lexicons of the two groups within a board topic may be noted, reflecting social and cultural differences. In the food category, items such as *tofu, bubble tea, rice, dumplings, noodles* and *soup* that are commonly served in Oriental countries and particularly in Taiwan are mentioned more often by the Taiwanese participants, while these are not found with a high frequency in the British students' texts. In contrast, items such as *pasta, fish and chips, crisps, pizza* and *cheese* occurred more frequently.

Table 4.5

Items Within Semantic Categories (and their Raw Frequencies)

Semantic category (tag code)	Items within the category (in descending order of frequency, excluding the ones occurring less than five times)	
	Taiwanese participants	British participants
Food (F1)	food (65), eat (33), chocolate (23), tofu (13), snacks (12), tea (11), ice cream (9), fruit (8), rice (7), dumpling (7), tomato (5), moon cake (6), fried chicken(5), omelet (5), noodles (5), beef (5), strawberry (5), restaurant (5)	food (59), eat (24), chocolate (23), breakfast (15), dinner (14), chips (8), pasta (8), crisps (8), cook (7), toast(6), restaurant (6), curry (5), pizza (5), fish and chips (5), sweets (5), cereal (5), ice cream (5), fudge (5), cheese (5), cake (5)
Education in general (P1)	school (140), study (32), teacher (28), class (15), exam (14), P.E. (12), student (12), lesson (12), test (8), homework (8), math (5)	school (151), lesson (36), college (15), geography (12), P.E. (12), tutor(8), student (6), teacher (5), homework (5), classroom (5)
Sports (K5.1)	sports (22), games (20), basketball (11), badminton (10), swimming (9), tennis (9), baseball (8), skating (5), jogging (5), stadium (5), soccer (5)	sports (24), badminton (10), rugby (9), cricket (8), riding (8),tennis (6), football (6), jogging (6), swimming (5), skate (5), skiing (5)

Furthermore, in the domain of education in general (P1), interestingly, the

Taiwanese students used *exams* and *tests* 22 times in total, while those two words were each mentioned less than three times by the British students. This seems to reflect that tests and exams occupy an important place in school education in Taiwan, as can be seen in the following texts:

(1) I am still thinking about this [the New Year's resolution]...ha maybe hope the world is peaceful and hope every day less **tests** and **exams**

(2) hope each time takes a **test** has the good result is my new year new hope.

(3) I haven't get on the website for a long time.
How's everyone??? Hope evryone have a very nice **exam**.

(4) we just had our second **exam** this semester. How many **exams** do you have? We have three big ones in a semester, but we still have many **quiz** every day. I don't like **exams**...

(5) Math **exam** was so difficult. I hope I can pass. Social studies is hard, too. I didn't do well on the **test**, <BT02>.

(BATTICC-O)

In examples (1) and (2) the Taiwanese students put exams as one of their New Year's resolutions; in excerpt (3), he/she made such a greeting and blessing probably because testing is one of the important issues that he/she is really concerned about; the descriptions in (4) and (5) show that school life in Taiwan is filled with many tests every day.

Popular sports in Taiwan and the UK are rather different. As can be seen in Table 4.5, basketball, baseball and swimming are more popular among the Taiwanese, whereas football, rugby and cricket, with relatively more occurrences in the British texts, are not commonly mentioned in the young Taiwanese students' texts. I also observed that the participants occasionally asked members of the other group a variety of questions, demonstrating a willingness to engage with otherness and a

curiosity for discovering different perspectives from a culture other than their own. For example, Taiwanese participants sometimes asked what rugby and cricket are because they had never seen or even heard of these; on the other hand, the British participants often queried some culturally related terms used by Taiwanese learners, especially when they talked about oriental foods. This is where socio-cultural learning occurs, which makes CMC a potentially beneficial tool to support collaborative learning. It appears that culture is embedded in language, and in this virtual third space, language and culture are even more tightly intertwined (Byram, 1997; Kramsch, 1993; Liaw, 2006; Warschauer, 1997).

4.5 Key semantic domains in FTF communication: BATTICC-F

Using *WMatrix* to analyse the semantic domains of the spoken discourse, for $p < .01$ with 1 d.f., the cut-off of 6.63 presented 102 USAS tags that are either significantly overused or underused between the BATTICC-F and the BNC Sampler Spoken data. At the $p < .001$ level, the critical value 15.13 revealed 53 significant USAS tags. The top 10 tags (with the largest LL values) in this set are shown in Table 4.6, including 7 overused (with +) and 3 underused (with -) domains. As can be seen, the significantly overused categories include Geographical names (Z2), Like (E2+), Education in general (P1), Degree: boosters (A13.3), Food (F1), Geographical terms (W3) and Weather (W4), and this indicates the topics commonly discussed among participants, as compared with a general spoken interaction. On the other hand, the domains that were underused by the participants are Negative (Z6), Existing (A3+) and Likely (A7+).

Table 4.6

Ten Most Significant Differences at the Semantic Level between BATTICC-F and BNC Sampler Spoken

Rank	Semantic code	BATTICC-F		BNC Sampler Spoken		Overuse or underuse	LL	Semantic domain
		Freq.	%	Freq.	%			
1	Z2	234	1.62	3541	0.36	+	331.32	Geographical names
2	E4.1+	118	0.78	782	0.08	+	313.39	Happy
3	E2+	171	1.25	2255	0.23	+	292.14	Like
4	P1	142	0.99	1928	0.20	+	224.40	Education in general
5	A13.3	240	1.67	5457	0.56	+	202.64	Degree: boosters
6	Z6	92	0.64	19932	2.03	-	185.93	Negative
7	F1	186	1.29	3914	0.40	+	176.04	Food
8	W3	75	0.52	699	0.07	+	163.28	Geographical terms
9	W4	48	0.32	379	0.04	+	113.59	Weather
10	A7+	95	0.66	15034	1.53	-	89.96	Likely

The domain of Geographical names (Z2) is the most commonly mentioned by the participants. In particular, the words *Taiwan, England, Taiwanese* and *UK,* involving the two countries where the participants come from, are the top four items within this category. The reason for this is probably that they talked about their own cultures and contrasted the differences between the two countries. At the same time, they demonstrated a willingness to learn from their international peers, showing interest in language, culture and life experiences. For example:

(6) <TW13>: In ***England*** what kind of food ... ***in England***?
<BT16>: What kind of food is there? Erm ... mainly fish and chips and ...

(7) <BT01>: How do we say turtle in ***Taiwanese***?
<TW01>: Pardon?
<BT02>: Turtle.

(8) <BT19>: What festivals do you have in ***Taiwan***?
<TW17>: We have Moon Festival.
<BT19>: What's that?

(9) <BT19>: Erm ... What food do you ***like*** in ***England***?
<TW17>: Erm ... I ***like England*** candy. It's very good.
<BT19>: The candy, yeah. ... In Taiwan, I ***liked*** the stuffed dumplings. The food is very, very different.
<TW17>: Do you ***like*** the stinky tofu?
<BT19>: I didn't try it.
<TW17>: Ah ... it's very good.

(BATTICC-F)

In addition, comparing the results of the key semantic domain analysis for the BATTICC-O and BATTICC-F, namely Table 4.4 and Table 4.6, it can be seen that Education in general (P1), Food (F1) and Like (E2+) are found with a high frequency in both datasets. It seems that these three topics were frequently talked about by participants in online communication and face-to-face interaction. This is probably because food and school life are quite different between Taiwan and the UK, and the participants shared what they like in both their home and the other country.

However, by further examining the items in concordance lines within the domain Like (E2+), which includes the items such as *like, love, enjoy, popular,* etc., what needs to be stressed is that some instances of *like* were not used as verbs to express speakers' feelings of enjoyment or to show that they find something/someone pleasant or attractive. Instead, they function as fillers or discourse markers (DMs), which should be attributed to the Grammatical bin (Z5) in *WMatrix* instead of this category (E2+). In this semantic domain analysis, within the category E2+ the use of *like* as a DM was found in 16 occurrences out of 171 concordance lines. This means in this case that only 90.64% of the items in E2+ were automatically tagged

correctly. In light of this finding, careful manual checking of concordances and interpretation of results are obviously required. Such a limitation of automatic tagging has been mentioned in a number of previous studies that employ *WMatrix* as the main research tool (Culpeper, 2009; Rayson, 2008).

On the other hand, the discourse marker *like* has been examined in a number of linguistic and discourse studies which have highlighted its particular pervasiveness in teenage talk (e.g., Andersen, 1998, 2000; Hellermann & Vergun, 2007; Tagliamonte, 2005). This can be seen in the following instances from BATTICC-F:

(10) <BT12>: Yeah, they don't .... over here they take lots of ***like*** care in the presentation ***like*** being clean and people have ***like*** a lot of respect for them.

(11) <BT07>: I wasn't gonna ***like*** perform until ***like*** ... erm ... before the performance ***like*** just before we went to sit down.

(BATTICC-F)

Clearly from the excerpts above *like* is used very frequently in speech in that it often occurs multiple times in a single utterance. Some occurrences serve to introduce new information or elements of utterances, and they do not seem to have a meaningful or syntactic function. That is, *like* sometimes acts as a filler that simply allows speakers to buy time to think what they are going to say. However, research has reported its functional complexity, serving as quotative marker, focus marker, approximator, exemplifier, hedge, discourse link or hesitational device (Adolphs, 2010; Anderson, 2000; Hellermann & Vergun, 2007). This will be further discussed in detail in 5.5.5.

By further examining their discourse concerning the two commonly discussed topics, food and likes, the young people showed their intercultural competence,

including attitudes of curiosity and openness, relational knowledge, skills of discovery and interaction and critical culture awareness (Byram, 2000, 2012; Fantini, 2012; Liaw, 2006), as discussed in Chapter 2. An attitude of curiosity and openness refers to the interest in learning about other people's way of life and introducing one's own culture to others. As can be seen in (12), the Taiwanese student TW09 was curious about what the British people usually eat in their daily life and why they look much taller than Taiwanese people in general. Furthermore, some entries show evidence that the participants have knowledge about their own culture and the others' culture for intercultural communication. As in (13), for example, the Taiwanese participant TW07 has knowledge of some facts about the English school system, so this makes communication easier and maintains the discussion about the differences in school life. With regard to skills of discovery and interaction, the participants discovered the differences between Taiwanese and British food through the social interaction and sharing experiences.

(12) <TW09>: So, what do you usually eat you know you're so tall and don't look like a junior high school student [laughter]

(13) <BT15>: Oh. What grade are you in?>
<TW07>: I'm in 7th, but in your country I'm in 8th grade. I'm 13...
(BATTICC-F)

In BATTICC-F, participants also showed their critical cultural awareness, "an ability to evaluate critically and on the basis of explicit criteria perspectives, practices and products in one's own and other cultures and countries" (Byram et al., 2002, p. 13). As in (14), the participants were talking about the cleaning job in schools, in which they discussed the different situations in the two countries. In Taiwan, cleaning the hallways, classrooms and campus is typically the first thing that all the students have to do every day when they arrive at school, and they usually do it at least twice a day. In the British schools, however, students do not

normally do that. As a result, at the beginning the British participants thought the Taiwanese students were just like child labourers. After some discussion on this issue the British interlocutors became aware of the differences between the two countries and their own assumptions and preconceptions. They also interpreted the different values on this issue. For example, they noticed that the Taiwanese students show more respect to their schools and the environment.

(14) <BT07>: No, no, no as in like kids and that because there was kids like cleaning the hallways and stuff and in England you have .... it's like child labour ... erm ....
<BT09>: Like using a child for a job which an adult should be doing.
<BT08>: But that's not the best way to describe this. We just like see that as like a bad thing to do but here I think they're like quite proud of their schools and stuff but we just like see it as it's not our job to clean up so we're not going to do it.
<BT09>: Training kids for like adult life.
<BT07>: But in the UK, like, no-one has respect for the schools, like, you know, there's chewing gum all over the floors, and there's like rubbish. [...]
<BT07>: But it's just not as like respected as it is here. That's what we're trying to say through all of that long conversation.

(BATTICC-F)

Sections 4.4 and 4.5 have displayed the semantic categories that are key in BATTICC-O and BATTICC-F respectively, illustrating a set of semantic domains that are specific to these particular interactions among the participants and also revealing the themes that the young people discussed commonly on the online discussion board and in face-to-face intercultural interaction. The words in each semantic category also display a great number of cultural and social differences in terms of lexical choices by different groups of teenagers. In the following sections the attention will be turned to key domain analysis at the POS level.

4.6 Key part-of-speech analysis: BATTICC-O

Applying the keyness method at the POS level, the two sets of texts from the Taiwanese and British participants were compared for their relative use of grammatical categories. For $p < .01$, the cut-off of 6.63 indicated that 58 POS tags are significantly overused or underused between the two sets. At the 99.99% level ($p < .001$) there are 18 significant POS tags. Table 4.7 displays the top 10 grammatical tags (with the largest LL values) that are key in both sets of texts. We can see that the domains of singular letter of the alphabet (ZZI) and coordinating conjunction (CC) are most underused by the Taiwanese participants as compared with the British ones, but they tend to overuse the general adjective (JJ) and the modal auxiliary (VM). The following subsections will investigate these four categories in more detail.

Table 4.7

Ten Most Significant Differences at Part-of-Speech Level between Taiwanese and British Participants in BATTICC-O

Rank	POS code	Taiwanese participants		British participants		Overuse or underuse	LL	POS
		Freq.	%	Freq.	%			
1	ZZ1	35	0.24	250	1.55	-	163.67	singular letter
2	JJ	1267	8.62	846	5.26	+	126.71	general adjective
3	CC	357	2.43	720	4.48	-	94.23	coordinating conjunction
4	VM	318	2.16	194	1.20	+	56.28	modal auxiliary
5	VBZ	318	4.38	141	2.62	+	63.34	is
6	RT	52	0.42	186	1.12	-	54.61	quasi-nominal adverb of time
7	PPHS1	33	0.22	102	0.75	-	42.89	third person sing. subject
8	VVZ	54	0.37	129	0.95	-	38.14	-s form of lexical verb
9	VVI	592	4.03	434	2.70	+	37.77	infinitive
10	DA2	72	0.49	20	0.12	+	36.08	plural after-determiner

4.6.1 Singular letters of the alphabet

With regard to the domain of singular letter of the alphabet (ZZI), there are far more occurrences found in the British texts (250 tokens/1.40%) than in the Taiwanese ones (35 tokens/0.24%). When the items within this category were examined, various types of word substitution by a singular letter could be found. In particular, the letter *i* was most predominant, in that there was a strong tendency for the British teens to use lower case *i* to substitute for the conventional upper case *I*, while the Taiwanese learners rarely did so. A similar substitution was also found in the use of *n* for *and*, *u* for *you* and *r* for *are*, as the following example shows:

> i really wanted to do the stuffed toy thing n when i got there there were just a load of people standing round and one person from one school writing all the names in n i was like :O I wanted to do that lol yeah i was gonna do that too : S cant remember what i signed up for now ohh well
>
> (BATTICC-O)

This paragraph presents many instances of the singular letters *i* and *n*, an example of economical language production commonly found in CMC. Additionally, both Taiwanese and British texts show a high-frequency of use of singular letters such as *D, X, P, S* and *O*, which are widely used in emoticons such as *:D* and *X* (pleasure, humour, etc.), *:P* (joking, a face with the tongue stuck out), *:S* (confused) and *:-O* (shocked, amazed). Other emoticons widely used include simple "smilies" *:)* and "winkies" *;),* although these are not classified in this domain. As Crystal (2006, p. 39) claims, these emoticons, the "combinations of keyboard characters", are commonly employed in CMC to compensate for the absence of real facial expressions, gestures and conventions of body posture that are so crucial in expressing personal opinions and attitudes and in moderating

social relationships.

4.6.2 General adjectives

As shown in Table 4.7, General adjective (JJ) is the category that was most overused by the Taiwanese learners, with a log-likelihood (LL) value of 126.71. A further examination of the items within the category of JJ shows that a number of them are not used to describe something, but are a formulaic use of general adjectives in self-introduction. For example, all 39 occurrences of *junior* are used to describe the schools the participants are studying at (e.g., *I am from ABC Junior High School*). It was also found that some items within this category were categorised incorrectly. For example, Chinese is often referred to as a subject or a language instead of a general adjective (e.g., *Among all the subjects, I only like Chinese and math/ I can teach you Chinese*). In light of this finding, careful manual checking of concordances and interpretation of results are obviously required (Culpeper, 2009; Rayson, 2008).

Despite this problem, the category JJ comprises many positive emotion words (e.g., *good, nice, great, favourite, interesting, happy,* etc.) used by both groups to express their positive attitudes toward the intercultural exchange. According to Matsumoto, Yoo and LeRoux (2009) and Liaw and Master (2010), emotions play an important role in intercultural communication, since the ability to regulate emotion is one of the keys to effective intercultural communication and adjustment. The positive emotions demonstrated in the process of CMC may have smoothed the communication, contributed to the success of the project and maintained participants' motivation.

4.6.3 Coordinating conjunctions

The domain of coordinating conjunctions (CC) is one of the most underused items by the Taiwanese participants in online communication, with an LL value of 94.23. At this point the proportionate use of such items within this domain in the British pupils' texts is almost double that in the Taiwanese students' texts (2.43% compared to 4.48%). Table 4.8 presents the frequencies of different items within this domain derived from the two different texts. As can be seen in the table, the much more frequent use of *and* and *or* by English-speaking pupils is particularly striking, with the Taiwanese and British texts having 322 versus 645 instances of *and* respectively, and 35 versus 64 instances of *or* respectively.

Table 4.8

Frequencies of Different Coordinating Conjunctions in BATTICC-O

Taiwanese participants			British participants		
items	Frequency	%	items	Frequency	%
and	322	2.19%	and	645	4.01%
or	35	0.24%	or	64	0.40%
			and stuff	7	0.04%
			plus	2	0.01%
			and everything	2	0.01%
total	357	2.43%		720	4.48%

On further examining the concordance lines of *and* it was found that most occur between sequences of clauses or sentences. In this way they provide important cohesive links between sentences (Carter & McCarthy, 2006), and often act as a discourse marker, which "introduces a separate message with its propositional content" (Fraser, 1999, p. 939), instead of purely functioning as a conjunction within a single message. According to Schiffrin (1987, p. 128), *and* is the most frequently used mode of connection at the local level of idea structure to

coordinate idea units and continue a speaker's action. A number of instances will serve to illustrate their function:

(15) my name is Sophie ***and*** i am 13 years old ***and*** my bestfreinds are Nastaha and Emma!
(16) she is an eight month old Poddle ***and*** she is very lively ***and*** happy , Here is a picture of me ***and*** my little brother, Darryl ***and*** then a picture of Tyra.

(BATTICC-O)

These examples show sentences or sequences of clauses linked by the coordinating conjunction *and*, with clauses strung together in a sequence of one clause unit being added to another and the units having equal status. In example (15), *and* works to link three pieces of information, i.e., name, age and best friends, into a section of personal introduction, while in example (16), all the information is about a pet dog and is linked with *and*. In these cases, *and* seems to be free of meaning as it is considered merely a marker in an idea structure (Schiffrin, 1987). The high-frequency use of *and* by the British participants is very noticeable. On the other hand, the Taiwanese students' texts do not appear to be as cohesive as the British students' since very few instances of the use of *and* to coordinate their ideas are evident. Consider the following examples:

(17) Hello, My name is Qianyulli. I am 13 years old. My birthday is on 11/6. I study in XX Junior high school in Hualien, Taiwan.
(18) Tyra is so lively. [My pet dog] Pico is timid. He is afraid other dogs.

Examples (17) and (18) present comparable discussions to (15) and (16). However, when describing themselves and a pet dog, the Taiwanese learners do not use *and* to structure their discourse. In the corpus as a whole, this also makes the average length of sentences by the British students longer than those by the Taiwanese

students. *WordSmith Tools* 5.0 was used to calculate sentence length, showing it to be 16.19 words in the British students' texts and 8.02 in the Taiwanese students' texts.

Table 4.9

Frequencies of Different Coordinating Conjunctions in BATTICC-F

Taiwanese participants			British participants		
items	Frequency	%	items	Frequency	%
and	128	2.43%	and	224	2.22%
or	24	0.46%	or	27	0.27%
			and stuff	11	0.11%
			or anything	7	0.07%
			and everything	3	0.03%
total	152	2.89%		273	2.69%

Although the differences in the use of coordinating conjunctions between the two groups of pupils is clear in the BATTICC-O, this is not the case in the BATTICC-F; that is, no significant differences are found between the two groups in the use of the coordinating conjunctions in the spoken data. As can be seen in Table 4.9, the overall percentage of this domain is comparable, with 2.89% versus 2.69%. In this regard, *and* was used frequently by both groups in face-to-face interaction.

When sorting the concordances in *WordSmith* to see how *and* was used in context, there is a strong tendency of it being used as a useful turn-initial resource for speakers. Evison (2008) called such use *a flexible instalment opener* because "its lack of specificity means that they can begin an instalment of talk without having to commit to a more complex relationship between upcoming and prior talk from the outset of the turn, thus momentarily easing their processing load and

simultaneously occupying the turn" (p. 223). A number of excerpts from the BATTICC-F demonstrate this as follows.

(19)<TW01>: So ... How is your country in er the UK? Do you have the view... the same view?
<BT03>: No. We have similar gorges but they're nowhere near as dramatic.
<BT02>: ***And*** big. They're like, bigger.
<BT01>: Yes, much bigger.
<BT01>: ***And*** cleaner, by far.

(20)<BT10>: Er ... I'm not sure.
<BT10>: ***An:d*** what languages can you speak?
<TW09>: Only two ... Taiwanese and English.

In example (19) the informants were talking about the views in two different countries. We can see that students BT02 and BT01 preface their ideas with *and* to convey that they have more to say. They then add more details to the previous comment initiated by BT03 in what became a jointly constructed explanation of the scenic views in Taiwan. In example (20) *and* seems to be used to shift to a new topic in the conversation. In this case, the participants had just finished a discussion, and BT10 used *and* to initiate another question so that their communication still continued.

As a result *and* exemplified its flexibility in terms of giving options for the further development of the upcoming talk, as Evison (2008) claims. The examples from the BATTICC-F show that both Taiwanese and British pupils frequently used *and* to structure their discourse and continue the conversation in face-to-face interaction, while in the BATTICC-O, Taiwanese learners significantly underused *and*. According to Schiffrin (1987), *and* is "a structural coordinator of ideas which has pragmatic effect as a marker of speaker continuation" (p. 152). In this way,

Taiwanese learners' discourse appears to be less fluent and less cohesive than the British participants'. Carter and McCarthy (2006, p. 168) also note that in real-time communication "utterances are linked ... as if in a chain" rather than built into sentences, and coordinating conjunctions are therefore widely used to link clauses and sentences in a non-hierarchical way. As a result, this is the kind of language that students should aim at when taking part in an informal conversation, in both face-to-face and online communication.

Finally, I will revisit Table 4.8 and Table 4.9, where it can be seen that a number of multi-word expressions with coordinating conjunctions used by the British students are not found in the Taiwanese students' discourse, such as *and stuff, and everything, or anything* and *or something*. Fernandez and Yuldashev (2011) label these phrasal expressions general extenders (GEs), which are further divided into adjunctive and disjunctive GEs. The former usually begin with adjunctive coordinator *and,* implying that more detailed information could be given without actually saying so. The latter, on the other hand, typically begin with *or* and basically indicate "the existence of alternatives" (Overstreet & Yule, 2001, p. 50).

4.6.4 Modal auxiliary verbs

Modal auxiliary verbs are the most overused items by the Taiwanese participants in online communication. As can be seen in Table 4.10, the total number of modals used by the two groups out of the total of 31,910 words in the corpus of online discussion is 318 (2.16%) for the Taiwanese students and 194 (1.20%) for the British students. In CANELC the use of modal verbs accounts for 1.33% of the whole corpus, which is comparable to the use of modals in the British texts. A

study of modals in the even larger British National Corpus shows an average occurrence of 1.46% in overall spoken and written texts (Kennedy, 2002). It thus seems to be the case that these Taiwanese students use modals with much greater frequency than native English-speaking people.

Table 4.10 illustrates the comparative frequency of occurrences of different modal verbs. In CANELC and the British participants' texts the modals of volition and prediction outweigh all others, which is consistent with Montero et al.'s (2007) findings. In addition, the modal *would* is the second most used in the British participants' texts, and represents approximately a quarter of the overall incidence of modal verbs in both corpora of English-speaking people. In comparison, relatively little use of *would* is found in the Taiwanese discourse, with only seven tokens found in the corpus of 15,293 words.

Table 4.10

Frequency of Different Modals in Three Different Texts

Taiwanese participants		British participants		CANELC	
Modals	No.	Modals	No.	Modals	No.
can	182	will	58	will	1,669
will	88	would	56	can	1,605
must	17	can	55	would	1,271
may	10	could	8	could	636
could	7	should	6	must	559
would	7	may	5	might	366
should	6	must	4	may	353
might	1	might	2	should	262
Total:	318 (2.16%)		194 (1.20%)		6,621 (1.33%)

Table 4.10 also shows the Taiwanese students' preference for the modal *can*, with 182 occurrences out of 318 modals used. These results are consistent with those of

other studies (e.g., Montero et al., 2007; Yates, 1996), indicating that modals conveying ability and possibility (*can/could*) are the most widely used in CMC. Given the overuse of modals, one might hypothesise that the Taiwanese learners use them in sentences where a modal is not necessary, or that at least some of their use is not appropriate. By examining the concordance lines further, the following instances were found:

(21) Wow, your mum owns a restaurant. You are a good daughter because you ***can*** help your mum every day.
(22) My new year's hope is ***can*** learn swimming. What's your new year's resolution?
(23) Bubble milk tea is so great, you ***can*** love it.
(24) What kind of dance you dance? I ***will*** not dance any dance.
(25) Hi, I was born in Hualien, Taiwan. I ***will*** play the flute, But not very powerful.
(26) Yesterday he said he ***will*** help me do my homework.

(BATTICC-F)

In the first two examples above the modals do not seem obligatory; that is, the learners add a modal to a sentence that is not expected to have one. This type of use was found in more than 15 instances in the texts of Taiwanese learners. In example (23), however, the author may intend to express prediction and, as a result, *will* would be more appropriate. Examples (24) and (25) indicate the ability to carry out these activities and, consequently, *will* should be changed to *can*. In example (26) it seems that the author wanted to convey the meaning of prediction, but he/she forgot to choose the appropriate tense. These errors show that the Taiwanese students may well have difficulties distinguishing between *can* and *will*. The reason for this might be interference from their first language, namely Taiwanese Mandarin, because the modals *can* and *will* can both be translated as *hui* in Mandarin. In this case, learner errors provide evidence of the system of

language that is being used (Brown, 2007; Ellis, 2008), which it is very useful for EFL teachers to examine in order to better understand students' learning processes, for example, to discover how far towards the teaching goal the learner has progressed.

Table 4.11

Different Meaning Distribution of Modals (Taiwanese/British)

	Ability	Possibility	Permission	Inappropriate
can	29.4% / 32.2%	39.2% / 61.2%	16.7% / 6.5%	14.7%
could	80.0%/ 33.3%	0% / 50.0%	20.0% / 29.4%	-
	Prediction	Volition	Habitual actions	Inappropriate
will	55.8% / 44.1%	25.5% / 32.3%	4.6% / 23.6%	14.1%
would	0% / 40.7%	100% / 49.9%	0% / 9.4%	-

*'Inappropriate' here refers to the Taiwanese learners' inappropriate use of modals.

Given that the four modals *can, will, could* and *would* account for more than 85% of all the modal verb tokens in the present study, the distribution of these different meanings was further analysed. Interestingly, it is clear from Table 4.11 that the modal *could*, indicating possibility, is not found in the Taiwanese students' texts, with them mostly using *could* to express ability and permission. While the Taiwanese students only use *would* for expressing volition, the British participants use it to indicate three different situations, for example:

(27) It ***would*** be really nice to have a penpal to e-mail or write to
(28) I don't learn Chinese but I hope to learn some soon, I think it ***would*** be very fun to learn another language!
(29) I ***would*** like to be either a mechanic or a musician when im older.
(30) I ***would*** very much like to be friends with you.
(31) every Monday we ***would*** play the match for 35 minutes each way and have something to eat and set off back home at about 1pm.

(32) Hey Cindy, I ***would*** go jogging but I can't get up in the morning, I always sleep too much!

(BATTICC-F)

Examples (27) and (28) show the British students' prediction when they responded to the Taiwanese students' offers. In examples (29) and (30), *would* is used with *like* together to express their volition; *would* in examples (31) and (32) refers to habitual actions and events. However, in the Taiwanese texts, *would* is only used with *like* or *love* together to indicate volition, whereas no tokens were found for expressing prediction and habitual events. This indicates that the Taiwanese learners may have difficulties in the use of appropriate modal auxiliaries, and the underuse of *would* is clear. These results do not seem surprising since the Taiwanese learners in this project are at a beginner level. However, does this suggest that learners at this level have less chance of success in the acquisition of modal verbs? Although the reasons for this are complex, one of the most important factors might be their learning input, namely the learning materials.

To investigate further, I subsequently examined the textbooks used in Taiwanese junior high schools and some problematic aspects were found. For example, the presentation of modal verbs does not acknowledge that there are several contexts in which a modal verb is used; that is, most of the modals in the textbooks are presented in only one context. Their translation is also quite misleading, with only one or two meanings and no further explanations given. Learners, as a result, might not understand the polysemous nature of modal verbs and the subtle differences between them (see also Chang, 2003). Moreover, the modal *would* is generally presented at grade 8 level in most of the Taiwanese textbooks, which is

the sixth year of the compulsory English curriculum, and most other modal verbs are usually taught prior to this, such as *can, could, will, must* and *may*. Since *would* is one of the three most high-frequency modals, as shown in the native speaker data (see Table 6), I would suggest teaching it at an earlier stage. Similarly, research has suggested that frequent items should be taught earlier than less frequent ones (e.g., Römer, 2004; Schmitt, 2010); consequently, *would* should be among the first three modals to be presented in teaching materials. As a result, improvement of teaching materials and curriculum organisation are needed, and alterations may well lead to a better understanding and use of modals for language learners at a beginner level.

Celce-Murcia and Larsen-Freeman (1999) proposed an approach to presenting modals. They suggest that teachers present a range of modal verbs as constituents of a system so that the relationship between each modal is clear. For example, one of the uses of logical probability modals is to predict something, such as the chance of rain tomorrow. The example demonstrates to students what degree of prediction is expressed by each modal (or combination of modal and adverb) (ibid., p. 153):

(possibly)	weak, outside chance	It could/might rain tomorrow.
(perhaps)	stronger chance	It may rain tomorrow.
(probably)	even stronger chance	It may very well rain tomorrow.
(certainly)	certainty	It will rain tomorrow.

Presenting new modal verbs in this way would help EFL learners to understand the differences between different modal verbs and consequently students could learn the more precise meanings of each modal. Although this approach is helpful for learners to construct a system of modality, the presentation of modal verbs

should also be made appropriate to the student's level. The participants in this project, for example, are at a beginner-intermediate level, so using modality appropriately may be a complex task for them since on one hand, expressions of modality are often poly-functional, and on the other hand, EFL learners generally have limited opportunities to use modality in an authentic context (Montero et al., 2007; Römer, 2004).

This section has examined the key POS domains of BATTICC-O, in particular singular letter of the alphabet (ZZI), general adjective (JJ), coordinating conjunction (CC) and modal auxiliary (VM). These domains occurred with an unusual frequency in Taiwanese participants' discourse as compared with the British participants'. We now turn our attention to the investigation of key POS categories of the spoken data, BATTICC-F.

4.7 Key part-of-speech analysis: BATTICC-F

The keyness method applied at the POS level compared the relative use of grammatical categories of the two sets of texts from the Taiwanese and British participants' spoken communication. For $p < .01$, at 1 d.f., the cut-off of 6.63 indicated that 19 POS tags were significantly overused or underused between the two sets. At the 99.99% level ($p < .001$), this generated 10 significant POS tags. Table 4.12 displays these top 10 tags (with the largest LL values) that occur statistically unusually frequently (with +) and unusually infrequently (with -) in Taiwanese students' texts compared with British participants' data.

Table 4.12

Ten Most Significant Differences at Part-of-speech Level between the Taiwanese and British Discourse in BATTICC-F

Rank	POS code	Taiwanese participants		British participants		Overuse or underuse	LL	POS
		Freq.	%	Freq.	%			
1	VBDZ	9	0.13	128	1.07	-	67.36	was
2	UH	381	5.58	368	3.07	+	66.02	interjections
3	NN1	755	11.05	950	7.93	+	45.52	singular common nouns
4	VVN	19	0.28	119	0.99	-	35.23	past participle of lexical verb
5	PPH1	100	1.47	333	2.78	-	35.01	neuter personal pronoun: it
6	RR21	4	0.06	63	0.53	-	34.62	general adverbs – ditto
7	VVD	38	0.56	142	1.19	-	19.53	past tense of lexical verbs
8	VV0	268	3.93	338	2.82	+	16.05	base form of lexical verb
9	VBDR	2	0.03	29	0.24	-	15.38	were
10	VBZ	207	3.03	265	2.21	+	11.38	is

The table shows that the Taiwanese learners underused the categories of *was* (VBDZ), past participle of lexical verb (VVN), 3rd person singular neuter personal pronoun *it* (PPH1), general adverbs – ditto tags (RR21), past tense of lexical verbs (VVD) and *were* (VBDR), while they tended to overuse interjections (UH), *is* (VBZ) and singular common nouns (NN1) compared with the British participants. The list shows quite clearly that the majority of the underuse or overuse categories are related to the use of tenses. This seems to indicate that Taiwanese pupils demonstrate difficulties in gaining full control of appropriate use of tenses in their speaking, particularly present simple and past simple. We now turn our particular attention to the use of tenses in BATTICC-F.

4.7.1 Tenses

As can be seen from Table 4.12, the Taiwanese learners significantly underused the categories of *was* (VBDZ), *were* (VBDR), past tense of lexical verb (VVD) and past participle of lexical verb (VVN) and overused *is* (VBZ) and base forms of lexical verb (VV0) compared to the British students. This raises the question of whether the Taiwanese learners have difficulties with gaining full control of the appropriate use of tenses in their speaking, particularly for present simple and past simple tenses. Consequently, I further examined the scripts and found that many instances present the Taiwanese learners' mis-selections of the forms of appropriate tenses. For example:

(33) <TW03>: That Treasure Hunt ***is*** [was] pretty fun, right?
<BT04>: Yeah,
<BT05>: We tried to run to get back in time and we were ... we looked like

(34) <TW05>: Confusing why? I think the question ***is*** [was] boring.
<BT06>: They were a bit confusing.

(35) <TW07>: How about ... er ... remember today we ***are*** [were] climbing the mountain right?
<BT09>: Oh, that was really funny because I nearly fell over.

(BATTICC-F)

In examples (33) and (34) the Taiwanese students TW03 and TW05 used *is* to describe past activities, but their British peers responded with a past tense. Such a type of error by Taiwanese students was found in 12 instances when examining all of the concordance lines of *is* in BATTICC-F. Example (35) also shows the speaker's mis-selection of the appropriate verb form. Errors like this were found in 5 instances. Therefore, it seems that the Taiwanese speakers might have difficulties when using the appropriate forms of present simple and past simple tenses. In addition, the overuse of *is* (VBZ) by Taiwanese learners can be found in

the following sentences.

(36) <TW01>: Like our schools. Our schools ***is*** [X] have many clubs and student can

(37) <TW16> : The bacon and the sausage ***is*** [are]... it tastes like ... erm ... I just don't like it.

(38) <TW10>: Our Taiwanese name ***is*** ...our Taiwanese name ***are***

(BATTICC-F)

In example (36) the speaker TW01 used an additional *is*, which was not necessary; in (37) the speaker TW16 misused *is*, which was meant to be a plural form *are*. In (38), moreover, the speaker was possibly not sure whether he/she should use *is* or *are*, so the utterance was repeated with a changed verb. Another important phenomenon of tense-aspect marking by the Taiwanese speakers is the underuse of the past tense of lexical verbs (VVD) and the overuse of base forms of lexical verb (VV0). A closer look at the use of the lexical verbs in speech shows that in many cases it involves the non-marking of the past tense. For example:

(39) <TW03>: this morning .. er:: the first we ***go*** [went] to the trail...the trail ... Pretty tired and we ***climb*** [climbed] on rocks or something but we ***know*** [knew] we ... how to help each others and we ***have*** [had] fun in this activity, yeah ...
<BT04>: And we had to be sort of careful.. going up the rock face where people might fall.
<BT05>: I did fall off .... [laughter]
.....
<TW03>: And some difficult words .... we ... just ***listen*** [listened]....and ***hear*** [heard] you ... have ... we ***are*** [were] confused.
<BT04>: Yeah, like on the Treasure Hunt we went on later, there was the ... erm .... rhymes things and there was like ... there was confusing words on there that sometimes people wouldn't understand.

(BATTICC-F)

In this example the past time adverbial *this morning* by speaker TW03 has set the utterance in the past, namely the activities – trail walking and the Treasure Hunt –

which they had just done in the morning. Nevertheless, it can be noted that a number of lexical verbs used by the Taiwanese participant TW03 were in their default present tense, such as *go, climb, know, have, listen* and *hear*, which were meant to be used as *went, climbed, knew, had, listened a*nd *heard* respectively. This feature can be considered as an example of adjacent default tense (ADT), which means "if the overall tense of an utterance is marked in the context of the utterance, then, the 'adjacent' finite verbs in the utterance can (but may not necessarily) be set in their 'default' forms" (Xu, 2010, p. 69). Such "non-standard forms" (NSF) of tense-aspect marking were also identified by Kirkpatrick (2007, p. 157), particularly in Asian EFL learners' speech, which involves much non-marking of tense by using the base form of the lexical verbs.

As Ellis and Larsen-Freeman (2006) claim, "a prominent characteristic of interlanguage in English L2 acquisition is the lack of tense marking," and it "constitutes a significant hurdle to overcome in L2 learning" (p. 561). One of the most important reasons for the Taiwanese learners' difficulties in the use of appropriate tense of lexical verbs may well result from the influence of their first language (Brown, 2007; Ellis, 2008; Chen, 2010a; Xu, 2010). In this regard, Mandarin Chinese does not have a grammatical category of tense, and such a lack of tense inflexion in Chinese, therefore, has an impact on the frequency of verbal non-marking in Chinese-speaking learners' interlanguage. Chen (2010a), Xu (2010) and some other Chinese-speaking scholars have stated that this distinctive use of English has constituted one of the most significant syntactic features of Chinese English.

Although the Taiwanese learners appear to have some difficulties with using

appropriate tenses in speaking, their electronic discourse shows the opposite. On examination of more than 800 concordance lines of *is* in the Taiwanese text of BATTICC-O, few errors concerning the use of tenses were found (only four instances). That is to say, they used the present simple form *is* properly most of the time in online communication on electronic discussion boards, while when speaking they frequently made a tense error by using the present simple tense to describe past events. This can be partially explained by Krashen's (1981, p. 52) Monitor Hypothesis. As he claims:

> When we speak a second language, the forms we use come "first" from our subconsciously acquired competence. We then attempt to apply conscious rules, sometimes before we speak and sometimes not, sometimes successfully and sometimes not.

In this regard, "conscious learning acts as an editor, as a Monitor" when learners speak a foreign language (ibid., 1982, p. 16), and they are encouraged to utilise conscious rules to raise their grammatical accuracy in speaking when it does not interfere with communication. To that end, Krashen further suggests three conditions that need to be met in order for a learner to 'Monitor' successfully. The learners must:

> 1) have sufficient time to think about and use conscious rules effectively;
> 2) focus on the form of the utterance; and
> 3) know the appropriate rule that needs to be applied.

It appears that time is one of the most important issues as the speakers need enough time to access and use conscious rules. Hence in real-time communication,

due to time constraints, it may not be possible for learners to apply many rules to the utterances. In addition, even when having sufficient time, "we may be so involved in what we are saying that we do not attend to how we are saying it" (Krashen, 1982, p. 16). Like the face-to-face exchange activity in this study, tenses here were often neglected by Taiwanese learners in speaking as they might pay less attention to the monitor of different syntactic structures between Taiwanese Mandarin and English. Regarding writing on discussion boards, on the other hand, participants may feel less time pressure, and thus produce the language with higher accuracy. As in Jeng's (2010) study, discussed in section 2.2.2, the results revealed significant differences in the comparison of language outputs in CMC and FTF discussions among all variables, in which CMC discourse showed lower error rates, fewer dysfluency markers and higher percentages of using more sophisticated words.

4.7.2 Interjections

Apart from the use of tenses, the category that occurred with an unusual frequency is interjections (e.g., *yeah, oh, ah*), which normally refer to "exclamative utterances consisting of single words that do not easily fit into the major word classes (noun, verb, adjective, adverb)", and "all these items express positive or negative emotional reactions to what is being or has just been said or to something in the situation" (Carter & McCarthy, 2006, p. 224). They are particularly common in spoken discourse in that they provide speakers with useful interactional and organisational resources and serve to mark the boundaries of discourse units.

Yeah is by far the most frequent interjection overall in BATTICC-F. Jucker and

Smith (1998) claim that "the most frequent use of *yeah* is to acknowledge the receipt of information that is new to the discourse but consistent with current active information" (p. 179). In this regard, *yeah* serves as a "continuer response token" in that it helps to maintain the flow of the discourse and "encourage the current speaker to continue" (O'Keeffe & Adolphs, 2008, p. 84). Many cases of *yeah* in BATTICC-F occurred alone as a continuer response token, for example:

(40) <BT07>: And your recorder.
<TW07>: ***Yeah***.
<BT08>: That was really good. I was like how does.. I thought it was a recording.

(BATTICC-F)

Yeah may also function as a direct positive response to a question (Norrick, 2009), as can be seen in the examples (41) and (42):

(41) <TW07>: and er er in your country I know some .... some food is famous, right?
<BT07>: ***Yeah***, like fish and chips.
<TW07>: Fish and chips.

(42) <TW15>: Er ... we usually play basketball and volleyball in our school.
<BT17>: Ah .. volleyball?
<TW15>: ***Yeah*** – but I don't really like it.

(BATTICC-F)

Additionally, *yeah* may signal agreement with a statement in the foregoing turn (Norrick, 2009). As in the excerpt below, *yeah* simply means *I agree with you.*

(43) <BT09>: Our one was rubbish, but like the Taiwanese Dragon Boats are beautiful.
<BT07>: ***Yeah***, they really are. But ours are like several canoes tied together basically.

(BATTICC-F)

Yeah also occurred as "an initial transition word with no obvious positive response or agreeing function" (Norrick, 2009, p. 874), and as such, it was used to express

"a general acknowledgment of the previous interactive unit" (Jucker & Smith, 1998, p. 181).

(44) <BT10>: I'm not atheist, agnostic... I don't ...not quite sure if there is a God or not.
<BT11>: ***Yeah***, if God appeared then I'd say hi God, I'd believe. We're explaining agnostic.

(BATTICC-F)

It appears that *yeah* fulfills a number of different discourse functions. This confirms Fung and Carter's observation that it functions to "acknowledge, agree, affirm and mark continuation" (2007, p. 431), and fits with Evison's (2008) views on its versatility in academic talk. However, based on the key part-of-speech analysis in *WMatrix*, Taiwanese learners significantly overused the interjection *yeah* as compared with British pupils, with 2.29% and 1.51% respectively. Some of the instances in the BATTICC-F show that the Taiwanese learners frequently and sometimes continuously used *yeah* in response to a question or just to acknowledge the previous utterance as a back-channel token, as in (45) to (47):

(45) <BT09>: Have you lived in Haulian all your life?
<TW07>: ***Yeah, yeah.***
<BT07>: You know the dragon boat racing?
<TW07>: ***Yeah, yeah, yeah.***
<BT07>: What's it about?
<TW07>:..... You mean ... story, or ...

(46) <BT02>: I feel very happy because I won.
<TW02>: ***Yeah, yeah, yeah.***

(47) <BT18>: Do you like the fruit that we have here?
<TW16>: ***Yeah.***
<BT18>: The fruit, like strawberries.
<TW16>: ***Yeah.***

(BATTICC-F)

From these extracts it appears that *yeah* is pervasive in Taiwanese students'

utterances. Examples (45) and (46) display the continuous use of *yeah* in Taiwanese students' talk. Such reduplications were also observed in O'Keeffe and Adolphs's (2008) comparative study of response tokens in British and Irish English casual conversation[8]. Evison (2008) also reports that the repeats of interjections are frequent, remarking that *yeah yeah* is the most frequent turn-initiator cluster in academic talk. In BATTICC-F, there are more reduplications in Taiwanese discourse than in British English. Moreover, the overwhelming tendency of *yeah* to occur alone can also be found in Taiwanese participants' data, as in (47). In some extreme examples, Taiwanese participants even said *yeah* right after each sentence from their British interlocutors. These two phenomena can result in greater use of interjections in the Taiwanese dataset than the British one. A possible explanation for this might be that the Taiwanese learners probably have no idea how to continue the conversation, or possibly they do not even understand clearly the previous utterance, and thus reduplication and stand-alone *yeah* are the strategies they employed frequently to buy time for discourse planning.

Some other interjections that the Taiwanese participants significantly overused include *ah*, *wow* and *oh*, occurring roughly 2-3 times more frequently in the talk of the Taiwanese pupils than in that of the British participants. The *ah* in (48) probably means *why didn't you try it?*, indicating the speaker's surprise since *stinky tofu* is one of the most famous and must-eat foods in Taiwan. The *ah* in (49) signals agreement with the statement, which simply means *I see/I understand.*

(48)<TW17>: Do you like the stinky tofu?
<BT19>: I didn't try it.

[8] Their study remarks that there is more reduplication in Irish English than in British English.

<TW17>: ***Ah*** ... it's very good.

(49)<BT19>: No. We've had one earthquake but it was very small, it was just

<TW17>: ***Ah*** ... but our earthquake is always very big. (BATTICC-F)

Oh was also one of the commonest interjections, and this has assumed functions of "signaling a change in cognitive state" (Norrick, 2009, p. 868). In responses, *oh* is a signal of surprise, according to Aijmer (1987). However, Taiwanese learners use it appreciably more frequently than the British participants in BATTICC-F, neither signaling a cognitive change nor a surprised expression. As in (50), TW15 constantly used 5 *oh*s in a rather short conversation, and some *oh*s serve simply as a general acknowledgment of the previous utterance, which may signal the speaker's interest and involvement. Moreover, reduplications can also be found in the use of *wow*, as in (51). Although TW01 used four continuous *wow*s in his/her response, the meaning of these *wow*s is not at all clear from the context.

(50)<BT17>: What did you do in Australia?

<TW15>: ***Oh*** .. we went to school.... like local school and just stay in school. yeah

<BT17>: Did you visit anywhere?

<TW15>: ***Oh*** .. we visit the the Golden ... The Gold Coast.

...

<BT17>: ...we are all going when we are twenty and when we have a job.

<TW15>: ***Oh*** ... you can tell me when you're coming to Taiwan.

<BT17>: Yeah, I think I will tell everyone I have met.

<TW15>: And what's your school's name?

<BT17>: <School name01>.

<TW15>: ***Oh*** ... pictures everywhere.

<BT17>: Yeah. Are you coming, are you going to anymore schools in England?

<TW15>: ***Oh*** we're just going to ...

(51)<BT01>: What email address do you use?

<TW01>: *Wow, wow, wow, wow.*
<BT01>: Doing Facebook now.
<TW01>: Yes, that's good. (BATTICC-F)

This section has presented some examples of the overuse of interjections by Taiwanese participants. Although this did not result in misunderstanding or serious face-threatening situations in the samples of BATTICC-F, learners still have to be careful not to overuse interjections. Reber (2010) suggests that EFL teachers should teach the sound patterns and usages of interjections on the basis of naturally occurring discourse rather than referring to invented conversation examples. In this way, learners would gain some knowledge about when and how to display affectivity and appropriate responses in talk-in-interaction. This makes a valuable contribution to promoting the EFL courses and intercultural communication.

4.7.3 Other key parts-of-speech

Apart from the use of tenses and interjections, as can be seen from Table 4.12 there are some other key parts-of-speech. With regard to the domain of general adverb – ditto tags (RR21), many more instances were found in the British students' texts than the Taiwanese ones. Ditto tags encode the notion that a token is not an individual unit, but rather is a (somewhat non-compositional) part of a larger "idiom" (Dickinson, 2005, p. 46). The ditto tags, RR21 here, indicate the part of speech (RR – general adverb), the total number of elements in the idiom (2 in this case) and the position of the current word within the idiom (1 in this case). Examples (with the frequencies) within this category include *sort of* (15), *a lot* (12), *a bit* (10), *as well* (7), *of course* (2), *kind of* (2), *at all* (1) and *by far* (1) in British students' discourse, while only *as well* (2), *of course* (2), *a lot* (2) and *a bit*

(1) were found in the Taiwanese texts.

On the other hand, Taiwanese learners used singular common nouns (NN1) much more frequently than native speakers of English (11.72% compared to 9.06%). By investigating the concordance lines, however, nearly 15% of the instances show that learners failed to use plural nouns in required contexts, as in the following examples.

(52) We have many higher ***mountain*** and clean ***river***.
(53) They can join ***club***. So our ***student*** always have a fun evening in their ***club***.
(54) Yeah. You don't have ***earthquake*** in your country?
(55) Some ***school*** have football ***team***?

(BATTICC-F)

As shown above, many instances seem to show a strong tendency to omit *-s* from singular nouns in the Taiwanese learners' discourse. Research has also indicated that speakers of Asian languages such as Mandarin Chinese (which lacks the plural inflectional morphology system) have found incorrect plural morpheme use to be among the hardest grammatical errors to detect (e.g., Jia, 2003).

This section has presented the relative use of grammatical categories between Taiwanese and British participants' spoken discourse and has shown the significant differences between the two groups. The significance of the analysis of POS categories in BATTICC-O and BATTICC-F (presented in 4.6 and 4.7) lies in its ability to demonstrate that Taiwanese learner data display the extent to which grammatical categories were overused or underused, as compared to the discourse of native English-speaking people. Such findings that pertain to grammatical categories are not likely to be made when only frequency and keywords are

examined.

4.8 Summary

This study brings together the three levels of keyness analysis, namely keywords, semantic domains and parts-of-speech, in two case studies of adolescent intercultural communication, namely online and spoken discourse. Differences between Taiwanese and British participants' discourse were observed at each of the different levels. Comparisons at the POS and semantic levels reduce the number of key items that the researcher needs to examine, as was noted by Rayson (2008), thus addressing the most significant linguistic features that are key in both online and spoken discourse among young Taiwanese and British people. The semantic domain analysis revealed the themes that young people discussed commonly on the online discussion board and in face-to-face interaction, with a great number of cultural and social differences in terms of lexical choices. It is also evident that the two groups of participants asked each other questions when they were not sure about the vocabulary, some of which has culturally-determined meanings. Although paralinguistic strategies were underused by the Taiwanese students at the beginning of the project, they gradually learned from their British peers by observing these features in the online context. In this regard, the use of CMC provides authentic interlocutors and opportunities to join in authentic language and cultural practice in the target foreign language, taking students beyond classroom cultures and learner-to-learner communication (Hanna & de Nooy, 2003; Montero et al., 2007). Nevertheless, it is advisable that teachers choose online tools carefully, with a view to suiting their aims and their students' particular learning needs.

With regard to keyness analysis at the POS level, the use of modals by the Taiwanese participants deviated most prominently from the use found in their British peers' online discourse (BATTICC-O). The appropriate use of modal verbs appears to be among the most problematic areas for EFL learners. Examination of contextual use of modals suggests that part of the difficulty of English modal verbs for Taiwanese EFL learners is that they might not realise that most modals are polysemous, and consequently they tend to use a specific modal to express only a single meaning (Kennedy, 2002; Römer, 2004). This study has found that the most difficult modals for Taiwanese learners are *can/could*, indicating the meaning of possibilities, and *will/would*, indicating the meanings of habitual actions and volition. On the other hand, the POS analysis of spoken discourse (BATTICC-F) elucidates that the Taiwanese participants have difficulties in the appropriate use of tenses in their speaking. They also overused the categories of general adverbs – ditto tags (e.g., *sort of, a bit, as well*), while they tended to overuse *is*, singular common nouns and interjections, as compared with the British participants.

The results of the keyness analysis contribute to a better understanding of learner grammar and lexis and thus have important pedagogical implications for EFL teaching and learning. First, the analysis reveals a number of cultural and social differences in the use of words, semantic domains and parts-of-speech, and EFL teachers can therefore encourage their learners to observe these culturally relevant features and further develop their intercultural awareness and the skills of online communication (Montero et al., 2007; O'Dowd, 2007). Second, the keyness approach employs a macroscopic analysis to identity those linguistic features that deserve further attention, providing a reasoned basis for drawing learners'

awareness to linguistic features specific to the target text (Rayson, 2008; Tribble, 2000). Last but not least, keyness analysis is accessible and practically useful for EFL classroom teachers as a result of its ease and rapidity of use. While other techniques, such as Biber's factor analysis, require considerable expertise in data extraction and statistical analysis, the automated keyness method used in corpus analytical tools, such as *WMatrix*, allows users to automatically extract the most significant items without extensive training or effort.

To conclude, although keyness has been applied in a growing number of diverse studies, few have applied this approach in an EFL pedagogical perspective. While illustrating that manual checking of concordances and interpretation of results remain useful in addressing occasional errors in automated keyness analysis (Culpeper, 2009; Rayson, 2008), this chapter demonstrates its pedagogical merit via a case study. In doing so I have argued that keyness analysis can provide a better understanding of online and spoken discourse that allows learners a direct insight into the ways in which they and expert writers draw on lexical resources, and will help to inform EFL teaching for adolescent intercultural interaction.

CHAPTER 5

Online and Spoken Discourse: A Discourse Analytical Approach

5.1 Introduction

The previous chapter employed a keyness approach to identify particular lexical and grammatical features of online (BATTICC-O) and spoken discourse (BATTICC-F) by Taiwanese and British participants. While the keyness technique has proven to be useful in identifying statistically significant differences in language use between different groups of participants, this current chapter, based on a discourse analytical point of view, will add greater detail and depth of description of the language used in different communication modes. This analysis pays specific attention to the most distinctive linguistic features of online and spoken data that are not typically found in traditional written grammar. It first concentrates on the quantitative analysis of the primary linguistic features of online and spoken discourse by the two groups of participants. The linguistic patterns will then be examined in context to identify the pragmatic and discourse functions. The different use of these distinct features between Taiwanese and British participants in two different communication modes will also be presented.

5.2 Analysing online discourse

As presented in Chapter 3, the analytical framework for online discourse mainly includes four distinctive linguistic features of computer-mediated communication (CMC): (1) use of the upper and lower cases, (2) nonconventional spelling, (3)

emoticons and (4) punctuation omission and repetition. With regard to the use of upper and lower cases in BATTICC-O, I focus particularly on: (a) nonstandard capitalisation and (b) the use of lower case instead of upper case. To examine these features all the capitalised words were first generated using *Python* programming with the criterion of a minimum of two continuous capitalised letters (e.g., *NOW*), so that single-letter capitalised words, such as *I*, *Q* and *A*, would not be included. A number of generated words were then manually excluded, such as the ones that were not produced by the user (e.g., *SUBJECT, REPLY, RE*), the use of capitals for a heading (e.g., *CONNECTING CLASSROOMS*), acronyms (e.g., *OMG, BTW, BBC, UK*) and units of measurement (e.g., *KG, GB*). After the data cleaning was carried out, the instances of inconsistent capitalisation and minusculisation were counted to see their frequencies and how they were employed in context. Moreover, *WMatrix* was used to further identify the semantic and part-of-speech (POS) fields of all these capitalised words, exploring the extent to which participants preferred to capitalise particular categories of information in their messages.

The second distinct feature of CMC that I examine is nonconventional spelling, including (a) abbreviations (e.g., *pls* for *please*), acronyms (e.g., *BTW* for *by the way*) and substitution (e.g., *4* for *for* and *2day* for *today*) and (b) vocal spelling (e.g., *helooooooo*). Nevertheless, informal or spoken words that have been conventionally accepted and included in the major dictionaries were excluded from this category, such as widely used terms (e.g., *email* for *electronic mail*), the abbreviations for the week days (i.e., *Mon, Fri*), spoken words (e.g., *wanna* for *want to*) and so forth. Vocal spelling represents prosody or nonlinguistic sounds (e.g., *helloooo, ahhh)* and onomatopoeia such as *hehe, haha*, indicating an entry

that reproduces a sound. To identify the vocal spelling with repeated letters in BATTICC-O, the *Python* programme was employed to produce all the words with letters repeated at least three times so that I would not obtain all the words with normal double letters (e.g., *will, too*, etc.) which is not the focus of this analysis. However, not all the words produced within *Python* were included in this category. For example, *www* was found in eight instances in which the writer used it as part of a web address with hyperlinks to other related webpages, and *xxx* was used very often as an emoticon. The abovementioned repeated letters and any other random strings that included repeats were removed from this category.

I then analysed different types of emoticons used by the two groups of participants. I used the corpus extraction tool *WordSmith* to calculate the number of occurrences and each was then examined manually to decide whether an emoticon appeared intentional. For example, "My name:Peter" would have been recognised as an emoticon :P by the programme but was excluded in the total number of emoticons as it was not employed intentionally.

The last category I examined in this section is the use of punctuation in BATTICC-O, including (a) repeating punctuation marks (e.g., *!!!, ???*) and (b) apostrophe omission (e.g., *im* instead of *I'm, dont* instead of *don't*). I particularly investigated the repetition of exclamation and question marks, which is the most prevalent tendency in CMC as shown in previous studies (Cho, 2010; Frehner, 2008; Kalman & Gergle, 2009, 2010; Riordan & Kreuz, 2010). I then examined the use of apostrophes, which were very often omitted by CMC users (Riordan & Kreuz, 2010). The following section will discuss the analysis of each type of CMC feature in more detail.

5.3 The linguistic features in BATTICC-O

This section presents the four different types of distinctive linguistic features of CMC in BATTICC-O, including the use of the upper and lower cases, nonconventional spelling, emoticons and punctuation. The total instances and percentages of each type of feature in British and Taiwanese participants' discourse are summarised in Table 5.1.

Table 5.1
Total Instances of Different Types of Features in British and Taiwanese Datasets

Type of feature	Taiwanese		British		Sig.
	Number	per 1000 word	Number	per 1000 words	(LL)
Use of the upper and lower cases					
Nonstandard capitalisation	28	1.35	51	3.24	**-9.17
Lower case instead of upper case	83	4.89	884	57.22	***-852.9
Nonconventional spelling					
Abbreviation/Acronym/Substitution	28	0.47	63	2.78	***-29.85
Vocal spelling	65	3.82	122	7.90	***-23.52
Emoticons	140	8.23	181	11.71	**-9.90
Punctuation					
Repeating punctuation	88	5.18	164	10.61	***-31.11
Apostrophe omission	6	0.35	118	7.64	***-134.8

$^{**}p < .01$ $^{***}p < .001$

From the table it can be seen that the inconsistent use of upper and lower cases is the most prevalent in the data in terms of the total amount of use (967 instances). This is followed by the use of emoticons (321 instances), repeating punctuation (252 instances) and vocal spellings (187 instances). In contrast, the category of nonstandard capitalisation is used to a lesser extent, with a total of 79 entries in BATTICC-O. Using log-likelihood ratio (Rayson, 2008) to compare the cumulative frequencies reveals significant differences in the use of each category

between Taiwanese and British participants. Surprisingly, in all of the categories investigated the different use by the two groups reaches a significant level. This shows that the British native speakers of English tend to employ relatively more linguistic or paralinguistic cues in CMC as compared with the Taiwanese learners. These features will be discussed individually in more detail in the following subsections.

5.3.1 Use of the upper and lower cases

(a) Nonstandard capitalisation

The use of nonstandard capitalisation and minusculisation is one of the most common typographic features of CMC discourse where there is a great deal of variation. Due to the lower case default nature of a keyboard, a strong tendency to use lower case can be widely seen in online communication to avoid awkwardness (Crystal, 2011). This means any use of capitalisation delivers a marked form of communication. As has been shown in Table 5.1, there are 79 entries of nonstandard capitalised words found in BATTICC-O (0.22% of all words), demonstrating that both Taiwanese and British participants frequently employ capitalisation on specific information in a message. Nevertheless, the quantity of use by British participants is significantly more than the use by Taiwanese participants (LL=-9.17, p <.01), with 28 and 51 instances respectively. The following excerpts present how nonstandard capitalisation is used in BATTICC-O, in which (1)-(6) are retrieved from British participant discourse and (7)-(10) are produced by Taiwanese participants.

(1) Aw thank you cindy, and I **LOVE** the picture!! I'm going to print it off and keep it!! :D (BT)
(2) I also play the Clarinet in the school Orchestra which I **LOVE**! (BT)
(3) **PAIGE** is my bestfriend :D:D:D (BT)

(4) **HAPPY NEW YEAR!!!!!** (BT)
(5) I hate P.E, history, geography, maths, science, german, R.E, PSHE, ICT, music and Literacy.. Or you can just say I hate **EVERY SINGLE LESSON**! (apart from Art and Food Tech..) (BT)
(6) **9 WEEKS AND 1 DAY TO GO!!!!!!!!!! I AM SOOOOOOOOOOOOOO EXCITED!!!!!!!!!!** (BT)
(7) I like every holiday ~!! Because we can staying at home **ALL** day. (TW)
(8) My favorite day is~~**CHRISTMAS**!! (TW)
(9) **HEY** My English name is **SOLO**. (TW)
(10) This is **SOOOO BEAUTUFUL**!!! (TW)

(BATTICC-O)

From the above extracts it can be noted that capital words are used to serve different functions. They help to intensify authors' affect or attitude (e.g., (1), (2)), signal tone of voice (e.g., (4), (6)), or simply strengthen the particular content of the message, such as when referring to names (e.g., (3), (9)), festivals (e.g., (8)), or other information that the writer thinks is particularly important (e.g., (5), (7), (10)). For example, in (2), the word *LOVE* is capitalised as the writer intends to show intensified interest in the school orchestra. In (3), the wholly capitalised name *PAIGE* may indicate the author's intention to strengthen the importance of such information. Similar use can be found in Taiwanese learners' discourse: in (7)-(10) the authors' intention to emphasise the particular information in the message is indicated. Although messages wholly in capitals are considered to be "shouting" and netiquette guides generally suggest that they should be avoided (Crystal, 2006, 2011), examples like (4) and (6), which show the writer's strong feelings and excitement, are found in 7 and 5 instances in British and Taiwanese participants' discourse respectively.

I further categorised all the capitalised words to the related semantic and POS domains using *WMatrix*. Tables 5.2 and 5.3 illustrate the five most common semantic and POS categories of the nonstandard capitalisation in BATTICC-O.

Table 5.2

Five Most Common Semantic Domains of Capitalised Words in BATTICC-O

Rank	Code	Semantic domain	Taiwanese	British	Total	Examples
			Freq.	Freq.		
1	Z99	Unmatched	8	17	25	PICO, QUEGS, SOOOO
2	T1.3	Time: Period	6	8	14	SUMMER, DAY, MONDAY
3	E4.1	Happiness	3	5	8	HAPPY
4	Z4	Discourse Bin	4	4	8	HEY, WOW, BYE
5	E2+	Like	2	6	8	LOVE

Table 5.3

Five Most Common Part-of-Speech Domains of Capitalised Words in BATTICC-O

Rank	POS code	POS	Taiwanese	British	Total	Examples
			Freq.	Freq.		
1	NN1	Singular common noun	7	8	15	CHOCOLATE, BIRTHDAY
2	VV0	Base form of lexical verb	3	8	11	LOVE, HATE, WAIT
3	JJ	General adjective	4	6	10	HAPPY, GOOD, NEW
4	NP1	Singular proper noun	2	7	9	PICO, GAGA, MSN
5	RR	General adverb	1	7	8	REALLY, MUCH, ONLY

It can be seen in the semantic categorisation that unmatched words (Z99) are predominant since many of them include proper nouns or names for particular groups of people or organisations, as well as some nonstandard spellings (e.g., *SOOOO*). The analysis of semantic domains also indicates that participants show

very positive emotions in intercultural exchange in which the semantic categories of Happiness (E4.1) and Like (E2+) are very frequent. Regarding the parts-of-speech, the most common domain of capitalised words is singular nouns (NN1, NP1), followed by lexical verbs (VV0), adjectives (JJ) and adverbs (RR). While British participants used these five domains for capitalisation quite equally, Taiwanese learners particularly prefer to capitalise singular common nouns in order to strengthen the particular content of the message. It appears that participants generally have some kind of preference for the use of nonstandard capitalisation in their CMC messages, showing that it is not simply employed indiscriminately.

(b) Lower case instead of upper case

On the other hand, many messages are written entirely in lower case even when the upper case is expected. Such tendency to reduce the use of capitalisation can be linked to linguistic economy in CMC. Instances of this type account for nearly one quarter of the messages in British participants' discourse (152 out of the total 683 messages), as illustrated in (11) to (13). From these excerpts the use of lower case instead of upper case is apparent, such as for the first letter of a sentence (e.g., *his, it, do, what*), for names (e.g., *pico*), for the first-person pronoun *I* (e.g., *i*) and for proper nouns (e.g., *buddhist*).

(11) Thanks, **do** you have any pets? (TW)
(12) **his** name is very cool and so is he, **pico, what** a cool name. =] (BT)
(13) **it** is ok and today **i** am going to a **buddhist** meditation centre on a school trip, **i** think it will be fun and interesting. (BT)

(BATTICC-O)

Table 5.4 presents the total instances of lower case instead of upper case use in the British and Taiwanese datasets. The table clearly shows that there is a stronger

tendency to minimise capitalisation in the British data than the Taiwanese. For example, the British data generates 205 proper nouns using lower case instead of upper case, including personal names (125 out of 277 instances), geographical names (69 out of 236 instances) and other proper nouns (11 out of 24 instances). In contrast, only 4 instances of minusculisation for geographical names are found in the Taiwanese learner data. This indicates that the Taiwanese learners generally follow grammar rules strictly.

Table 5.4

Total Instances and Percentage of Using Lower Case Instead of Upper Case

Minusculisation	Taiwanese			British		
	upper	lower	rate	upper	lower	rate
First letter of a sentence	1502	40	2.59%	945	395	29.48%
First-person Pronoun *I*	920	39	4.07%	709	284	28.60%
Proper nouns	465	4	0.85%	332	205	38.18%

5.3.2 Nonconventional spelling

Spelling practices in CMC often suggest "loosened orthographic norms" (Herring, 2013). This section presents nonconventional spelling from two perspectives: (a) lexical reductions, in which users economise on typing effort (i.e., abbreviation, acronyms and substitution) and (b) creative use of language, in which users mimic spoken language features or express themselves creatively (i.e., vocal spelling).

a) Abbreviation, acronyms and substitution

It is accepted that abbreviation, acronyms and substitution are effective economical means of communication online. In the following excerpts some abbreviations can be seen in the messages in which words are shortened by removing one or more phonemes or morphemes. For example, some of them are

phonetically motivated, such as *wud* for *would*, *wat* for *what* and *hav* for *have*. Such pseudo-phonetic spellings are found in 8 instances in BATTICC-O.

(14) My Favee Sweet Is Gummy Bears!!! (BT)
(15) what is everyone else frm cockermouth bringing? (BT)
(16) Wat do u like to do after school? Plz respond i wud love to here wat u hav to say (BT)
(BATTICC-O)

Moreover, some of the abbreviations found in BATTICC-O represent the omission of vowels, as in *plz* for *please* in (16) and *frm* for *from* in (15). Such consonant spelling is one effective way of communicating economically in CMC, and as Thurlow (2002) claims, consonants usually have more semantic detail/value than vowels. Nevertheless, the consonant spelling is not very frequent in my data, with only 5 instances found in British participants' discourse.

The substitution of numbers or letters for words or parts of words is common in BATTICC-O in that the CMC user can reduce keystrokes and/or symbolise a playful communication style or social identity (Herring, 2013). For example:

(17) nice to see every1 too!! (BT)
(18) i agreeeeeee with u m8 I signed up 4 the earthquake monitoring (BT)
(19) r u guys wearing shorts or skirts (BT)
(20) I'm going to print it off n keep it!! haha (BT)
(21) How are u everyone? (TW)

(BATTICC-O)

Substitution can be achieved by the use of a number whose phonological content is equal to a word or a part of a word, as in (17) and (18) where the number homophones 1, 8 and 4 are used to replace *one*, *ate* and *for* due to the same pronunciation. Similarly, a word is abbreviated to a single letter so that a letter homophone can stand on its own and refer to a single word. For example, the

letters *r* and *u* in (19) refer to *are* and *you* respectively, and the letter *n* in (20) refers to the coordinating conjunction *and*. Such lexical reductions can be found in 51 instances in my data, where the homophone *u* is most prevalent (18 instances), followed by *n* (9 instances) and *r* (7 instances). In addition, the use by Taiwanese and British participants differs markedly. The Taiwanese learners only use the homophone *u* for substitution (6 instances), while *n* and *r* are not found in their discourse for such use.

With respect to the overall instances of abbreviations, acronyms and substitutions, as presented in Table 5.1 a total of 28 and 63 entries were found in Taiwanese and British participants' datasets respectively. The log-likelihood analysis elucidates significant differences regarding the amount of use by the two groups of participants (LL=29.85, p <.001). While the British participants produced significantly more items classified as this type of feature (0.278%), the actual amount of use is somewhat less than the findings of previous research on this CMC feature employed by native speakers of English, such as the 0.587% in Frehner's (2008) study. In addition, some common forms of nonconventional spelling in other text-based discourse cannot be found in BATTICC-O. For example, *g*-clippings, which indicate a tendency of users to omit the final letter *g*, such as in *goin* for *going* or *darlin* for *darling*, can easily be seen in CMC. Frehner (2008) reported that 36% of the letter *g* in the words ending with *ing* is clipped off in text messaging, but the participants in the study rarely use *g*-clippings.

b) Vocal spelling

Another type of nonconventional spelling is vocal spelling, which is not for the

purposes of economical language use in CMC; instead, it requires more keystrokes from the users. Vocal spelling represents prosody or nonlinguistic sounds, emulating a stretched out syllable in spoken discourse. In all 187 instances of this feature were found, a total of 5.76 entries per 1000 words in BATTICC-O. As can be seen in Table 5.1, the amount of vocal spelling used by British and Taiwanese young learners differs significantly (LL=23.52, p <.001). Within this type the majority of repeating letters indicate the stretching of a lexical word, with the intention of adding more emotional emphasis. Harris and Paradice (2007) labeled these *emotional cues*, which indicate the type and degree of emotion that the message sender intends to convey. For example:

(22) I think school is sooooooooooooooo boring and silly. (BT)
(23) Mariah's voice is sooooooooo good. (TW)
(24) My school is gooooood!!!!!!!! (TW)
(25) I agreeeeeee with u m8 (BT)
(26) have you done much today ezzzzzer (BT)
(27) My bestfriend is NATASHAAAAA (BT)

(BATTICC-O)

As in (22) and (23), the adverb *so* with *o* repeated is used to indicate the writer's strong feeling about the issue he/she is talking about. Such use of *so* is the most common word used for vocal spellings in my data, with 16 instances found. Moreover, interestingly, names are often written in a vocal spelling in BATTICC-O. In this way the writer seems to address the message to a particular person and intends to catch his/her attention, as in (26), or this is done simply for emphasis, as in (27), which is often in conjunction with other cues like capitalisation. Such a grapho-phonemic play with vocalisation and voicing by the participants gives their utterances more of a pattern-reforming flourish. This also indicates self-dramatisation of the CMC user, which may result from the detached

nature of friendly interpersonal exchanges (Carter, 2004, p. 197).

POS and semantic categories of vocal spelling were also generated using *WMatrix*. In BATTICC-O vocal spelling occurs most often in the POS domain of Degree Adverb (RG) (e.g., *sooooo, toooo*), followed by the domains of Interjection (UH) (e.g., *hellooooo, ahhhh*), Base Form of Lexical Verbs (VV0) (e.g., *waiitttt, looovveee*) and Singular Common Noun (NP1) (e.g., *schooool, ezzzzer*). With regards to the semantic categories, Degree: Boosters (A13.3) (e.g., *soooo, veeery*) is the most common, followed by Discourse Bin (Z4) (e.g., *heeey, ooooh*) and Like (E2+) (e.g., *looovveee*). This seems to indicate that vocal spelling exhibits emotional cues (Harris & Paradice, 2007), tempo, pitch, prosody or other paralinguistic elements (Kalman & Gergle, 2009; Riordan & Kreuz, 2010).

With regard to the total amount of repeated letters as vocal spelling in BATTICC-O, 60.9% (179 items) are of vowels and 39.1% (115 items) are of consonants. We can also see that certain letters are more frequently repeated than others. In particular, *o* is the most prevalent, with 118 instances found in all the vocal spelling words. This is followed by *e* (49 items), *h* (30 items) and *m* (19 items). These four highest-frequency repeated letters are the same as the ones found in the large general e-language corpus CANELC, with *o, e, h* and *m* in 692, 303, 299 and 264 instances respectively. It seems that *o* (e.g., *soooo, noooo*) and *e* (e.g., *pleeeeeze, agreeee, seeeee*) are the vowels repeated most frequently, while the most commonly repeated consonants are *h* (e.g, *ahhh, ohhh, yeahhh*) and *m* (e.g., *Mmmmm, hmmm, yummmm*). While vowels are more commonly repeated as vocal spelling in BATTICC, the use of repeated consonants (54.4%) is slightly more frequent than vowels (45.6%) in CANELC. This notwithstanding, the

average of repeating letter instances per vowel are much higher than consonants, and continuant consonants (e.g., *m, h*) can be more commonly found than plosive consonants (e.g., *p, b*). This is perhaps due to the audible and articulable nature of vowels and continuant consonants, which allow the continued flow of air and are more easily continuously articulated than other sounds to reflect vocal intonation (Kalman & Gergle, 2010). Riordan and Kreuz (2010) note that communicators tend to repeat letters that they would stress while speaking rather than repeating just any letter in a word.

While the majority of the repeating letters were used to add additional stress, to mimic spoken language features or simply to achieve visual emphasis, nearly one third of the items identified in vocal spelling indicate an entry that reproduces a sound. For example:

(28) I think I can as long as there are no buiscuits in the house :/ ***ohhh*** dear (BT)
(29) Hey My names Emma.. but my friends call me Em (I hate that name!) ***Hahahaaaaa***. (BT)
(30) but its open today and I have a maths test!! ***Ahhhh***.I am really bad at maths haha (BT)
(31) Haha thanks 149kepti, plus that time with indya when she got kn bella and she took one step and indya screamed '***ahhhhh***! (BT)
(32) cant wait to get out there 25 day ***whooooooooppp*** :P (BT)
(33) ***Haha*** it is very funny. I laugh and laugh..... (TW)
(34) the teachers were like going keptical in a very quiet way! ***Haha*** very funny indeed!! (BT)
(35) wow its so close now!! ***Haha*** :D (BT)

(BATTICC-O)

Among all the instances of vocal spelling, *haha*, which represents the sound of laughter, was the most common. It is pervasive in BATTICC-O, as can be seen in (33)-(35), where *haha* is mainly used to express a positive emotion. With regards

to the placement in the textual utterances, the majority of *haha*s appear at the end of the utterance, accounting for 64% of the total instances, but in many instances they also act as turn-openers, as in the first *haha* in (33). This turn-opening function was found in 29 instances, which account for 19.1% of the total instances of *haha* found in BATTICC-O. Moreover, 14.1% of the *haha*s occur in the middle of the message, connecting two clauses or phrases, as in (34), and 2.8% occur by themselves as individual utterances.

In many cases, however, the vocal word *haha* may not be used as the "real" sound of laughter, but rather as an indication of the illocutionary force of the textual utterances that they accompany. As can be seen in (36)-(40), the *haha* may not always indicate happiness or the sound of laughter in the context, but rather it can be pragmatically specialised as a downtoner in that many instances indicate a preference for collocating with negative statements.

(36) I can't swim inside very much because chlorine is ***not*** great for me ***haha*** (BT)
(37) I used to be very fluent when I was little but now I'm ***not*** very good ***haha*** (BT)
(38) I I like long distance because I am ***not*** a very good runner ***haha*** :D
(39) she used to run under peoples feet so she got stood on quite a bit which ***wasn't*** great ***haha*** (BT)
(40) are they flowers? I'm ***not*** sure ***haha*** (TW)

(BATTICC-O)

Such a use of *haha* can be found in 19 instances in my data. This is probably because the writer is not serious about the content of the message and intends to develop a more informal and relaxed tone of conversation, as well as to demonstrate a humorous way of indicating the negative aspect or uncertainty.

5.3.3 Emoticons

Throughout the BATTICC-O, 321 instances of emoticons of 29 different types were found, accounting for 0.99% of the whole dataset. Table 5.3 presents the variety and total instances of each emoticon used by Taiwanese and British participants. From the table it can be seen that *:D*, representing a big smile, was the most used emoticon, followed by the general happy smiley *:)* or *:')* and the rhetorically playful emoticon *:P* or *>P*. They are pervasive in my data in that they can be inserted in different places in the message: at the beginning or the end of the sentence, as in (41), or in the middle of a sentence or between two clauses, as in (42)-(44).

(41) =), no I don't learn keptic but I hope to learn some soon, or in Taiwan, I think it would be very fun to learn another language!! :P (BT)
(42) just a load of people standing round and one person from one school writing all the names in n I was like :0 I wanted to do that lol (BT)
(43) he is naughty :p but I'm sure he won't do it again now! (BT)
(44) I love cake :P and chocolate :P (TW)
(45) Haha yes, that would be so funny :P (BT)
(46) My birthday is coming :D I am very excited :D (TW)

(BATTICC-O)

Emoticons can also substitute some forms of punctuation, especially full stops at the end of a sentence, as in (44)-(46) and commas between clauses, as in (44). The data shows a strong tendency of the participants to leave out full stops and use emoticons instead, which is found with 48.1% (87 of the 181 instances) and 41.4% (58 of the 140 instances) in the British and Taiwanese datasets respectively, while the substitution of commas or other punctuation occurs only in 10 instances in total. Such an observation was also made in Frehner's (2008) study of e-mail and text messages, which showed that 52-76% of all the full stops were replaced by emoticons in the corpora examined, while less than 1% of exclamation and

question marks were substituted by emoticons. Accordingly, the findings of the present study along with previous studies illustrate the tendency that emoticons can serve a certain function of punctuation in a CMC context. In my data analysis they replace nearly half of the full stops in both British and Taiwanese participants' discourse, which may result from the economical nature of CMC.

Table 5.5

Numbers of Each Emoticon Type by British and Taiwanese Participants

Emoticons	Translation	Taiwanese	British	Total
:D or ;D	Big smile	34	49	83
:) or :') or :-)	Happy/smile	43	36	79
:P or >P	Playfulness	13	19	32
x sequence	Kisses	0	20	20
XD	Big smile	9	7	16
=] or :] or :-]	Happy/smile	2	11	13
x	A kiss	0	12	12
^^ or ^_^	Happy/smile	18	0	18
=)	Happy/smile	5	3	8
:/	Sceptical smiley	0	7	7
:S	Confusion	0	6	6
:L	Laughing	0	5	5
=-O or :O	Surprise	3	2	5
;) or ;-)	Wink/Joking	0	4	4
(^(00)^)	Little pig	4	0	4
T_T	In tears	4	0	4
:(or :-(	Frowning/sad smiley	3	0	3
:目	Laughing	1	0	1
\(ㄒ.ㄒ)/	In tears	1	0	1
Total		140	181	321

With regard to the total quantity of emoticons, British participants generally use them more frequently than Taiwanese participants. As shown in Table 5.5, the difference regarding the total instances between the two groups of young learners reached a significant level (LL=9.90, $p<.01$). The number of some types of

emoticons differs markedly between the two groups of users. For example, *:D, :P* or *>P* and single/multiple *x* are used appreciably more frequently by British participants, while relatively fewer instances of such types are found in Taiwanese participants' discourse. Some emoticons are not even used at all by the Taiwanese learners, such as the wink/joking emoticon *;)* or *;-)*, the confused expression *:S*, the sceptical emoticon expressing uncertainty *:-/* or *:/* and kisses *x* and *xx*.

In particular, the use of single or multiple *x* was found to be high frequency in British participants' data, as shown in the following excerpts (47) to (51), while no instances of such emoticons were found in Taiwanese learners' discourse. From the excerpts it can be seen that *x* or repeated *x* mainly appears at the end of the textual utterance and shows a positive and a happy emotional state of the user.

(47) lol hahahahahaha funni !!! x (BT)
(48) what type of music do you listen to in Taiwan? x (BT)
(49) whats yr fav subject ????? xxxx (BT)
(50) Well, we like sunday roast and we like fish and chips xxx (BT)
(51) PAIGE is my bestfriend :D:D:D xxxxxxxxx (BT)

(BATTICC-O)

While it is widely accepted that emoticons serve as non-verbal indicators of emotion, in many cases they act as indications of the illocutionary force of the textual utterances that they accompany (Dresner & Herring, 2012). That is, some emoticons are not to convey emotion but rather pragmatic meaning, showing the writer's intention in producing that message. In the excerpts (47) and (51) the happy expression of the writer is apparent, while in (48)-(50) the use of emoticons is less straightforwardly affective. For example, although (48) and (49) are in a questioning form, the use of emoticons attached to the textual utterances is likely to present a rather informal setting in that the writer is not too serious about the

content of the message. This may further constitute a less face-threatening speech act. According to Dresner and Herring (2012):

> These uses of emoticons do not contribute to the propositional content (the locution) of the language used, but neither are they just an extra-linguistic communication channel indicating emotion. Rather, they help convey an important aspect of the linguistic utterance they are attached to: what the user intends by what he or she types. (p. 62)

In addition, ^^ or its derivations, such as ^_^ or ^_____^ serve a similar discourse function as *x* and *xx*. However, ^^ is widely used by Taiwanese participants, as shown in the following excerpts (52)-(55), while no instance of such a type of emoticon is found in British participants' discourse. It appears that cultural variation in the use of emoticons can be found in BATTICC-O, in which Taiwanese young learners prefer to use ^^ or its derivations to express a happy emotion or a particular illocutionary force, while the British participants generally tend to use *x* and *xx*.

(52) Hope you can come here and taste the stinky tofu.^^ (TW)
(53) I love chirstmass^^ (TW)
(54) Hi, My name is Cindy. I am 13 years old. ^_^ (TW)
(55) Please go to Hualien. *^_____^* (TW)

(BATTICC-O)

Another emoticon that was used by Taiwanese learners but not found in British participants' discourse is *(^(00)^)*, which looks like a cute little pig smiling and tends to show the users' playfulness and creativity, as in (56) and (57).

(56) Hello! Nice to meet you. (^(00)^) (TW)
(57) How about you? What kind of chocolate do you like? (^(00)^) (TW)

(BATTICC-O)

On the other hand, several emoticons, such as *:/, :S* and *;)* or *;-)*, were not used by Taiwanese learners, but they were common in the British participants' dataset. Examples of these emoticons are illustrated in (58) to (62). The emoticon *:S* is mainly used to express confusion, as in (58) and (59). Moreover, *:/* may indicate the writer's uncertainty. It can be seen in (60) that the description from the friends of the writer seems to contradict his/her own opinion about his/her appearance so that the emoticon *:/* is used followed by his/her opinion to present the uncertainty, collocating with the hedging *I think* and the laughter *haha* to soften and downtone the assertive force of the utterance. Similar pragmatic functions of the emoticon *:/* can be seen in (61) and (62). We can see *I think* and *haha* in (61) and *a bit* in (62). This also shows the writer's attitude and serves a pragmatic function in the sense that they are used as part of interpersonal strategies to hedge the assertion.

(58) could be these ones but not sure :S (BT)
(59) and oh' it all seems so confusing :S (BT)
(60) my friends describe me as tall~ish and slim but *i think* i am fat :/ *haha.* (BT)
(61) *I think* I can as long as there are no biscuits in the house :/ ohhh dear *haha* (BT)
(62) wow that sounds cool our school uniform is a blue, white and really dark blue tie, white shirt, black trousers or skirt and a black blazer. :/ its ok but its *a bit* uncomfortable :/ (BT)

(BATTICC-O)

It appears that emoticons can be seen as a pragmatic or paralinguistic device that conveys the writer's emotion and illocutionary force and further facilitates the understanding of the message. Lo (2008) reported that most of the participants in his CMC project could not determine the writer's emotion, attitude and intention when they are simply shown with messages without emoticons. However, when emoticons are added in the same context, the reader's perception of the messages

change significantly, which indicates that the use of emoticons helps to decrease the ambiguity of text-based messages.

5.3.4 Punctuation

a) Apostrophe omission

Another remarkable feature in BATTICC-O is the omission of apostrophes, which also results from the economical nature of CMC discourse. As can be seen in the following examples, the apostrophes in *I'm*, *can't*, *it's* and *don't* are intentionally deleted. As shown in Table 5.1, 118 instances of apostrophe omission are found in British participants' discourse, while there are only 6 entries in the Taiwanese dataset. For example:

(63) ***Im*** sure it will still be fun (BT)
(64) at the moment I ***cant*** play outside because of all the snow!! (BT)
(65) ***its*** funny, my dad wakes me up because my mum sleeps even more than me (BT)
(66) I ***dont*** know that (TW)

(BATTICC-O)

Table 5.6 presents the four items whose apostrophes are most frequently omitted by the participants. It can be noted that in the total instances of the four items, on average more than half (58.4%) of the apostrophes are omitted by British participants. This high percentage of omissions may be for the purpose of saving typing effort. From the table, in particular the word *it's* always occurs in its apostrophe omission form in the British dataset, namely *its* (91.2%). Although it may refer to the third person possessive pronoun, no instances of this form are found in the British dataset, while all of the instances of *its* in the Taiwanese dataset are the third person possessive pronoun.

Table 5.6

Apostrophe Omission by Participants

	Taiwanese			British		
	apostrophe	omission	omission	apostrophe	omission	omission
I'm	95	2	2.1%	40	40	50.0%
it's	50	0	0.0%	3	31	91.2%
can't	12	1	7.7%	19	18	48.6%
don't	34	1	2.9%	18	14	43.8%

b) Repeating punctuation

Both British and Taiwanese participants' discourse shows evidence of punctuation repetition, which manifests itself mainly in exclamation marks and question marks. Lee (2003) states that "multiple question marks or exclamation points signal the reader to intensify the degree of rising intonation in a question or the loudness of an assertion" (p. 320). The grammatical function of exclamation marks is described as indicators of "emotive force" (Quirk, Greenbaum, Leech, & Svartvik, 1985, p. 1,633), or as a means to demonstrate that a "preceding word, phrase or sentence is an exclamation or strong assertion" (McArthur, 1992, p. 394). As such, they are generally used for exclamatives and typically occur after interjections (Carter & McCarthy, 2006).

(67) Really!!!!! (TW)
(68) WOW!!! Taiwan looks amazing!!!! (BT)
(69) Aw no!!! has he been found yet?? (BT)
(70) how can you hate ketchup!!!!!! (BT)

(BATTICC-O)

As in (67)-(69), the interjection ends in consecutive exclamation marks to express an emphasis of extra exclamation. An exclamation mark is also intended to show astonishment. As in (70), the writer may feel astonished about the message saying that someone does not like ketchup. Although a single mark is the norm, in CMC

discourse multiple exclamation marks are pervasively used for additional emphasis, particularly in an informal situation (Carter & McCarthy, 2006; Kalman & Gergle, 2010).

Repeated exclamation marks are occasionally simply used as an intensifier, which is used to emphasise the previous textual utterance, as can be seen in (71) and (72).

(71) I absolutely cannot waiitttt !!!!!"!!!!!!!!!!!! (BT)
(72) I am sooooooooo excited!!!!!!!!!!!!!!!!!!!!! (TW)
(73) ONLY 64 DAYS TO GO NOW!!!!!!!!!!!!!!!!! (BT)
(74) My favourite food is chocolate!!!!! (BT)
(75) i LOVE CHOCOLATE!!!!!! (BT)
(76) HAPPY BIRTHDAY!!!!! (TW)
(77) Merry christmas to you too!!! (TW)

(BATTICC-O)

The use of multiple marks as paralinguistic cues in CMC often occurs in conjunction with other cues, such as adverbs *so* and *absolutely* for emphasis and capitalisation, as in (71) to (73). The combination of different cues in CMC can indicate stronger feelings of the writer. Moreover, repeated exclamation points can be used to indicate high volume (shouting), as in (76) and (77). It seems that the repeats of punctuation indicate a sense of emotion by using a different pitch, prosody or other paralinguistic elements. They can also achieve visual emphasis (Kalman & Gergle, 2010).

The analysis of punctuation omission and repeats of punctuation marks indicates that the users of text-based CMC usually hold a rather lax attitude towards the use of punctuation and orthographical correctness. The omission of punctuation in online discourse may result from the economical nature of CMC; the repeating of

punctuation can endow emphasis and intensify the message. With the multiple punctuation marks, the text-based CMC can evoke a rather dramatic effect and thus becomes more semantically marked. The same message would not be as effective without the repeated punctuation (Frehner, 2008; Thurlow, 2002).

5.3.5 Use of cues over time

Figure 5.1 and 5.2 below illustrate the use of different CMC features by Taiwanese and British participantss across the four phases of the intercultural exchange programme. From the figure it is apparent that in general increasingly large numbers of the cues are employed over time, except the use of emoticons, which occur frequently all through the exchange programme. In the comparison of the total amount of use of each cue or feature in the first two and last two phases, two domains reach a significant difference (p <.01): nonstandard capitalisation and nonconventional spelling, with log-likelihood ratios of 7.33 and 13.59 respectively. This shows that the total numbers of instances of these two cues in the last two phases are significantly greater than in the first two phases, which indicates that Taiwanese participants used increasingly large numbers of these two cues over time during the exchange project. This may be due to the fact that most of the Taiwanese learners rarely employed such cues or strategies in online communication, but during the exchange programme communicating with the British participants they noticed how these features were employed by their international peers and gradually started using them. As was noted by Crystal (2006, 2011), although the members of an online community come from different backgrounds and write in different styles, they tend to accommodate each other, and as such their contributions progressively develop a shared linguistic character. On the other hand, however, as shown in Figure 5.2, comparison of the use of

these CMC cues by British participants across the four phases of the exchange did not show significant changes between the first two and last two phases. Although the change is not significant, we can see from the figure that the first phase, which is critical for relationships building, include slightly more instances of cues than other phases.

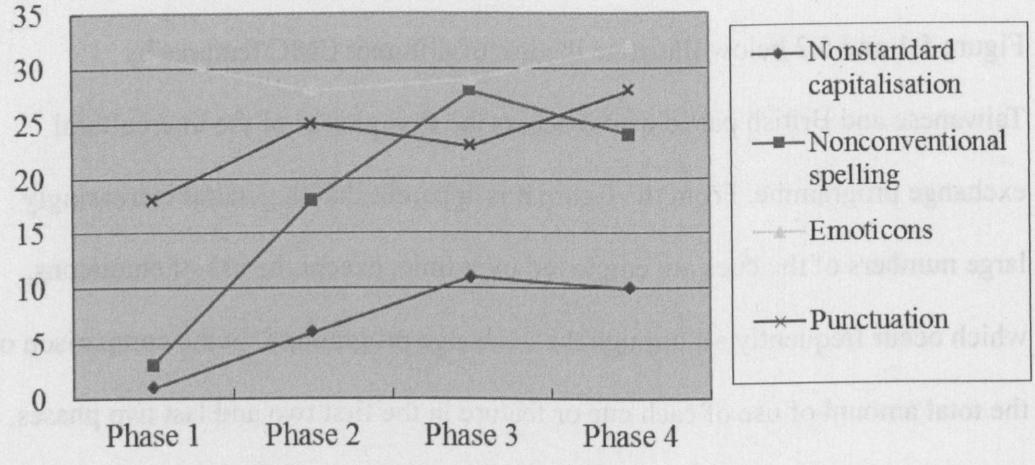


Figure 5.1

The Use of Different Features Over Time by Taiwanese Participants

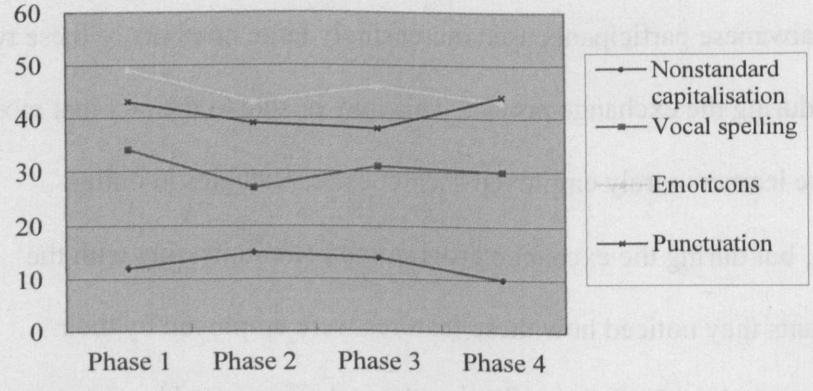


Figure 5.2

The Use of Different Features over Time by British Participants

While the Taiwanese learners increasingly employed more of these CMC features, which are helpful for facilitating online communication, these distinctive linguistic characteristics deviate significantly from the normative usage taught in formal education. They are often considered as incorrect language use by some parents, educators and media and, in their opinions, frequent use of these features may further negatively affect the development of spelling and literacy skills of youngsters (e.g., "Oxford Learning," 2006). Some studies also indicate that young people are increasingly using the CMC features in their offline writing (e.g., Pew, 2009). Therefore, in general these kinds of language usage are often not encouraged in formal writing or other forms of informal written communication. This notwithstanding, previous research (e.g., Plester, Wood, & Joshi, 2009; Varnhagen et al., 2010) has provided evidence that there is no clear relationship between the use of CMC features and conventional written language, indicating that this type of written communication does not have a harmful effect on literacy skills; rather, knowledge of CMC language was statistically associated with vocabulary, word reading and phonological awareness measures. Some studies (e.g., Department for Education, 2012; Plester et al., 2009; Sternberg, Kaplan & Borck, 2007) reported that this language use benefits students in terms of encouraging creativity in written expression, increasing literacy and word reading ability and improving speaking fluency. Clark and Dugdale's (2009) study also shows that blog owners and youngsters with a social networking presence are reported to be significantly better writers compared to those who don't use blogs or social networking sites. In addition, a number of researchers (e.g, Crystal, 2011; Lewis & Fabos, 2005; Varnhagen et al., 2010) suggest that this phenomenon simply represents contemporary language use, a process in the evolution of the English language. Rúa (2007) suggests that "an incursion into a subcode with

which youngsters feel so identified could be extremely beneficial, not only for L1 but also for L2 learning" (p. 178).

This also shows that the CMC features exhibit important pragmatic and interpersonal functions, as shown in 5.3.1-5.3.4, and excluding them from online texts may result in potential ambiguity in their messages and the participants may risk being perceived as rather domineering or pedantic. As Averianova (2012) reports, the potential hazards of inappropriate use of CMC may lead to significant communication problems, such as exclusion, flaming and general lack of comprehensibility (p. 14). However, pedagogical research and practice have not sufficiently addressed this particular type of communication and its importance in intercultural interaction. Averianova (2012) notes that the ability to communicate in different electronically-mediated formats comprises "a new type of literacy required of foreign language learners in the new millennium" (p. 17). Teachers accordingly can make their students aware of these types of language usage in the situations where Internet informality would be more appropriate and those where it would not. As Rúa (2007) notes, learning the CMC features may result in a better understanding of "the notion of linguistic appropriateness and at the same time gain an insight into the functioning of languages and their flexibility to adapt themselves to different communicative situations" (p. 165).

5.3.6 BATTICC-O vs. CANELC

The distinct features found in BATTICC-O can also commonly be seen in a large e-language corpus (CANELC). Table 5.7 presents total instances of different types of features found in BATTICC-O and CANELC. We can see that in general BATTICC-O presents more instances of these distinctive CMC features or cues,

and most of the categories reach a significant difference. These differences can be explained in part by the nature of the two corpora. In BATTICC-O, for example, participants communicate in order to build and maintain a good relationship with other international peers; many of these interpersonal and informal CMC features therefore would be very useful to achieve such goals. CANELC, on the other hand, includes a wide range of online discourse, including blogs, discussion boards, Tweets and e-mails, some of which do not display a great deal of informality. Although many more instances of CMC cues were found in the BATTICC-O than CANELC, the use of abbreviation/acronym/substitution in nonstandard spelling did not occur at a high frequency in BATTICC-O. This is probably because in the intercultural exchange project the participants may realise that their international peers may not understand some of the abbreviations, acronyms or other substitutions and therefore they tended to use fewer nonstandard spellings so that the participants in the group would understand each other more easily.

Table 5.7

Total Instances of Different Types of Features in BATTICC-O and CANELC

Type of features	BATTICC-O		CANELC		Sig.
	Number	per 1000 words	Number	per 1000 words	(LL)
Use of the upper and lower cases					
Capitalised word for stress	89	2.74	1,504	3.01	-2.77
Lower case instead of upper case	967	29.80	352	0.70	3924.9
Nonconventional spelling					
Abbreviation/Acronym/Substitution	51	1.57	494	0.10	8.80
Vocal spelling	187	5.76	703	1.41	219.69
Emoticons	334	10.29	2,436	4.87	136.17
Punctuation					
Repeating punctuation	252	7.77	633	1.27	432.35
Apostrophe omission	124	3.82	96	0.19	404.53

*** $p < .001$

5.3.7 Brief summary

We have seen a wide variety of types of distinctive language characteristics in adolescent online communication (BATTICC-O) and how these linguistic features are employed by Taiwanese and British participants in an intercultural setting. This language usage contains a wealth of cues in CMC in the form of capitalisation/minusculisation, nonconventional spelling, emoticons and punctuation omission and repetition, with a great deal of variation within each category. The examination of these remarkable features demonstrates different preferences by the participants for different purposes. For example, nonstandard capitalisation commonly occurs in singular nouns (e.g., *SUMMER*), followed by lexical verbs (e.g., *LIKE*), adjectives (e.g., *GREAT*) and adverbs (e.g., *REALLY*), and the words semantically related to happiness (e.g., *HAPPY*), like (e.g., *LOVE*) and summoning (e.g., *HEY*); vocal spelling is frequently employed in the use of degree adverbs (e.g., *sooooo, toooo*) and interjections (e.g., *helloooo, yeahhh*), and the related semantic domains of describing degree (e.g., *soooo, veeery*), discourse markers (e.g., *heeey, ooooh*) and like (e.g., *looovveee*). This suggests that these features are not simply employed indiscriminately.

This phenomenon of language usage that contains a wealth of cues in CMC displays a high level of informal interaction. Although many of them do not contribute any specific content or propositions, they have important interpersonal functions and particularly appeal to young people. This may be due to the fact that these particular CMC features help the users to show a high degree of intimacy, informality and in-group membership, all of which serve to develop and maintain good relationships. The employment of these features is therefore given the opportunity to flourish among this population (Grinter, Palen, & Eldridge, 2006).

It can also be noted that in many cases one message contains several of the distinctive features at once. This concurrent usage of cues in text-based CMC is perhaps understandable as using different nonverbal cues simultaneously is extremely common in face-to-face meetings. Therefore, the usage of multiple cues in text-based CMC can be very helpful to compensate for the lack of nonverbal cues in a text-based setting. In addition, although it is accepted that the nonverbal cues in face-to-face interaction such as eye contact, gaze, vocal intonation and gestures are absent in text-based CMC, it is evident that the distinctive characteristics examined in this section have been shown to serve similar functions in text-based CMC with a view to providing information, regulating interaction, stressing the message content and expressing intimacy and emotion. This adds to a growing body of recent literature on the use of cues in CMC as demonstrated in a range of studies (Cho, 2010; Dresner & Herring, 2012; Kalman & Gergle, 2009, 2010; Riordan & Kreuz, 2010; Varnhagen et al., 2010). In particular, expressing intimacy and emotion is most prevalent in BATTICC-O, in which various types of emoticons and punctuation are highly frequently used in the adolescent intercultural exchange, and many of the features such as nonstandard capitalisation and vocal spelling include a good variety of cue-laden words of affect, expressing a different level of emotion. Harris and Paradice (2007) reported that the recipients of the message perceived a higher degree of the sender's emotions when increasing the number of cues. This shows that the cues in text-based CMC do influence message interpretation and the importance of them should not be neglected.

5.4 Analysing spoken discourse

The previous section has presented the most distinctive lexical and grammatical features commonly occurring in online discourse (BATTICC-O). In this section I examine the most important linguistic features of spoken discourse (BATTICC-F), paying particular attention to the analysis of vague expressions, approximations and hedging (5.4.1), situational ellipsis (5.4.2), headers and tails (5.4.3), pausing and repeating (5.4.4) and discourse marking (5.4.5).

The analytical framework for analysing spoken discourse has been presented in 3.3.4. I first looked at vague expressions, approximations and hedging. These items are deliberately used by speakers to refer to people and things in an imprecise way (Carter & McCarthy, 2006; O'Keeffe et al., 2007), as in *around six*, *a couple of days ago*, *and so on* and *and things like that*. Syntactic features include situational ellipsis, headers and tails. Situational ellipsis involves the deliberate omission of items such as subject pronouns and verb complements. Headers (e.g., *the teacher*, he is very nice) and tails (e.g., They're really nice, *my teachers*) are the features that alter the word order of traditional written grammar.

Concerning discourse features, I first examine pausing, repeating and recasting. Pausing is identified as (...) or (..) in the transcripts, or it can be filled by a vocalisation such as *er* and *erm*. Repeating can be one word (e.g., *I'm I'm not sure*) or phrases/sentences (e.g., *We're meant to be talking ..er.. we're meant to be talking about the walk.*). Recasting is identified as instances of reformulating words, phrases, clauses or sentences. (e.g., Before we start ... *before we go into that level of detail*, I'm going to write it on the OHP). Finally, I looked at the use

of discourse markers (DMs) containing four main functional domains: interpersonal, referential, structural and cognitive categories. In this analysis, the use of each DM in its discourse contexts is examined to identify the primary function in my datasets. However, it should also be noted that each DM may perform more than one of these functions, as can be seen in the examples above. The following sections will report on the analysis of spoken discourse in BATTICC based on the discourse analytical framework.

5.5 The linguistic features in BATTICC-F

Table 5.8 presents the total frequencies and percentages of different features of spoken discourse found in BATTICC-F, showing the different amount of use by Taiwanese and British participants.

Table 5.8
Spoken Grammar in BATTICC-F

Spoken features	Taiwanese		British		Sig.
	Number	per 1000 words	Number	per 1000 words	(LL)
Vague categories	21	3.91	83	7.36	**8.75
Approximation	16	2.84	52	4.61	3.09
Hedging	2	0.36	31	2.75	***14.40
Situational ellipsis	36	6.40	53	4.70	2.00
Headers and tails	5	0.89	9	0.80	0.04
Pauses	632	112.38	915	81.15	***38.67
Repeating and recasting	37	6.58	58	5.14	1.34
Discourse marking	432	76.81	968	85.85	3.75

*p <.05. **p <.01 ***p <.001

From the table it can be seen that British speakers generally produce more instances of vague categories, approximation, hedging and discourse marking, especially hedging and vague categories, which reach a highly significant level.

On the other hand, situational ellipsis, headers and tails, pauses, repeating and recasting are more commonly used by Taiwanese students, and the use of pausing between the two groups achieves a significant difference. These features will be discussed in more detail in the following subsections.

5.5.1 Vague expressions, approximations and hedging

Vague language is frequently used by speakers to convey information that is softened in some way so that utterances are slightly more indirect and less assertive (Carter & McCarthy, 2006, p. 202). In this section the use of vague language falls into three aspects: vague categories, hedges and approximations.

Table 5.9
Number of Different Vague Expressions

Vague expressions	Taiwanese		British		Sig.
	Number	per 1000 words	Numbe	per 1000 words	(LL)
Vague categories					
(and) stuff (like that)	2	0.36	15	1.33	$p<.05$
(that/this) sort of (thing/like)	0	0.00	12	1.06	$p<.01$
(or) anything (like that)	2	0.36	10	0.89	
(or/and) something (like that)	7	1.24	12	1.06	
(that/this) kind of (thing)	3	0.53	3	0.27	
(and) everything	3	0.53	7	0.62	
(and/but) thing(s) (like that)	5	0.71	24	2.13	
Total	21	3.91	83	7.36	$p<.01$
Hedges					
sort of	0	0.00	17	1.51	$p<.00$
a bit / a little bit	2	0.36	14	1.24	
Total	2	0.36	31	2.75	$p<.00$
Approximation					
about / around	6	1.07	6	0.53	
lots of / a lot	10	1.78	34	3.02	
loads of	0	0.00	8	0.71	$p<.01$
a couple of	0	0.00	4	0.35	
Total	16	2.84	52	4.61	

Table 5.9 presents the total number of instances for each type of vague expression, which will be discussed in the following three subsections. The table also shows statistically significant differences in the amount of use between the two groups of participants revealed by using log-likelihood ratios (Rayson, 2008) to compare the cumulative frequencies of each item.

a) Vague categories

Vague categories refer to vague use of categories of items. The speakers in this case do not necessarily convey precise and concrete information, and the hearers in most cases know what their vague expressions refer to. One highly important function of these is to indicate assumed or shared knowledge and to mark in-group membership (Carter & McCarthy, 2006; O'Keeffe et al., 2007). In BATTICC-F 104 instances of expressions indicating vague categories were found. As can be seen in Table 5.9 they typically include words and phrases such as *thing, stuff, like, or something, or anything, kind of* and *sort of*, which are found in 21 and 83 instances in the Taiwanese and British datasets respectively. Tests of log-likelihood revealed a significant difference in the use of *(and) stuff (like that)*, *(that/this) sort of (thing/like)* and the cumulative frequencies of vague categories between the two sets of data.

Sort of is one of the most commonly used vague expressions in the British data, while no instances were found in Taiwanese learners' discourse. The following extract presents how it is used in context. In (78) BT13 and BT14 are talking about gift ideas for their fathers. BT14 used the vague expression *that sort of thing* twice, and BT13 may well know what he/she means although no explicit reference is given.

(78) <BT14>: Right, because all of them .. all the presents I've made .. you know what I mean, like I made all the key rings they're more for Mum then you know ... my Dad doesn't like ***that sort of thing***.
<BT13>: Yeah, I bought a load of rope bracelets for my Dad.
<BT14>: My Dad's not into ***that sort of thing***. I was going to get him like a model ***or something*** ... If I do, I'll get him some alcohol from duty free ...

(BATTICC-F)

The first use of *that sort of thing* may well refer to the presents that the speakers have made during the cultural exchange programme, and this reference appears to be a marker of shared knowledge and experience that they can draw on. The second use of *that sort of thing* refers to the gift that BT13 bought for his/her father so the speaker BT14 does not necessarily need to repeat the noun phrase *a load of rope bracelets,* and this in turn further asks the hearer to construct the relevant ideas of buying a gift. Moreover, the use of *or something* basically indicates an alternative category of gifts, and such usage simply "keeps options open" (Carter & McCarthy, 2006, p. 202). Such use of vague language describing categories of items is sometimes referred to as a "vague category identifier" (Channell, 1994), which is made up of an exemplar (i.e., *a model*) plus a vague tag (i.e., *or something*), where the exemplar directs listeners to identify the category referred to.

This use of vague language is sometimes given different labels, such as "general extenders" (Overstreet & Yule, 2002), "extension particles" (Dubois, 1992), "vagueness tags" (De Cock, 2004), "vague category identifier" (Channell, 1994) and "vague category markers" (O'Keeffe et al., 2007; Evison, McCarthy, & O'Keeffe, 2007). Some more instances of vague expressions retrieved from the BATTICC-F are presented in the following excerpts:

(79) <BT07>: I know that they're a lot more like ... the girls have to have their hair out of their faces and they can't wear make-up ***and stuff*** and ...

(80) <BT07>: But then we might not have chance to come again because of like money ***and stuff***.

(81) <BT23>: So .. er ... what sort of different things have you been noticing in our culture and traditions ***and stuff***?

(82) <BT18>: Yeah, so it's like a lot fresher and generally .... do you find that we have fresher ... erm ... fresh vegetables ***or anything like that***?

(83) <BT17>: ...but ***things like*** you know, your hair or your shoes ***or anything*** – they're not really bothered about it.

(84) <BT08>: Erm .. .. it wasn't like amazing ***or anything***. I'm just like weird anyway so people laugh at it.

(BATTICC-F)

The *and stuff* shown in (79) may be inferred to mean the kinds of girls' make-up, dress and accessories; the use of the same phrase in (80) was used for stating the reasons that might discourage BT07 from having another chance to visit Taiwan, which might include funding, school rules and other complicated restrictions that are not easy to explain in detail. Such use of vague expressions serves interpersonal functions of conversation and is probably preferred by interlocutors as it may have distanced the speakers from the interlocutors if they had used more formal terms, such as cosmetics for (79) or economic hardship for (80). Vague category identifiers sometimes occur with an interrogative manner, i.e., being used as a tag question, as in (81) and (82), which likely leave room for the interlocutors to add their own description of the situation (Adolphs, Atkins, & Harvey, 2007). In these cases, BT23 and BT18 seem to not only ask the interlocutors to describe the differences between British and Taiwanese cultures or the different fruits or vegetables they have found, but also direct them to consider an entire category of cultural differences to share with the speakers. In addition, the vague expressions with disjunctive coordinator *or* typically indicate an alternative, such as *or anything* in (82), (83) and (84).

In some cases vague categories can be used to exemplify the explanations (Koester, 2007), as in (79) and (83). For example, the use of *your hair or your shoes or anything* in (83) refers to the *things* in the utterance, where the speaker BT17 explains the different items related to school dress code. BT17 even puts stress on the *or anything* in the utterance, with an explicit meaning that there is no specific requirement concerning hairstyle, make-up and attire in British schools, while the Taiwanese schools normally set a strict dress code. Another situation of using vague language happens when BT08 won the award for the talent show. As in (84), people praise and admire BT08 for his/her excellent performance, and the use of *or anything* seems to soften his/her response and function as a disclaimer used to forestall negative evaluation by others. As Overstreet & Yule (2001) claim, such use of *or anything* may "support the speaker's attempt to make sure that the co-participant does not adopt the possible negative interpretation of behavior being disavowed" (p. 51). This use seems to downtone or hedge the utterance, which I will examine further in the following subsection.

b) Hedges

Another important function that vague expressions serve in my data is to hedge the commitment of the speaker to what he or she asserts. As shown in Table 5.9, *sort of* is the most prevalent example of such use (17 instances), although most of the instances of *sort of* function to indicate vague categories. As can be seen in the following extract involving the British speaker BT18 and the Taiwanese TW16 talking about the differences between Taiwanese and British food, *sort of* is used three times in one utterance.

(85) <BT18>: Okay. Yeah, your food generally is a lot more ***sort of*** ... erm ... traditional and special than ours. Ours is just ***sort of*** simple, ***sort of***, ...

<TW16>: It's okay, I like it ... it's your culture actually.
<BT18>: Yeah.
<TW16>: But I am .. I don't like the traditional breakfast because it's too salty and the flavour is too strong.

(BATTICC-F)

The speaker BT18 is likely trying to hedge the assertion by frequently using *sort of* when giving comments on Taiwanese food so that the statement sounds less direct. This is perhaps explained by the uncertainty of the speaker BT18 about his/her own assumption, and he/she thus intends to be less assertive; on the other hand, the speaker TW16's response *I don't like the traditional breakfast because it's too salty and the flavour is too strong* seems much more direct compared to BT18's statement. Miskovic-Lukovic (2009) calls such use of vague expressions "positive politeness strategies" (p. 622). These help to "*downtone* the force of the utterance" and to "*mitigate* against any potential threat to face" (O'Keeffe et al., 2007, p. 174). Moreover, the pervasive use of *sort of* in BT18's utterance seems to indicate a certain level of hesitance in the planning of speech and searching for appropriate words in that the expression *sort of* functions as a filler or a time-buying device in the discourse, which might further develop speaking fluency in general. However, as can be seen from the extract, the Taiwanese speaker TW16 uses very few fillers in his utterances.

Interestingly, vague quantifiers such as *a bit* and *a little bit* found in my data pragmatically function as downtoners, which is exemplified in the following cases. Such use of vague quantifiers as hedges can be found in 16 examples in BATTICC-F.

(86) <TW05>: I think the question is boring.
<BT06>: They were ***a bit confusing***.

(87) <BT17>: Because we go back to school the day I do my birthday. It is ***a***

bit annoying but it is okay.

(88) <BT13>: Uh...thanks. How would you write all of that? ... I'm not sure. Just ***a little bit difficult***. Maybe just that word.

(89) <BT07>: Don't mind these two they're ***a bit weird***.

(BATTICC-F)

In these cases *a bit* or *a little bit* is commonly prefaced to different adjectives, such as *confusing, annoying, weird, cheeky, strange* and *uncomfortable,* most of which seem to be used with negative situations. Such use of vague quantifiers seems to downtone and hedge the utterances, which is highly likely to be more appropriate in conversation, and as such this is considered as possessing more "pragmatic adequacy and integrity" in informal contexts (O'Keeffe et al., 2007, p. 71). Adolphs et al. (2007) describe this as "a modification which serves to reduce the negative assessment" by the speakers (p. 72). For example, in (86), BT06 identifies with the negative assertion previously provided by TW05 and reformulates it, which presumably helps to minimise the negative emotion of the interlocutor and constructs and maintains a relaxing tone of conversation.

c) Approximations

Similar to vague expressions, approximations, particularly used with numbers, quantities or some other measurable units, are frequently introduced by speakers in informal situations to downtone what might otherwise sound overly precise (O'Keeffe et al., 2007). This is sometimes described as "vague additives" (Channell, 1994) or "vague approximators" (Koester, 2007). In BATTICC-F a wide range of expressions can be found. The most prevalent item of this type is *lots of/a lot*, which can be found serving this function in 44 instances. Other uses of approximations include *about/around* (12 instances), *loads of* (8 instances) and *a couple of* (3 instances). In the following extracts derived from BATTICC-F, we

can see how the approximation is used in conversation:

(90) <TW07>: Really?
<BT07>: Not all the time – for ***a couple of*** days – and then there's ***a couple of*** months and it's quite warm.

(91) <TW01>: So how is the weather in the UK?
<BT01>: Rubbish.
<BT03>: Raining.
<BT01>: If you come over, just pack ***loads of*** jumpers.

(92) <BT15>: Harry Potter?
<TW11>: Yeah.
<BT15>: Yeah. I ... erm ... I know someone who's read the entire series ***about*** fourteen times.

(BATTICC-F)

The extracts present approximations found in spoken discourse involving the use of vague quantifiers such as *about, a couple of* and *loads of.* These items seem to indicate the absence of precision, and they have the same interpersonal functions as the vague expressions in that young learners tend to engage in a more conversational style and avoid being absolutely precise and perhaps being considered pedantic (Carter & McCarthy, 2006). As in (90), *a couple of* occurs twice to refer vaguely to the amounts of days and months as the exact amount of time might not be relevant. As Anderson (2000) claims, exactitude would not benefit the hearer as it requires additional and unnecessary cognitive effects

Although three categories of vague expressions were discussed individually, they usually co-occur at the same time. Some more examples of vague language are derived from BATTICC-F, as shown in the following excerpt (93). The conversation takes place between two British speakers BT07 and BT09 and one Taiwanese learner TW07, mainly talking about trains in the UK. The use of both vague expressions and approximations are pervasive in the discourse, where the speakers regularly insert hedges and monitors of shared knowledge.

(93) <TW07>: Yeah, so how did you feel er .. to take this train?
<BT07>: Yeah, the trains ***like*** are more on time.
<BT09>: Yeah, it's ***a lot*** cleaner on the trains here. Sorry if that's scared you about English trains.
<BT07>: ***I think*** they're ***like*** ... in England they're ***sort of like*** ... ***well*** I've missed it I'll ***just*** wait half an hour and get the next one.
<BT09>: It's every 15 minutes by the way.
<BT07>: Yeah, they don't .... over here they take ***lots of like*** care in the presentation ***like*** being clean and people have ***like a lot of*** respect for them.
<BT08>: Yeah, everyone has ***loads of*** respect for different ***like*** ... ***you know***, trains ***and stuff*** but in the UK, it's ***sort of like*** no one has as much respect for ***things*** as you do over here, ***like*** for trains or buses ***or anything*** because they ***just*** ... I don't know why ... they ***just*** don't have as much respect.

(BATTICC-F)

This example presents nearly 20 instances of vague words or phrases in that the multi-word expressions in italics seem to mark a purposive vagueness to hedge the assertions by allowing interlocutors to downtone what they say. Some of them also indicate assumed or shared knowledge and mark in-group membership (Carter & McCarthy, 2006; O'Keeffe et al., 2007). Although these items are vague in nature, they are interpreted successfully by the hearer because the referents of the expressions can be assumed to be known by the interlocutors. As Carter and McCarthy (2006) state, such use of language presents that young learners tend to engage in a more conversational style and avoid being absolutely precise and perhaps being heard as pedantic.

5.5.2 Situational ellipsis

As discussed in Chapter 2, ellipsis involves the deliberate omission of items such as subject pronouns and verb complements that may not be necessary in the

utterances since they contain enough information for the purpose of the conversation (Carter & McCarthy, 2006; Thornbury & Slade, 2006). The BATTICC-F represents 36 and 53 instances of situational ellipsis in the Taiwanese and British datasets respectively. Table 5.10 shows the different elements of ellipsis in the participants' discourse. From the table it can be noted that the Taiwanese participants generally used more situational ellipsis in face-to-face communication than the British learners, although the difference between the two groups does not reach a significant level (LL=2.00, $p>.05$).

Table 5.10
Elements of Ellipsis in BATTICC-F

Element of ellipsis	Taiwanese		British	
	Number	per 1000 words	Number	per 1000 words
Initial *I* (+*be*)	8	1.422	8	0.710
Personal pron. *they, he/she, you, we*	5	0.889	8	0.710
Interrogatives (+*be*)	3	0.533	4	0.355
It and demonstratives (+*be*)	9	1.600	18	1.596
Copular verb *be*	0	0	3	0.266
Existential *there*	4	0.711	2	0.177
Auxiliary verb *do, does, did*	2	0.356	3	0.266
Preposition	1	0.178	3	0.266
Final ellipsis	4	0.711	4	0.355
Total	36	6.401	53	4.701

Some of the examples of ellipsis are shown in the following excerpts. The bold italics in the sentences in parentheses illustrate a hypothetical understanding of what speakers have chosen to 'omit' or have left unsaid.

(94) <BT01>: (***I'm***) Doing Facebook now.
(95) <BT18>: Erm (***I'm***) Not sure really.
(96) <TW01>: And how is the view?
<BT02>: (***It's***) Beautiful.
(97) <BT23>: Do you like French then?

<TW22>: Er erm (***it's***) pretty hard … pretty hard but ...
<BT23>: (***Is it***) Harder than English?

(98) <TW18>: (***What's***) Your school's name? (***I'm***) Sorry I forgot.

(BATTICC-F)

It can be noted from the excerpts that the majority of the situational ellipses involve the initial elements of the clauses or sentences. This is probably because the information at the beginning of the utterance is usually incorporated – "information that is more readily recoverable from the context" (Thornbury & Slade, 2006, p.84). From the excerpts the omission includes initial *I* plus copular verb *be* (e.g., *I'm*) in declaratives, as in (94) and (95), subject pronoun *it* or other demonstrative pronouns plus *be* (e.g., *It's*), as in (96) and (97), and interrogatives (e.g., *What's*), as in (98). Although the subject, verbs or other grammatically essential elements are omitted, as Carter and McCarthy (2006) state, "in reality nothing is 'missing' from elliptical messages" (p. 181). In (96), for example, it is easy for the hearer to understand that the elliptical element refers to *the view* in the previous utterance. Also, an interrogative without auxiliary, verbs or subject can be found in my data, such as *Harder than English?* in (97), which can be clearly understood as *Is French harder than English?*

Although the situational ellipses mainly involve the omission of *It* and demonstratives (27 instances) and first person pronoun *I* (16 instances), some other subject pronouns, such as *we, he* and *they*, are often unnecessary in a situational context, as in (99) and (100). In addition, existential *there* (and its accompanying verb *be*), copular verb *be* (i.e., *are*) and prepositions are occasionally omitted in speech, as in (101) to (104).

(99) <BT07>: we need to get paid back for those coats ...
<BT06>: Yeah – (***we***) need a refund [laughter]

(100) <TW07>: Maybe people in your country is very busy, so ... so (***they***) should take a train on time

(101) <BT02>: Yes, mountains (***are***) higher.

(102) <BT02>: ... Hi mate, (***Are***) you all right, yeah?

(103) <TW07>:So girls ... girls always put some make-up to go to school?
<BT09>: But (***there are***) not lots.

(104) <BT11>: The nearest one's (***in***) Liverpool.

(105) <TW16>:Yeah, I have not tried... I don't try fish and chips yet.
<BT18>: Okay, you should (***try fish and chips***).

(BATTICC-F)

Final ellipsis can also be found in the discourse, as shown in (105), where the ellipsis avoids a repetition of the phrase *try fish and chips* by the interlocutor. These elements may not be necessary in these cases as the reference is obvious. From all of the excerpts shown above it appears that situational ellipsis may involve understood references to a range of expressions relating to people and different items and it is pervasive in BATTICC-F. However, although elliptical forms are commonly used by both native speakers and learners in informal interrogatives, they are sometimes considered as 'grammatically incorrect' forms for EFL classroom teaching (Mumford, 2009). Mumford further suggests that teachers should demonstrate to learners these significant features of spoken grammar, thereby training them for efficiency in speaking.

5.5.3 Headers and tails

Headers refer to fronting "adjuncts, objects and complements, and noun phrases before the pronoun" (Cutting, 2011, p. 160), which is often used to emphasise what the speaker thinks is particularly important (Carter et al., 2011). In BATTICC-F, 11 entries of headers are found, with 6 and 5 instances in British and Taiwanese participants' datasets respectively. Some of the examples are presented as follows:

(106) <BT21>: Yeah, we ... I don't know ... ***our lakes*** ... I don't think ***they***'re very clean but maybe they look clean.

(107) <TW01>:Because ***Eric, he*** want us to prepare present for tomorrow's social activity.

(108) <BT08>: ***Me and my friend*** ... ***we*** honestly tried to do a proper high five and we went like this

(109) <TW07>:So today's ... er.. ***today's activity*** er ***which part*** do you like best?

(110) <TW07>:Yeah, and ***the man*** ... maybe ***he***'s thinking...

(111) <BT16>: ...because the road is very slippery. And last year, in ***a car, it*** spun around and I was screaming [laughter]

(112) <BT19>: Yeah. Erm ... And what is your school like?
<TW17>: Pardon?
<BT19>: ***Your school***, what's ***it*** like?

(BATTICC-F)

From the examples above it can be seen that headers generally include a noun or a noun phrase, which performs the orienting and focusing function, followed by a pronoun that refers back to the noun or noun phrases previously mentioned. For example, in (106) BT21 seems to put *our lakes* at the front to provide orientation for the listener, and the pronoun *they* is then used to refer back to *our lakes.* In this way, by using headers, his expression can be easily understood. Such use can also be found in Taiwanese participants' discourse, as shown in (107), (109) and (110). In (109), the use of headers by TW07 is used in an interrogative sentence, which is slightly different from other examples. The fronted element *today's activity* prior to the interrogative *which part* in the utterance seems to lead listeners to the focus of the question. It can be seen even more clearly in (112) in that the first utterance *what is your school like?* by BT19 is rephrased to *Your school, what's it like?*, which may be more easily understood by the hearer. From the examples it appears that headers provide orientations and emphasis for listeners, "serving to include information which speakers consider relevant to their listeners and

attempting to do so economically" (Carter et al., 2011, p. 94).

Tails, on the other hand, normally occur after clauses, which allow speakers to "express attitudes, to add emphasis, to evaluate and to provide repetition for listeners" (Carter et al., 2011, p. 81). In BATTICC-F there are only 3 instances found in the British participants' discourse. In the following examples (113) and (114), the noun phrases *the weather here* and *climbing up hills* clarify, repeat and/or emphasise the referent of the pronoun *it* in the sentence that comes before them.

(113) <BT23>: Okay. .. is it ... is ***it*** a shock difference.. ***the weather here***?
(114) <BT08>: But ***it*** was quite fun, ***climbing up hills***.

(BATTICC-F)

In these excerpts it can be seen that tails serve "an essentially recapitulatory function" (ibid., p. 82). In this case, speakers are sensitive to listeners' reactions and employ clarifying noun phrases following the utterances to ensure cohesion and to facilitate the flow of communication. As such, Carter et al. (2011) describe tails as one of the elements of "an *interpersonal* grammar" in that the speaker attempts to involve the listener by using expressions that reflect personal attitude, feelings and listener-sensitiveness (p. 82).

5.5.4 Pausing, repeating and recasting

Other characteristics that typically feature in naturally occurring speech are pausing, repeating and recasting.. Cutting (2011) categorises these characteristics under the general heading *disfluency features*, and Corley and Stewart (2008) include them as *hesitation disfluencies*. Nevertheless, Tottie (2011) argues that the term *disfluency* is based on an ideal world of fluent speech production and is "a

rather negative and uninformative default term that says nothing about the discourse functions" (p. 193). She further proposes the term *planners,* in a more positive vein. As Kjellmer (2003) points out, they have important discourse functions, helping "to organise the utterance for the listener, who will more easily realize its structure and its main point and be able to follow the argument" (p. 190).

In Carter and McCarthy's (2006) framework, pausing can be categorised into unfilled or filled pauses. Unfilled ones are simply a silence labeled as a sequence of dots, such as (..) for a brief break or (...) for a longer pause in the transcripts of BATTICC-F. Filled pauses are identified by a vocalisation such as *er* and *erm.* They are typically used to fill pauses between the elements of utterances and mark a hesitation or uncertainty on the part of the speaker (Carter & McCarthy, 2006). Research has shown that the two items are extremely pervasive in naturally occurring speech. As shown in Carter and McCarthy's (2006) analysis, *er* represents the seventeenth most common word in the Cambridge International Corpus. Moreover, according to Fox Tree's (1995) study, approximately 6% of words uttered are, or are affected by, some form of hesitation devices. These items are also common in BATTICC-F, as can be seen in the following examples.

(115) <BT07>: You know the dragon boat racing?
<TW07>: Yeah.
<BT07>: What What's it about?
<TW07>: You mean story, or ...
<BT07>: Yeah, yeah the story of it.
<TW07>: Oh the story ... okay I think ... A long, long time ago ...***erm*** ... was ... ***erm*** ... how do you ***er*** say .. you say King ... in China they say ... they say like King.

(116) <TW08>:Why don't you eat hot food?

<BT07>: ***Erm*** ... we do but there's like ... ***erm***... not many people like it as much over ... as like over here, so in the UK people like ...

(117) <TW10>:Do you have any festival in England?

<BT10>: ***Er erm*** Mayday ... the first day of May ... ***erm*** ... New Year's Day.

(BATTICC-F)

In (115) BT07 asks TW07 about the origin of dragon boat racing, while TW07's discourse, which includes three filled pauses and nine unfilled pauses, seems to indicate that he/she has limited knowledge about it. These pauses may further mark the hesitation and uncertainty of the speaker. In such cases the speaker might not have much to say about a difficult topic, therefore many instances of pauses are found. A similar example can be found in the British participants' data, as shown in (116). The speaker BT07 might not have many opinions about TW08's question, and consequently a number of pauses are used in the response. In (117) TW10 asks BT10 about the festivals in England. Three filled pauses (i.e., *er, erm*) and a number of unfilled ones (i.e., ...) can be found in BT10's utterance. They fill the gaps between the elements of the utterance and allow the speaker more thinking time. In this case, *er* or *erm* is used as a time-buyer to select a more appropriate lexical choice.

From the excerpts it can also be noted that *er* frequently collocates with *erm* and/or other unfilled pauses (i.e., ..or...). They all co-occur significantly to indicate the speaker's lack of certainty. This is supported by Kjellmer's (2003) study, showing that one of the top collocates of *er* is *erm*, and, correspondingly, one of the top collocates of *erm* is *er* (p. 182). These two items are extremely pervasive in the Taiwanese participants' data. As can be seen in Table 5.11, regarding the total number of pauses by the participants, the Taiwanese learners

generally employ significantly more pauses than the British participants (LL=38.28, p <.001), although the use of *erm* does not reach a significant difference (LL=0.84, p >.05). Such a substantial use of pauses in the Taiwanese participants' discourse indicates a hesitation or lack of confidence in speaking English. This may be partly because they have not had many opportunities to use the language and might therefore feel anxious when they speak a foreign language. The questionnaire data also confirms this, showing the high level of anxiety of Taiwanese participants during the conversation. Although the items are widely used, most of the time they do not disrupt the conversation.

While pauses are sometimes considered as a sign of hesitation, they also have other functions in face-to-face communication. One such example, the filled pause *er* or *erm,* can function to signpost speaker turns (Kjellmer, 2003). As can be seen in (118), *er* or *erm* indicates its use for turn taking, turn holding and turn yielding, occurring at the initial, middle and end of the message respectively. In my data the majority of filled pauses are found as utterance or turn initiators. This can be clearly seen in the excerpts (118) to (120).

(118) <TW09>:***Erm*** what do you usually do in the holidays?
<BT10>: ***Erm*** Sleep. [laughter] – sleep, ***erm*** meet up with friends...***er...***
<TW09>:***Er*** do you like sports?

(119) <BT15>: ***Erm*** ... what sort of damage is caused by the typhoons?
<TW11>:***Er*** ... yeah ... you know the tree was blown away, and and the car ..

(120) <TW09>:Fish and chips?
<BT10>: Yeah, it's very nice. It's famous English food.
<BT09>: And full English breakfast.
<BT10>: ***Erm*** that's like bacon, beans, eggs, tomatoes, hash browns, mushrooms and sometimes onion and it's really nice. (BATTICC-F)

In many cases, *er* or *erm* not only serves as a turn initiator, but indicates that the speaker wants to take over the turn. As in (120), BT10 tries to break into the conversation and add extra information to the response. Similarly, some instances of *er* or *erm* serve the function of turn-holding, keeping the floor while formulating the utterance (Kjellmer, 2003). This indicates that the speaker is preparing a new piece of information to be uttered, intends to go on speaking and may not be willing to yield the turn.

(121) <TW13>: In England, there have a afternoon tea?
<BT16>: Afternoon tea, yes, I love afternoon tea. ***Erm*** When we go to afternoon tea, you dress up nicely, and it's usually in a big house, an old house, ***and erm*** you have many sandwiches, and cakes and tea obviously [laughter] it's very nice.

(BATTICC-F)

As can be seen in excerpt (121), the first *erm* occurs at the end of the first sentence and the speaker intends to keep the floor and continue. The second *erm*, following the coordinating conjunction *and,* indicates the speaker's intention to add new information. In contrast, if the speaker does not have something substantial to say, he/she would yield the turn by means of *er* or *erm*. As in (118), the speaker might have no information to add, and the use of *er* signals that he/she is yielding the turn.

In addition to pauses, in naturally occurring spoken discourse repeating and recasting are common under the pressures of real time communication. Repeating can be just one word, as in (122) and (123), or can be whole phrases/sentences, as in (124) to (126). This is very common in both British and Taiwanese learners' discourse. Sometimes it is extremely common when hesitating in speaking. This can be clearly seen in (125), where the speaker repeats *do you like* three times,

collocating with other hesitation devices such as filled pauses (i.e., *erm*) and unfilled pauses (i.e., ...).

(122) <TW11>:In in typhoon, it's ***very*** ... ***very*** bad, you know ... ***it's it's*** wet be=

(123) <TW13>: But it's very difficult to me. Er... ***I I I*** only learn a song.

(124) <TW07>:So you spend much time to prepare this talent show?
<BT07>: No, five minutes [laughter] ***I wasn't*** .. ***I wasn't*** gonna do it ....

(125) <TW11>:Have you ***ever read*** ..er. ***ever read*** Harry Potter?

(126) <BT18>: Erm ... ***do you like*** ... erm ... ***do you like*** general, like ... ***do you like*** fish and chips?

(127) <BT23>: Okay. .. ***is it*** ... ***is it*** a shock difference, the weather here?

(BATTICC-F)

Recasting is identified as instances of reformulating words, phrases, clauses or sentences. As can be seen in the following excerpts (128) to (130), many instances of utterance reformulating may be the result of the real time pressure of face-to-face communication in that the speaker occasionally speaks too fast and would like to modify the utterance.

(128) <BT21>: Yeah ... I used to do it a bit but I'm in the right place basically but ... ***do you ever*** ... ***do you go*** .... ***where about do you go*** if you do go in Taiwan?

(129) <BT18>: ***Do you have*** ... erm ... ***you have carrots don't you***?

(130) <TW17>:Erm Do you like erm P.E. class?
<BT19>: Erm ... not very much [laughter]. ***It's*** ... ***I'm*** not very good at it. [laughter]

(BATTICC-F)

From the analysis of pausing, repeating and recasting, it seems that these strategies allow speakers to buy time for speech planning and keep the floor while formulating the following utterance, and meanwhile listeners are also provided with time to figure out what is going on and what will come next. Comprehension of communication as a result can be further facilitated (Munford, 2009; Tottie,

2011). As such, pauses not only indicate pure hesitation, but also have some particular discourse functions in communication. They guide and lubricate the conversation in that "they operate partly below the level of consciousness and can therefore be an unobtrusive and effective instrument in facilitating spoken interaction" (Kjellmer, 2003, p. 191). In addition, what needs to be stressed is that some instances of pauses found in the Taiwanese participants' discourse are *a ya* (3 instances), *ei* (4 instances), etc. These items seem to be their L1 equivalents, which sound very unnatural embedded in English conversation. As a result, the natural features of spoken language should be introduced and encouraged for EFL learners by demonstrating authentic data extracted from real-time communication.

Table 5.11
Number of Unfilled and Filled Pauses, Repeating and Recasting

Type of pauses	Taiwanese		British		Sig.
	Number	per 1000 words	Number	per 1000 words	(LL)
Short pauses (..)	313	39.50	424	34.83	2.84
Long pauses (...)	352	44.42	403	33.10	16.10
Er	92	11.61	25	2.05	74.93
Erm	70	8.83	93	7.64	0.84
Total pauses	827	104.37	945	77.62	38.28
Repeating	31	3.91	32	2.63	2.47
Recasting	7	0.88	18	1.48	1.43

5.5.5 Discourse marking

Discourse markers (DMs) investigated here fall into four categories: interpersonal, referential, structural and cognitive DMs. Table 5.12 presents the total numbers and relative frequencies per 1000 words of four different types of DMs found in the Taiwanese and British datasets in BATTICC-F.

Table 5.12

Discourse Markers in BATTICC-F

Four types of DMs	Taiwanese		British		Sig.
	Number	per 1000 words	Number	per 1000 words	difference
Interpersonal DMs	**159**	**28.27**	**233**	**20.67**	$p<.01$
yeah	132	23.47	153	13.57	$p<.001$
oh	25	4.45	43	3.81	
sort of	0	0.00	23	2.04	
you know	2	0.36	14	1.24	
Referential DMs	**144**	**25.60**	**378**	**33.53**	$p<.01$
and	66	11.74	171	15.17	
but	32	5.69	78	6.92	
so	31	5.51	77	6.83	
coz/because	15	2.67	52	4.61	$p<.05$
Structural DMs	**95**	**16.89**	**113**	**10.02**	$p<.001$
so	49	8.71	43	3.81	$p<.001$
okay	40	7.11	29	2.57	$p<.001$
then	3	0.53	31	2.75	$p<.001$
right	3	0.53	10	0.89	
Cognitive DMs	**34**	**6.05**	**244**	**21.64**	$p<.001$
well	5	0.89	13	1.15	
like	12	2.13	201	17.83	$p<.001$
I think	12	2.13	25	2.22	
you know	5	0.89	5	0.44	
Total	**432**	**76.81**	**968**	**85.85**	

Within each category only the four most frequent items in BATTICC-F are presented and counted. As can be seen in the table, although the differences in the total numbers of DMs in the two datasets are not statistically significant, the accumulative frequencies of each type of DM reach a significant difference, in which Taiwanese learners use significantly more interpersonal and structural DMs, while referential and cognitive DMs are used significantly more often by the British participants. Furthermore, the numbers of high-frequency DMs in the two datasets differ significantly. For example, the frequencies of interpersonal *yeah*

and structural *so* and *okay* are significantly higher in the Taiwanese discourse, while structural *then* and cognitive *like* are used significantly more frequently by British participants.

a) Interpersonal DMs

The main functions of interpersonal DMs are marking shared knowledge, indicating attitudes and showing responses (Fung & Carter, 2007). In BATTICC-F the most widely used DMs of this type are *yeah* (285 instances) and *oh* (68 instances), showing responses and feedback in a conversation. As for shared knowledge marking, *you know* is the most typical and common form (28 instances), and most of the instances of *sort of* indicate attitude of the speaker (23 instances). The pervasive use of *yeah* can be seen in the following excerpt. Most of them serve as an interjection, which has been discussed in detail in 4.7.2.

(131) <BT18>: Do you like fish and chips? Have you tried that yet ... you haven't tried that yet have you?
<TW16>: ***Yeah***, I have not tried... I don't try yet.
<BT18>: Okay, you should.
<TW16>:***yeah yeah*** Er I think the fishes smell not very good.
<BT18>: ***Yeah***, no I don't like fish.
<TW16>: ***Yeah***, I don't like fish too.

(BATTICC-F)

In (131) most uses of *yeah* are not equivalent to a direct positive response *yes* or an agreement with a prior statement. Rather, they serve to express "a general acknowledgment of the previous interactive unit" (Jucker & Smith, 1998, p. 181). That is, they are commonly used by listeners as "back channels" to signal that what is being said is followed and supported. In this way, interpersonal DMs indicate active participation and positive listenership (Fung & Carter, 2007), and they further help "stake out interpersonal territory, focus on the other in speaking

and listening and are essential for successful communication" (Carter, 2008, p. 15). Besides *yeah*, common response tokens found in BATTICC-F include *okay*, *right*, *alright* and *I see*, as can be seen in the following excerpts.

(132) <TW15>: Ah ... I think ... erm ... Have you ever been here before?
<BT17>: ***Yeah***, I have been up here twice.
<TW15>: Twice... ***okay right*** ... that's cool. I think here is very beautiful.
(133) <BT21>: Young-Min Shan
<TW19>: Young-Min mountain, ***yeah***, with my parents and my dog.
<BT21>: ***Alright, okay***, do you ... okay so you go as a family then.
<TW19>: ***Yeah***.

(BATTICC-F)

In addition to *yeah*, *oh* is an even more common discourse marker and interjection in the participants' conversation, in particular "to respond to new information or to indicate that a speaker has just discovered something surprising" (Carter & McCarthy, 2006, p. 115). It "pertains primarily to the information state, signaling some change in the speaker's cognitive state" (Norrick, 2009, p. 875), and it is usually used to "express receipt of new information" (Fraser, 1996, p. 172), as in (134)-(135):

(134) <TW07>: How about ... er ... remember today we're climbing the mountain right?
<BT09>: ***Oh***, that was really funny because I nearly fell over.
(135) <TW01>: so there are still questions. How are you today? Good, good?
<BT01>: Even better since we talked to you.
<TW01>: ***Oh***, really?

(BATTICC-F)

The *Oh* in these examples seems to convey the message that the speakers have just received new information and understood it. Heritage (1984) characterises the interjection *oh* as "a particle ... used to propose that its producer has undergone some kind of a change in his or her locally current state of knowledge,

information, orientation or awareness" and it provides "a fugitive commentary on the speaker's mind" (pp. 299-300). Moreover, *oh* sometimes occurs with other interjections or discourse markers. In the example below, *oh* is used with *yeah*, *no* and *well* in the initial turn position.

(136) <TW13>: Do you read the Harry Potter?
<BT16>: ***Oh yeah***, I like the Harry Potter books. I have, you know, the first one ...

(137) <TW11>: Ohhh... okay I know. So we have to talk typhoon?
<BT15>: ***Oh no***, this is just something from ... because I'm doing a weather project over here because we don't get typhoons in the UK so we've been asked to find out about them.

(138) <BT20>: Your nickname is big mountain.
<TW18>: No.... it's my real name ... my real name is big mountain.
<BT20>: ***Oh well.*** [laughter]

(BATTICC-F)

The use of *oh yeah* and *oh no* in the excerpt seem to be just an intensifier of *yeah* and *no* respectively. Some of the examples of *oh no* were used as a self-initiated repair. In example (139), BT08 answered BT07's question (*I did*) but did so incorrectly. She suddenly realised her own mistake (*oh no*) and then replaced her prior answer (*I didn't*).

(139) <BT07>: You didn't do Awkward Giraffe?
<BT08>: I did, ***oh no*** I didn't – I forgot about Awkward Giraffe. And I forgot to say how you hold your cucumber.
<BT07>: I know I was so upset.

(BATTICC-F)

Another commonly used DM serving interpersonal function is *you know* (26 instances), which generally marks statements as representing assumed shared knowledge or experience between speakers and hearers (Carter & McCarthy, 2006; Jucker & Smith, 1998). It is particularly common in casual conversation, ranking as the most frequent two-word sequence in most of the corpora of informal spoken

discourse. As such, in a way it makes speech more casual and marks a high degree of intimacy and in-group membership (O'Keeffe et al., 2007). Östman (1981) proposes that the highly frequent use of *you know* is to show that "[t]he speaker strives towards getting the addressee to cooperate and/or to accept the propositional content of his utterance as mutual background knowledge" (p. 17). For example:

(140) <BT13>: Hey Aiden – ***you know*** last night at the meeting thing –
<BT14>: Yeah.
<BT13>: ... did you see that cat man who was there?

(141) <TW11>: In in typhoon, it's very .. very bad, ***you know***, it's it's wet =
<BT15>: Yeah.
<TW11>: =because it's raining and it's cold.
<BT15>: Windy as well.

(142) <BT07>: I would like to know ... are you used to like how hot it is. Like we find it really like warm, ***you know***, like the weather?
<TW07>: No. .. erm... do you find Taiwan hot or like cold?
(BATTICC-F)

In both cases, *you know* is used by speakers to invite addressee inferences based on their shared experience or knowledge. In (140) both BT13 and BT14 might be familiar with what BT13 said *last night at the meeting thing*; in (141) TW11 is talking about typhoons and is appealing to BT15's shared understanding about them. In the conversations it can also be seen that *yeah* is used as an acknowledgement in the two cases, and this is expected since in inviting inferences participants in conversation normally back-channel to show their understanding. In addition, from the excerpts, the use of *you know* also indicates that the speaker may not only want to appeal to the shared knowledge but also desire the interlocutors to participate and share more about their own ideas. As Jucker and Smith (1998) argue, *you know* does not just simply indicate that the recipient knows the information, but it often serves as "a device to aid in the joint

construction of the representation of the event being described ... *you know* invites the addressee to recognise both the relevance and the implications of the utterance" (p. 194), thereby making communication more interactive, involving and informal (Fung & Carter, 2007).

b) Referential DMs

Referential DMs indicate relationships between utterances. Fraser (1999) states that they "impose a relationship between some aspect of the discourse segment they are a part of ... and some aspect of a prior discourse segment" (p. 938). The most common DMs of this type in BATTICC-F include coordinative, i.e., *and* (237 instances), contrastive, i.e., *but* (110 instances), consequential, i.e., *so* (108 instances), causal, i.e., *cos/because* (69 instances), disjunctive, i.e., *or* (7 instances) and digressive, i.e., *anyway* (4 instances). As can be seen in the following examples, most of the DMs relate the discourse segment they introduce (e.g., *I get to see her do it again* in (143)) with the prior segment (e.g., *I feel very happy*). It is worth noting, however, that not all of the items in bold in the extracts function as a DM. To take (147) for example, the first *and* purely serves as a conjunction within a message instead of introducing "a separate message with its propositional content" (Fraser, 1999, p. 939). Such a use of *and* is therefore excluded from the total amount of DMs in this analysis. Examples of the use of *and* as a DM have already been discussed in 4.6.1.

(143) <BT08>: I feel very happy ***coz*** I get to see her do it again...
(144) <BT16>: then I tried the drums and I was good ***so*** I like it [laughter].
(145) <TW15>:Oh ... I like running, ***but*** I like team sport better.
(146) <BT09>: Yeah, are you used to the weather ***or*** do you complain?
(147) <BT17>: I liked walking around with ... I walked around with Aiden ***and*** Katie ***and*** it was very fun.

(BATTICC-F)

In (143) and (144) it is apparent that *coz* and *so* are markers of cause and result, in which BT08 gives the reason (i.e., *I get to see her do it again*) that causes him/her to feel very happy, and BT16 explains why he/she likes the drums. Moreover, *because/coz* and *so* occasionally co-occur in the same utterance, which is not generally accepted in traditional written grammar. In (148), *because* is used twice by BT15 to initiate two reasons for TW11's query, and the *so* is used to draw a conclusion upon the two reasons.

(148) <TW11>: Ohhh... okay I know. So er we have to talk about typhoon?
<BT15>: Oh no, this is just something from ... ***because*** I'm doing a weather project over here ***because*** we don't get typhoons in the UK er ***so*** we've been asked to find out about them.

(BATTICC-F)

Nevertheless, it can also be noted that there is a *so* in TW11's utterance in (148), which is not in bold due to the fact that it is not considered a referential DM. Rather, it may well serve a discourse function of topic transition and organisation, which will be further discussed in the next section on structural DMs.

c) Structural DMs

Structural DMs "provide information about the ways in which successive units of talk are linked to each other and how a sequence of verbal activities... are organised and managed" (Fung & Carter, 2007, p. 420). In BATTICC-F DMs like *so* (92 instances), *okay* (62 instances), *then* (30 instances) and *right* (16 instances) are most frequently found to serve such functions. One common use of structural DMs is to signal the opening or closing of a segment of conversation. For example:

(149) <BT21>: ***Okay***. Erm ... have you enjoyed today?
<TW19>: Yeah. As .. yes, I never go hiking with my friend.

(150) <BT01>: The time's very different in England ***so*** ...
<TW01>: ***So*** let's talk about er your performance.... what do you think about your performance.

(BATTICC-F)

Other than functioning as a response token or an interpersonal DM, *okay* is also found to be exploited as a structural DM, indicating turn opening, as can be seen in (149). On the other hand, in (150) the *so* in BT01's utterance may act as a turn yielding marker, marking the speaker's readiness to relinquish a turn, and such use of *so* is described as a "turn-transition device" (Schiffrin, 1987, p. 218). It can also be noted that the DM *so* in TW01's utterance clearly indicates the speaker's intention to change a topic in the conversation. Although *so* is one of the most common referential DMs, in more cases in my data it is considered a structural DM, a point that has been raised by Bolden (2009), Carter and McCarthy (2006) and Schiffrin (1987). Carter (2008) maintains that *so* very commonly acts as a DM, which indicates the beginning or end of a topic or a transition from one topic or bit of business to another (p. 14).

Another important function of structural DMs is to logically sequence the segments of talk. In BATTICC-F, *then* is the most common item of this type with 34 instances, most of which collocate with the coordinating conjunctions *and* (13 instances) and *but* (7 instances), as shown in the following excerpts.

(151) <TW16>: Yeah, have you tried to use chopsticks?
<BT18>: Er ... yeah. Yeah, I was sort of getting used to them by the end of my trip.
<TW16>: Really?
<BT18>: ***And then*** I got back home ***and then*** I tried using them and I couldn't really ... [laughter]

(152) <BT03>: We have after school clubs that you can go to. You take them in your own time.

<TW01>: I see.
<BT03>: ***But then*** you do have to balance that with exams, which we do ...
<BT03>: Yes, we're just doing exams and so on.
<BT02>: ***And then*** you've got homework.

(BATTICC-F)

Schiffrin (1987) states that *then* indicates "temporal succession between prior and upcoming talk" (p. 246). In (151) two instances of *and then* are used to signal the sequence of the talk and mark successive event time, showing the temporal relationship among the different activities mentioned by BT18. In addition, *and* or *but* in conjunction with *then* is frequently exploited as a turn initiator, as in (152); more precisely, Fung and Carter (2007) labeled them *continuers*, providing the prior speaker with a conversational space to expand upon. In this case, the additional utterance can be from the same speaker, as with BT03 in (152), where *but then* connects the two utterances from him/her. The continuers can also connect two utterances from different speakers. In (152), for example, *and then* preceding BT02's utterance indicates that he/she has something to say, adding more details to the previous comment initiated by BT03 in what became a jointly constructed explanation of school life in the UK. As Bolden (2009) states, "[the] discourse marker is a resource for establishing discourse coherence and, more fundamentally, accomplishing understanding" (p. 996).

d) Cognitive DMs

Cognitive DMs serve to denote the thinking process; they reformulate, elaborate and mark hesitation (Fung & Carter, 2007). The most widely used items include *like* (213 instances), *I think* (37 instances), *well* (18 instances) and *you know* (4 instances). For example:

(153) <BT18>: And, do you have, ***well*** you have ***like*** .. .you have more of ***sort of*** – yeah you have more ***sort of*** exotic fruits than we do.
<TW16>:Oh, really.

(154) <TW07>:Do you feel wow – why this weather in Taiwan is .... why this weather in Taiwan ... so hot.
<BT09>: ***Well***, we all complained like on the first day but then you gradually get used to it.

(155) <TW07>: So girls ... girls always put some make-up to go to school?
<BT09>: But not lots.
<BT08>: ***Well I think*** there's some people that take it a bit like too far ....

(BATTICC-F)

A number of different DMs denoting the thinking process can be found in my data, such as *well*, *like* and *sort of*. This is often connected with difficulties in speech production (Miskovic-Lukovic, 2009), which indicates a certain level of hesitance, planning of speech and searching for appropriate lexical items in that these expressions function as a filler or a time-buyer device in the discourse. As can be seen in (153), various types of cognitive DMs can be found. As Tsui (1994) claims, these perform a local coherence function and thus may well further develop speaking fluency in general. In particular, *well* is commonly found in the turn initial position, as in (154) and (155). Aijmer (2011) describes this as "primarily a "mental state" interjection" that can be associated with the speaker's deliberation (p. 235). Similarly, cognitive DMs sometimes co-occur to signpost the thinking process. As in (154), the turn is initiated by *well I think*, which indicates a hesitation and allows the speaker time to plan and keep a turn in an interaction. This may be due to the fact that an answer to the question asked by <TW07> is not immediately available. Such use of DMs may well also soften the expressions to some degree so that they do not appear too direct or unduly authoritative and assertive (O'Keeffe et al., 2007).

Although *you know* functions as an interpersonal DM, signaling shared knowledge, it might not always be the case that speakers and hearers have shared knowledge. In BATTICC-F the speakers occasionally use it for reformulating, repairing and exemplifying (Fox Tree & Schrock, 2002; Schiffrin, 1987). This use of *you know* is particularly common in Taiwanese learners' discourse. For example,

(156) <TW09>: Er Do you go to any cram school ... cram school?
<BT10>: Cram school?
<TW09>: Yes, cram school. ***You know***, like guitar or ....
<BT10>: Oh ... music lessons.

(157) <TW09>: So erm what do you usually to eat? ***you know*** you are so tall and I don't think look like a junior high school student?

(158) <BT09>: If you can't be bothered to go to work, you call a sickie.
<BT07>: Yeah, you just phone in and say that you're ill. ***You know***, you just don't go but like over here even like schools and stuff it's so much more ...

(BATTICC-F)

In these cases shown above, *you know* invites interlocutors to refer to the speakers' previous information. In (156) TW09 is asking a question about *cram school*, but BT10 seems to have no idea about what it is. TW09 then explicates and exemplifies the term by using *you know* turn-medially to elicit an addressee response. In (157) *you know* marks the speaker TW09's reformulation and modification of his/her question, and thus clarifies the intention of the speaker. Moreover, as in (158), *you know* also functions to highlight a particular point in the utterance (Fox Tree & Schrock, 2002). As in (158), BT07 reformulates his/her previous statement and further emphasises it.

From the excerpts discussed above, it seems that *you know* does not simply act as a filler or time-buyer; both Taiwanese and British learners use it as a pragmatic

marker for interpersonal, attitudinal and organisational purposes, which is broadly consistent with earlier research (e.g., Fung & Carter, 2007; Hellermann & Vergun, 2007; Jucker & Smith, 1998; Schiffrin, 1987). However, House (2009) argues that the functional use of *you know* by EFL learners and native speakers is markedly different in that EFL speakers use *you know* predominantly as a self-serving strategy to improve coherence rather than inviting addressee inferences or cooperating with their interlocutors. Although the results of this study do not fully support her conclusion, it is evident that relatively fewer instances of *you know* are found in Taiwanese learners' discourse, and they mainly use it as a cognitive DM.

The DM *like* is the most prevalent in my data, with a total of 213 instances in BATTICC-F. Such use of *like* has been proven to be particularly common in teenage talk (Andersen, 1998, 2000; Tagliamonte, 2005). Previous research has also reported the functional complexity of *like*. This can be seen in BATTICC-F in that the instances of *like* serve many different discourse functions, such as a quotative marker, focus marker, approximator, exemplifier, hedge, discourse link or hesitational device. One important function that has achieved much attention in the literature is as a quotative marker for introducing reported speech (Adolphs, 2010; Anderson, 2000; Hellermann & Vergun, 2007). According to Hellermann & Vergun (2007), "Quotative *like* is semantically the equivalent of 'say', except that it can be used to introduce inner monologue, speaker attitude, or non-verbatim renditions of dialogue" (p. 366). Adolphs (2010) also notes that *like* stands in the place of "*said that* plus quoted speech" (p. 182), as in the following instances:

(159) <BT16>: No, always ... in England I'm always ***like*** "Mum please buy me some" ...

(160) <BT09>: I was going to do some acting but then we were ***like*** ... "oh we can't be bothered".

(161) <BT07>: And we started pretending that we had all the [laughter] ... and we're ***like*** "where's the tree?".

(162) <BT07>: I think they're ***like*** ... in England they're sort of ***like*** ... "well I've missed it I'll just wait half an hour and get the next one".

(BATTICC-F)

The word *like* in (159)-(162) seems mainly to be used to introduce speech reports by the speakers. For example, in BT16's speech, *I'm always like* seems to be semantically similar to *I always say*, and the quoted speech then follows. Nevertheless, *like* might not be simply the equivalent of 'say', as was claimed by Adolphs (2010), as it "serves to dramatically highlight what follows and sets the stage for a speech report which is marked by its quotability, especially by its intensity and by the very prosodic contours which are reproduced" (p. 183). As such, the speech reports would become more vivid reproductions.

Another frequent use of the DM *like* is as a focus marker in that new information or the focus of the utterance is often followed by *like* (Fuller, 2003; Hellermann & Vergun, 2007). The following four examples illustrate this function:

(163) <BT01>: So you've got to have, ***like*** ... you can't have people behind you to see what you're really doing.

(164) <BT12>: Yeah, they don't .... over here they take lots of ***like*** care in the presentation ***like*** being clean and people have ***like*** a lot of respect for them.

(165) <BT07>: Yeah .... [laughter] We're not really used to ***like*** ... really spicy foods.

(166) <BT12>: The next thing you know it's gonna be ***like*** take your shirt off. [laughter]

(167) <TW03>: Er ***like*** the first we go to the trail...the trail ... Pretty tired and=

(BATTICC-F)

In these cases the information directly after *like*, which signals the element of focus in the utterances, could be phrases, as in (164) and (165), or complete sentences, as in (163), (166) and (167). Most of the *like*s in the excerpts are generally used to introduce new information and also the main idea that the speaker intends to convey. Underhill (1988, p. 236) considers such use, namely "*like* as a new information marker" the most salient function of *like*. Nevertheless, Anderson (2000) argues that it cannot only be considered as a new information marker, but it "plays the role in the process of utterance interpretation" (p. 228), and thus it is more socially accepted, particularly in the context of conversation among teenagers. Furthermore, it can also be noted that some *like* tokens in the excerpts seem to indicate more than one function. To take the use of *like* in (166) for example, the elements after *like* are clearly the focus of the utterance, while it also acts as a quotative, introducing the quotation *take your shirt off*. It appears that in this case *like* functions as both a quotative and a focus marker.

In addition, a number of instances of *like* act as an approximator, which is normally added to modify the following numeral phrases or other measurable units. As in the following excerpts, *like* may have a similar meaning to *roughly*, *approximately* or *about*, as the examples below demonstrate.

(168) <BT13>: I fell asleep in ***like*** half the films.
(169) <BT16>: ... Yeah, there was ***like*** 50 people who came to Dobby's funeral.
(170) <BT13>: I know I've got ***like*** £96.97 for $4,000 so I've just said I've got $4,000.
(BATTICC-F)

These examples illustrate *like* as an approximator. That is, for example, speaker BT16 does not necessarily mention the precise number of people who came to

Dobby's funeral, and 50 is an approximate number. O'Keeffe et al. (2007) state that "speakers frequently introduce approximators to downtone what might otherwise sound overly precise" (p. 177). But, as was shown in (170), the numeral unit following *like* is very precise. In this case, the number *£96.97* should be explained as the focus of the utterance, instead of an approximation marker. However, such use of *like* seems to have more than one function. For example, the *like* in (170) by BT13 approximates the amounts of money and at the same time the element introduced by *like* is also clearly the focus of the utterance. As Fuller (2003) notes, *like* indicates "looseness of meaning, or focus, or both" (p. 369). That is, *like* tokens can act as both a focus marker and an approximator at the same time, or they can clearly have one usage or the other.

The next excerpts derived from the BATTICC-F illustrate another function of the discourse marker *like* as an exemplifier. In these cases, speakers use *like* to support or illustrate their ideas by giving examples, based on shared knowledge and personal experiences. This gives listeners a clearer picture of what speakers are trying to convey, as in (171).

(171) <TW16>: Erm ... no, we don't get lots of berries in Taiwan. Yeah, but we have ***like*** water melon and banana.

(172) <TW01>: Okay. So er er::m what do you like to do after your school life?
<BT02>: Some jobs?
<BT01>: ***Like*** after school clubs?

(BATTICC-F)

Also, interlocutors sometimes co-construct a closer description of a particular point of reference by exemplifier *like* (Adolphs, 2010). In (172), for example, the speakers BT02 and BT01 collaboratively extend and clarify the question advanced

by TW01. Furthermore, *like* co-occurs very frequently with various types of vague expressions to convey exemplification and comparison, such as *things like that* and *sort of like*. This confirms the earlier observation in 5.5.1.

Like can also function as a hesitational or discourse linking device, indicating planning difficulties, false starts and self-repairs (Anderson, 2000). These three aspects indicated by *like* can be seen in the following examples.

(173) <BT23>: So ... erm.... you have ***like*** .... you still have ***like*** ... coz we went to Taroko in actual fact that's lots of mountains but not sort of the same – quite different.

(174) <TW03>: That's ... this experience, I'm very, very.. erm ... ***like*** ...very proud of it.

(175) <TW17>: Yeah. You don't have earthquake here?
<BT19>: No. We've had one earthquake but it was very small, it was just ***like***...
<TW17>: Ah ... but our earthquake is always very big.

(BATTICC-F)

From (173) to (175), *like* commonly co-occurs with pauses (i.e., ...), which indicates speaker engagement in thinking and a certain level of hesitation. This also allows speakers to buy time to think what they are going to say. In addition, they present a fitting paraphrase, as in (173), where self-repairs and false starts can be seen in the two instances of *like*. The first *like* shows that the speaker cuts off the utterances and resumes another, which presents the same syntactic structure with a minor correction and self-repair. In contrast, the sentences preceding and following the second *like* are syntactically unrelated in that the speaker resumes talk with a new syntactic structure, and this counts as a false start (Anderson, 2000). Moreover, in some cases, *like* occurs clause-finally, as in (175), where the speaker cuts off the utterance without resuming a new one. In this respect, the

speaker may intend to continue, but in light of planning difficulties or maybe interlocutor interruption, the speaker yields the turn.

Last but not least, *like* is occasionally used for hedging, which can mitigate the directness of utterances and operate as a face-saving device (Carter & McCarthy, 2006; O'Keeffe et al., 2007).

(176) <BT06>: In a way it was ***like*** a bit boring because we had to ...
(177) <BT08>: Erm .. .. it wasn't ***like*** amazing or anything. I'm just ***like*** weird anyway so people laugh at it.

(BATTICC-F)

It can also be noted that such use of *like* often occurs with other phrases marking hedging, which are often referred to as "vague language" (Carter & McCarthy, 2006). In (176), for example, the speaker BT06 uses *like* with *a bit* to hedge the statement. A similar situation can be found in (177). The first *like* occurs with the vague expression *or anything* and the other one is preceded by *just*. They function together as a discourse marker for hedges.

5.6 Does spoken grammar exist in CMC discourse?

The previous section has explored a number of distinctive lexical and grammatical features that commonly occur in spoken discourse (BATTICC-F). With these particular characteristics in mind, the attention is now turned to investigate the extent to which these features exist in the CMC discourse (BATTICC-O), focusing on the analysis of vague expressions, situational ellipsis, headers and tails, hesitation and discourse marking.

a) Vague expressions

As shown in Table 5.13, vague expressions can be commonly found in BATTICC-O, with 101, 24 and 136 entries for vague categories, hedges and approximation respectively. Regarding the distribution of vague expressions across online and spoken discourse, one clear difference we can see is that the frequencies of each type in BATTICC-F are generally higher than the ones in BATTICC-O, and in particular the different amounts of use of vague categories and hedges between the two datasets reaches a statistically significant level.

Table 5.13

Vague Expressions in BATTICC-F and BATTICC-O

Vague expressions	BATTICC-F		BATTICC-O		Sig.
	Number	per 1000 words	Number	per 1000 words	(LL)
Vague categories	**105**	**6.21**	**101**	**3.11**	***24.23
(and) stuff (like that)	17	1.01	12	0.37	
(that/this) sort of (thing/like)	12	0.71	4	0.12	
(or) anything (like that)	12	0.71	8	0.25	
(or/and) something (like that)	19	1.12	16	0.49	
(that/this) kind of (thing)	6	0.36	12	0.37	
(and) everything	10	0.59	3	0.09	
(and/but) thing(s) (like that)	29	1.72	46	1.42	
Hedges	**33**	**1.95**	**24**	**0.74**	***13.26
sort of	17	1.01	0	0.00	
a bit / a little bit	16	0.95	24	0.74	
Approximation	**68**	**4.02**	**136**	**4.19**	**-0.08**
about / around	12	0.71	35	1.08	
lots of / a lot	44	2.60	93	2.87	
loads of	8	0.47	5	0.15	
a couple of	4	0.24	3	0.09	
Total	**206**	**12.19**	**261**	**8.04**	***19.44

*** $p < .001$

The following examples taken from the BATTICC-O and CANELC present the use of vague expressions in online communication, particularly indicating a vague

category of specific items. Such use in most cases in my study involves the use of words or phrases such as *stuff, thing, sort of, kind of, everything, something* and other related vague expressions. For example:

(178) i love summer but then i dont think it get to hot and all the bee ***and stuff*** are horrible x In the summer i go camping with my dad.

(179) I occasionally will do that. I have a shower or bath and then get into bed and watch Wild at Heart and then I will sort my bag ***and everything*** out for school and then go to sleep.

(180) cool, I want to be a doctor because I have always liked ***that sort of*** job.

(BATTICC-O)

(181) it would be enough wouldn't it... to write ***something like that***. Even just once...
(CANELC)

(182) a Big Mac or a bucket of KFC?! It's usually the ***sort of thing*** we eat from the bag rather than taking it home and serving it up on a plate.
(CANELC)

In example (178) the use of *and stuff* after the word *bee* indicates an assumption on the part of the speaker that shared personal experiences in summer exist, therefore the people involved in the conversation understand what is included in *and stuff*. In example (179) *and everything* appears to have a basic function similar to *and stuff*, making "a call upon familiarity with assumed common ground" (Overstreet & Yule, 2002, p. 787). In this case most teenage participants would know *and everything* to involve the things that need to be prepared for school as they are all currently junior high school students. In making that call, as in (179), the informer may not necessarily need to list every specific item that he/she would sort out, and as a result these vague expressions are often used as "a marker of intersubjectivity" with the implicit meaning that "there is more to say on this, but I don't have to because you know what I mean" (Overstreet & Yule,

2002, p. 787).

In addition, as discussed in 4.1.2.1, vague quantifiers (e.g., *a bit*) also pragmatically function as downtoners, exhibiting illocutionary force with the textual utterance they accompany. Such use of vague language as hedging can be found in 24 instances in BATTICC-O, which is exemplified in the following cases.

(183) she is very lively and happy, she like to do as she is told but sometimes is ***a bit cheeky***. *:P*

(184) ir looks ***a bit strange*** but i looovveee oysters so i think it would be very very nice!!

(185) our school uniform is a blue, white and really dark blue tie, white shirt, black trousers or skirt and a black blazer. :/ its ok but its ***a bit uncomfortable*** :/

(BATTICC-O)

In these excerpts *a bit* is commonly prefaced to the different adjectives, such as *cheeky, strange* and *uncomfortable,* most of which seem to be used with negative situations. This confirms the results shown in the analysis of spoken discourse, as in 4.1.1.2. Such use of vague quantifiers serves to downtone and hedge the utterances, which is highly likely to be more appropriate in an informal context. In excerpt (184) for example, when the informer sees the food that he/she has never seen or had, although it might in fact be slightly odd for the informer, *a bit* is used adverbially to attenuate the negative assessment of the adjective *strange*. It is also noted that 17 instances of *sort of* serve this function in BATTICC-F, while such use of *sort of* is not found in BATTICC-O.

With respect to the use of approximations, a wide range of expressions can be found in BATTICC-O, with a total of 136 instances. The most prevalent item of

this type is *about* or *around*, which can be found serving this function in 41 instances. Other uses of approximations include *lots of* (24 instances), *a couple of* (6 instances), *loads of* (6 instances) and *for ages* (4 instances). The following extracts derived from BATTICC-O illustrate how the participants use approximations in online communication:

(186) Today in Cumbria there is ***loads of*** snow on the ground. It is ***around*** 15cm deep. It is really fun!!

(187) my worst school memory was when i didnt get a part in the school play in primary school, ***a couple of*** years ago now.

(188) Who knows how many people are acctually going to taiwan? is ther ***LOADS or what***. The question has been bugging me ***for ages*****.....**

(BATTICC-O)

Such items – *about, around, a couple of, loads of* and *for ages* – indicate the absence of precision. As in (187), *a couple of* refers vaguely to the amounts of years and *loads of* quantifies the amount of snow, and the exact amount of time or snow might not be relevant in these cases.

b) Situational ellipsis

Another distinct feature of spoken discourse is situational ellipsis, which can also be found in BATTICC-O. Herring (2011) classifies this as one of the most salient syntactic features of electronic language, which deviates from standard syntax and is sometimes described as "telegraphic" and fragmented (p. 5). A usual reason given for such syntactic reductions in online communication is to save keystrokes. In BATTICC-O the instances of situational ellipsis can be found in both groups of participants' data in that (189)-(193) were derived from the British participants' discourse, while (194)-(197) were Taiwanese leaners' language.

(189) (***I'm***) SO excited!! woohoo!

(190) (***I'm***) looking forward to meeting you all

(191) that time with indya when she got kn bella and she took one step and indya screamed 'ahhhhh! It moved!!! Haha (***They are***) good memories!! XD

(192) (***It***) looks a bit strange but i looovveee oysters so i think it would be very

(193) What (***do***) u find cool about your country?

(194) Wow, (***you are***) so talented!! I can't play any of them.

(195) (***It's***) Such a cute dog!

(196) (***It's***) Just not around my house.

(197) I think in Taiwan, (***it***) only snow in mointains....

(BATTICC-O)

Similar to the spoken discourse, the situational ellipses found in online discourse mainly occur at the beginning of utterances and involve the initial subject and sometimes its accompanying copular verb *be*. As can be seen in Table 5.14, the most common occurrences are the omission of subject pronouns such as *it* and demonstratives (22 instances), as in (195)-(197), and first person pronoun *I* (11 instances), as in (189) and (190). Some other subject pronouns such as *you, we and they* are sometimes intentionally omitted (3 instances), as in (191) and (194). Additionally, as in (193), the ellipsis of auxiliary verbs can be found. The two instances of this type found in my data do not include the omission of the subject. Although these elements in their sentences are omitted, they can be recovered by referring to preceding elements in the discourse so that there is no misunderstanding in CMC. However, the ellipsis of existential *there* and copular verb *be*, which can be commonly seen in BATTICC-F, are not found in BATTICC-O.

Table 5.14

Elements of Ellipsis in BATTICC-F and BATTICC-O

Element of ellipsis	BATTICC-F		BATTICC-O	
	Number	per 1000 words	Number	per 1000 words
Initial *I* (+*be*)	16	0.95	11	0.34
Personal pron. *they, he/she, you, we*	13	0.77	5	0.15
Interrogatives (+*be*)	7	0.41	3	0.09
It and demonstratives (+*be*)	27	1.60	22	0.68
Copular verb *be*	3	0.18	0	0.00
Existential *there*	6	0.36	0	0.00
Auxiliary verb *do, does, did*	5	0.30	2	0.06
Preposition	4	0.24	1	0.03
Final ellipsis	8	0.47	4	0.12
Total	89	5.27	48	1.48

c) Headers and Tails

While 11 entries of headers and 3 entries of tails are found in spoken discourse, BATTICC-F, only 2 instances of headers are found in online discourse, BATTICC-O, as in the following examples.

(198) But ***my father, he*** was the kind of prevent me from going to sports class

(199) ***My friends, they*** usually say that I am tall.

(BATTICC-O)

As in the examples above, the word or phrase that is fronted and comes first seems to indicate the emphasis of what the writer considers especially important, and this further provides readers with an orientation to the main points of the information. This is similar to their use in spoken discourse, although the difference in the frequencies of use is significant. This also shows that spoken language is much more flexible than written forms are with respect to the word order in an utterance (Carter & McCarthy, 2006; Mumford, 2009). This may be due to the fact that spoken discourse is constructed in real time and follows the order of ideas emerging from a speaker, while asynchronous communication is still restricted to its written nature and thus fewer instances of headers and tails are

found in BATTICC-O.

d) Pausing, repeating and recasting

As discussed in 5.4.4, pausing, repeating and recasting are types of characteristics that typically feature in naturally occurring speech. In the transcripts of BATTICC-F they are simply a silence labeled as a sequence of full stops, such as (..) for a brief break or (...) for a longer pause. However, these features are abundant throughout the BATTICC-O, as in the following examples.

(200) that's a really good resoluton....haha ...but I don't think I can do it....6 months is too long.... I can do it for only 6 hours.....haha

(201) well we have quite alot really....but the main would be christmas n easter...our easter holidays are coming up...we have near enough 3 weeks off coz of the royal wedding and may day.

(BATTICC-O)

In BATTICC-F both Taiwanese and British participants' discourse present repeated full stops in their messages, as in (200) and (201) respectively. In the extracts the repeats of full stops seem to replace most of the punctuation in the messages. This supports the earlier findings in 4.4 showing that the users of text-based CMC usually hold a rather lax attitude towards the use of punctuation and orthographical correctness. In BATTICC-O the repeated periods mainly substitute full stops (48%) and commas (43%), and only 9% them of are used for replacing other punctuation, such as exclamation marks, question marks and colons. These triple-dot punctuation marks seem to indicate a slight pause among the sentences or clauses in a message.

In addition, repeating and recasting are common in naturally occurring spoken discourse under the pressures of real time. Although many instances of repeating

and recasting are found in BATTICC-F, there are only a few examples in BATTICC-O. Repeating often marks a hesitation or planning on the part of the speaker in BATTICC-F, while in BATTICC-O it is used for additional emphasis, as in:

(202) I stopped when I was eight because the ice rink was too far from where i lived but I hope i'm going to start again soon!! I ***love love love*** it!!

(203) Thank you ***very very very*** much....

(BATTICC-O)

From the examples it is apparent that the writers purposefully use the consecutive words to add more emphasis to their messages instead of showing a kind of hesitation or planning their next utterances. It appears that discourse functions indicated by repeating in natural spoken and online discourse are different.

e) Discourse marking

As shown in Table 5.15, the DMs commonly used in BATTICC-F can also be found in online discourse BATTICC-O, with a frequency of 49.74 entries per 1000 words. From the table, one clear point is that DMs are generally more frequently used in spoken data, which is particularly evident in the use of interpersonal, structural and cognitive DMs. Within the category of interpersonal DMs, *oh* is most prevalent. Its major function as a DM in real-time speech is to "respond to new information or to indicate that a speaker has just discovered something surprising" (Carter & McCarthy, 2006, p. 115), as in (204) and (205).

(204) ***oh*** wow cindy thank you very much!! that is really cool!! I love it

(205) ***oh***, It looks delicious. I like to drink Pearl Milk Tea, too. Haha

(BATTICC-O)

But *oh* does not always indicate a positive emotion. As can be seen in (206) and

(207), the authors' statements seem to indicate a kind of disappointment. In (207) particularly, *oh* is written in a repeated vocal spelling, which intensifies the tone of voice.

(206) I like to play basketball too and badminton haha >P but at the moment I cant play outside because of all the snow!! ***oh*** how i wish it would go away!!

(207) I think I can as long as there are no buiscuits in the house :/ ***ohhh*** dear

(BATTICC-O)

Table 5.15

Discourse Markers in BATTICC-O and BATTICC-F

DMs	BATTICC-O		DMs	BATTICC-F	
	Number	per 1000 words		Number	per 1000 words
Interpersonal DMs	**52**	**1.60**		**392**	**23.20**
oh	25	0.77	*yeah*	285	16.86
yeah	17	0.52	*oh*	68	4.02
actually	5	0.15	*sort of*	23	1.36
sort of	5	0.15	*you know*	16	0.95
Referential DMs*	**1279**	**39.42**		**522**	**30.89**
*and**	890	27.43	*and*	237	14.02
*but**	250	7.70	*but*	110	6.51
coz/because	104	3.21	*so*	108	6.39
so	35	1.08	*coz/because*	67	3.96
Structural DMs	**139**	**4.28**		**198**	**11.71**
*then**	114	3.51	*so*	82	4.86
so	10	0.31	*okay/OK*	69	4.08
okay/OK	10	0.31	*then*	34	2.01
finally	5	0.15	*right*	13	0.77
Cognitive DMs	**190**	**5.86**		**278**	**16.45**
like	102	3.14	*like*	213	12.60
*I think**	75	2.31	*I think**	37	2.19
well	11	0.34	*well*	18	1.07
I mean	2	0.06	*you know*	10	0.59
Total	**1660**	**51.16**		**1400**	**82.85**

* Item has a higher frequency in BATTICC-O than in BATTICC-F

Referential DMs are the only category that is significantly more common in online discourse than in spoken data. In particular the frequency of *and* in BATTICC-O significantly outnumbers that in BATTICC-F, and it is obviously the most frequent DM among the whole dataset. Section 4.1.5.1 has offered a detailed discussion of its discourse functions, which mainly serve an important cohesive link between sentences so utterances are linked as if in a chain (Carter & McCarthy, 2006). Similarly, other items such as *but* and *so* are frequently used in online communication, and these are mainly used to introduce a separate message with their propositional content (Fraser, 1999, p. 939), as in the following excerpts.

(208) haha when Tyra was little she used to run under peoples feet ***so*** she got stood on quite a bit which wasn't great haha ***but*** she is ok now ***and*** she won't run under your feet!!

(209) I have just had a maths exam ***but*** I don't know how I did yet ***but*** I am not very good at maths ***so*** I don't know haha

In the use of structural DMs, *then* is the most frequent in BATTICC-O, and this is also the only DM that is used significantly more often in online than in spoken discourse. As can be seen in the following excerpts, *then* often occurs multiple times in a single message.

(210) I have Art and English ***then*** break at 10.45am I ***then*** have I.C.T and Science ***then*** dinner at 12.45pm after that I have Geography and Maths I ***then*** go home at 3.15pm

(211) at christmas i open my christmas presents ***then*** i have a meal with my family. ***then*** i go to my other dads and have some more presents.

(BATTICC-O)

The predominant use of *then* can be attributed to the fact that in intercultural exchange participants very often share their personal life experiences with others

on electronic discussion boards. Since they frequently talk about their everyday lives and compare the differences between two cultures, *then* is commonly used to indicate a timeline in a sequence of daily events. Another item that needs to be stressed is *so* as a structural DM. In spoken data *so* serves rather equal referential and structural purposes, while in online discourse only three instances of *so* function as a structural DM, as shown in the following extracts.

(212) ***So*** everyone, do you have any new years resolutions you would like to share??

(213) My names Caitlin-Anne
I can't wait for Taiwan to try new things and meet new people :'))
So.... My question is what is the one thing You can't wait to do in Taiwan or What do you like doing in Taiwan?

(BATTICC-O)

As can be seen in (212) and (213), the two instances of *so* occur at the turn initial position, which indicates the beginning of a topic or a transition from one topic to another. With reference to cognitive DMs, *like* is most predominant in BATTICC-O. As has been discussed in 5.5.5, the DM *like* in spoken discourse serves as a quotative marker, focus marker, approximator, exemplifier, hedge, discourse link or hesitational device in BATTICC-F. However, there are not many functions of *like* found in the online context (BATTICC-O), with it mainly serving as a focus marker, as in (214) and (215), or an exemplifier, as in (216) and (217).

(214) that was soooo funny!! the teachers were ***like*** going balistic in a very quiet way! haha very funny indeed!!

(215) especially when it is snowy, ***like*** it is now. It is very cold.

(216) I was wondering what your school is like? Do you do anything ***like*** extra curricular clubs?

(217) There are many festivals in Taiwan, ***like*** Chinese New Year, Dragon Boat festival, Moon Festival....

(BATTICC-O)

The results presented in this section show that a large number of features seen in BATTICC-O approximate the spoken forms of language, although the medium of online communication is in written form and is text-based. As has been discussed in Chapter 2, "electronic discourse is writing that very often reads as if it were being spoken – that is, as if the senders were writing talking" (Davis & Brewer, 1997, p. 2). This may be due to the informal and interpersonal nature of the intercultural exchange project. It is also likely done by users to economise on typing effort (e.g., situational ellipsis), mimic spoken language features or express themselves creatively (Carter, 2004; Crystal, 2011; Herring, 2013). Carter and McCarthy (2006) note that text-based online communication will continue to create more new forms of spoken English in its written forms.

5.7 Summary

The analyses undertaken in this chapter have provided an in-depth examination of the examples of how a particular linguistic feature is employed in online and face-to-face intercultural communication using a discourse analytical approach. This approach to the analysis of BATTICC-O and BATTICC-F has also elucidated how British and Taiwanese teenagers employ these distinctive features in an intercultural setting.

This chapter first examined the distinctive features of CMC, including use of the upper and lower cases, nonconventional spelling, emoticons and punctuation omission and repetition. The examination of these remarkable features demonstrated different preferences by the participants for different purposes. Although the BATTICC-O contains a wealth of cues in CMC, it was found that

Taiwanese learners use significantly fewer cues than British participants. Excluding these important features of CMC may result in potential ambiguity of their messages and lead to communication problems. As has been discussed in 5.3.5, the underuse of the CMC features or cues may be due to the fact that they are not taught in the EFL classroom and most of the Taiwanese learners rarely have opportunities to employ these strategies in real-life communication. In fact the exploitation of linguistic features of CMC, i.e., E-grammar, in the L2 classroom can be very interesting to learners, and this may further be beneficial for them in terms of improving vocabulary use and acquiring a better understanding of linguistic appropriateness. As Averianova (2012) suggests, EFL teachers can introduce students to conventions of electronic discourse and invite them to critically evaluate different samples of CMC messages with regard to their comprehensibility, appropriateness, intercultural sensitivity and other relevant characteristics. Based on this concept, sample material for teaching E-grammar is presented in Appendix G, which includes authentic CMC messages from the discussion board (BATTICC-O) and displays a number of exercises for learners to practise the interpretation and codification of English online discourse.

This chapter then examined the most important linguistic features of spoken discourse (BATTICC-F), paying particular attention to the analysis of vague expressions, approximations and hedging (5.5.1), situational ellipsis (5.5.2), headers and tails (5.5.3), pausing and repeating (5.5.4) and discourse marking (5.5.5). It has been shown that British participants generally produce more instances of vague categories, approximation, hedging and discourse marking, especially hedging and vague categories, which reach a highly significant level. As has been discussed, these features of spoken grammar have very important

discourse functions, such as organising the utterances by "breaking up utterances into smaller 'meaning chunks'", which may actually aid comprehension (Gilmore, 2004, p. 369). They can also indicate turn-taking (Carter & McCarthy, 2006), helping speakers keep the floor while formulating their next utterance, or in some cases indicating that they are ready to relinquish the floor. Moreover, they serve important interpersonal functions (O'Keeffe et al., 2007), which are highly relevant to successful interaction in an informal communication setting. As a result, it would be helpful to include these important spoken features that commonly occur in authentic data in any EFL syllabus. On the other hand, situational ellipsis, headers and tails, pauses, repeating and recasting are more commonly used by Taiwanese students, and the use of pausing between the two groups achieves a significant difference. This can be explained by the fact that the Taiwanese participants generally do not have high-proficiency skills in speaking, and the "planners" therefore were employed more frequently than the British participants.

We also investigated the extent to which the most distinctive features of spoken discourse exist in CMC. It was found that a large number of features identified in BATTICC-O approximate the spoken forms of language presented in BATTICC-F, although the medium of online communication is in written form and is text-based. In addition, the language usage of CMC and spoken discourse displays a high level of informal interaction. While many of these distinctive features do not contribute any specific content or propositions, they have important interpersonal functions and particularly appeal to young people. For EFL learners who attempt to create and maintain good relationships in intercultural conversation, it would therefore be very helpful to be aware of and learn these features with a view to

becoming an intercultural speaker, and this is worthwhile and "should not be considered a poor imitation of native speaker competence" (Byram, 2012, p. 89).

In the next chapter I will turn my attention to the use of recurrent multi-word sequences, which have been an important part of the analysis in this and preceding chapters but which have not yet received specific analytical attention. Again, I will focus on the patterns of language use by Taiwanese and British participants in intercultural contexts, taking into account the different types of communication modes in which they are uttered.

CHAPTER 6

Online and Spoken Discourse: A Multi-word Sequence Perspective

6.1 Introduction

Chapters 4 and 5 of this thesis have explored the distinctive linguistic features in online and spoken intercultural communication, some of which indicate that many recurrent multi-word sequences are as frequent as or more frequent than single-word lexical items. This has also been highlighted in a range of previous research, as reviewed and discussed in 3.4.5. Given the phrasal nature of the English language, both written and spoken discourse contains a large proportion of highly recurrent sequences of words. As has been defined in Chapter 3, *multi-word sequence* is considered as an umbrella term to cover all types of "frequently occurring contiguous words that constitute a phrase or a pattern of use" (Greaves & Warren, 2010, p. 213). That is, multi-word sequences are simply sequences of word forms that commonly go together in discourse, regardless of whether they are grammatically or semantically complete.

This analysis focuses specifically on three-word sequences (e.g., *I don't know, I would like*) as they include sufficient contextual information for the assessment of units' discourse functions, and they are also analytically more manageable. In addition, I particularly investigate two important aspects of three-word units: functional and developmental perspectives. Functional use of three-word units concentrates on the most frequent items in the two different communicative modes, namely computer-mediated communication (CMC) and face-to-face (FTF)

interaction. The primary discourse functions of the high-frequency three-word units in the two settings are first identified. The differences in how two groups of participants – British and Taiwanese teenagers – used the three-word units for different functions and in different registers are then examined. The second part of this chapter investigates the extent to which intercultural exposure to native English speakers affects the use of recurrent sequences by young Taiwanese learners over one year of contact. The multi-word sequences frequently used by Taiwanese participants during the exchange programme were analysed in order to examine their approximation to those sequences used by native English-speaking participants. Research findings outline how multi-word sequence analysis can inform EFL teachers in relation to course design for intercultural communication.

6.2 Most frequent three-word sequences from CMC to FTF

The 10 most-frequent three-word sequences are derived from the participants' discourse as displayed in Table 6.1, showing their use over time from a three-phase CMC to FTF communication by the Taiwanese and British participants. In Phase One, it can be seen that the majority of the most frequently used sequences are in relation to personal introductions regarding names (e.g., *my name is*), ages (e.g., *am # years*), and birthdays (e.g., *my birthday is*). This may be due to the nature of the online community, which generally starts with personal introductions. From the table it can also be noted that there are two sequences that are frequently used by both groups of young people, namely *my favourite food* and *favourite food is*, which are both likely to be part of an extended sequence such as *my favourite food is*. The sequences surrounding the frequently-used lexical item *food* indicate that food culture was commonly discussed in Phases

One and Two of CMC.

When entering Phases Two and Three, on the other hand, the participants used more expressions concerned with the elicitation of opinions and knowledge such as *do you like*, *do you have* and *what kind of*. That is, the types of discourse tend to shift from basically transactional, transmitting factual information, to increasingly interactional, used for maintaining social relationships (Nattinger & DeCarrica, 1992). Such sequences can be found with an extremely high frequency in FTF interaction. For example, the top five high-frequency sequences (e.g., *do you have, I don't know*) mark the highly interactional nature of FTF communication.

Table 6.1

The Ten Most Frequent Three-word Sequences Over Time

	CMC Phase One		CMC Phase Two		CMC Phase Three		FTF interaction	
	Sequence	Freq.	Sequence	Freq.	Sequence	Freq.	Sequence	Freq.
1	my name is	64	I like to	43	I like to	34	do you like	41
2	I am #	48	a lot of	30	I am very	27	do you have	32
3	am # years	39	my name is	28	my name is	27	I don't know	24
4	I like to	38	my school is	26	I can't wait	23	do you want	20
5	my favourite food	30	Chinese New Year	26	it is very	22	what do you	19
6	favourite food is	28	I have a	25	do you have	21	you want to	18
7	junior high school	28	do you have	24	a lot of	21	I want to	18
8	and I am	25	do you like	23	with my friends	20	it was very	18
9	I have a	25	we have a	18	but I am	17	fish and chips	16
10	my birthday is	24	my favourite food	18	what kind of	15	go to school	16

From Table 6.1 different use of personal pronouns can be seen. In Phase Two, for example, the sequence *we have a* is ranked ninth with 18 occurrences. The use of *we* in this case indicates the level of focus on involvement with others, shifting

from self-identity to group identity (Liaw, 2010). Other sequences in Phases Two and Three involving the use of *we* and occurring more than 5 times include *we had a*, *when we are* and *we sometimes do*. The frequency information also reveals that the numbers of *we* in the three stages of CMC are 91, 141 and 138 respectively. Such increasing tendency of the use of *we* from Phase One to the second and third phases is probably a natural process of relationship building in that young people share their experiences and identify themselves as group members of an online community.

Table 6.2

Three-word Sequences Produced by the Participants

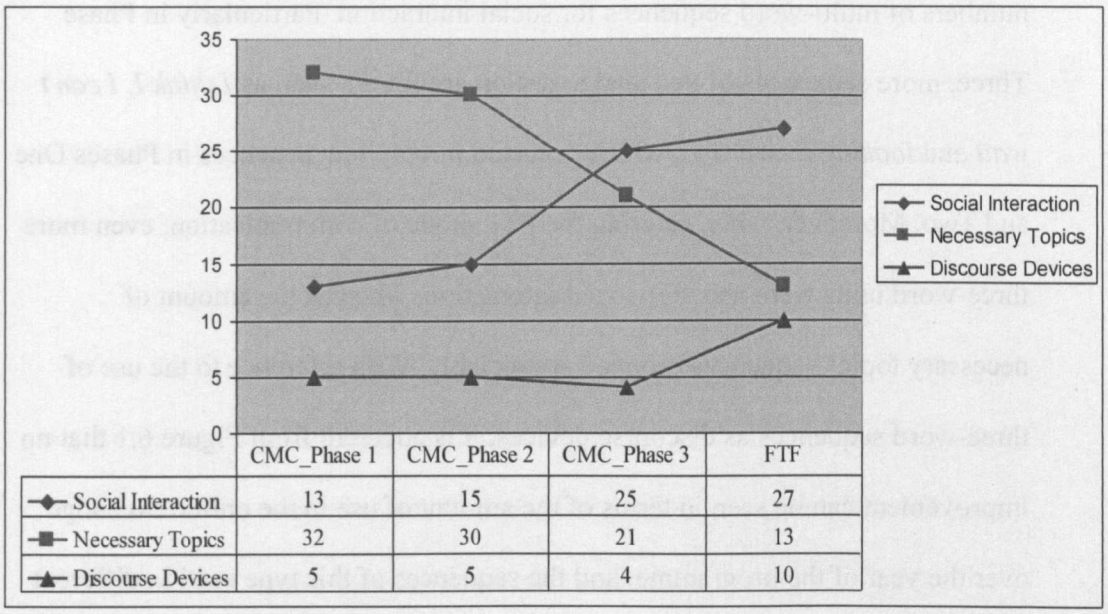
	Taiwanese participants				British participants			
	Phase One	Phase Two	Phase Three	FTF	Phase One	Phase Two	Phase Three	FTF
Number of words	5828	5642	5745	5238	5944	5771	5846	10746
Number of 3-word sequences	112	94	91	69	164	93	98	109
Percentage of 3-word sequences	13.99%	8.85%	7.68%	6.46%	13.53%	7.27%	6.98%	5.66%
Type/Token Ratio	18.39	21.22	21.48	16.61	18.03	20.68	21.06	13.45

Aside from illustrating the 10 most frequent three-word sequences, it is worth taking into account the cumulative use of three-word units over time by both groups of young people. The frequency cut-off of at least three occurrences in my data was established as the criterion for inclusion since the default cut-off figure set by *Wordsmith Tools* of two occurrences in any corpus is highly likely to return combinations that occur by chance. The numbers of three-word sequences occurring at least three times and their cumulative percentages are presented in Table 6.2. With regard to the percentage of use, the table demonstrates a steady decline in the use of three-word sequences in both Taiwanese and British

discourse over the three phases of online interaction, beginning at approximately 13% in Phase One of CMC, followed by a steady decrease in Phases Two and Three. The observed decline in this case could be attributed to two factors. First, it may be a result of the increase in variation in the participants' lexical and grammatical choices (Adolphs & Durow, 2004). A useful way of estimating the degree of productivity of a language sample is the lexical type/token ratio, calculated by dividing the number of different words (types) by the total number of words (tokens) (Abrams, 2003; Schmitt, 2010). *WordSmith Tools* display the increase of type/token ratio of the two groups in the three phases, rising from 18.39 to 21.22 and reaching 21.48 in the Taiwanese students' dataset and rising from 18.03 to 20.68 and reaching 21.06 in the British learners' discourse. Such increase is generally considered to indicate greater lexical diversity. Second, the decrease in the use of three-word units over time shows that the online exchanges of young people initially tended to rely more on specific multi-word formulaic expressions in order to build friendships and maintain relationships.

As shown in Table 6.2, an interesting aspect of the data is that while the percentage of three-word sequences used by Taiwanese learners is higher than for the British participants throughout the three phases, the reverse occurs in relation to the number of sequences employed. The fact that British pupils use a greater number of sequences than Taiwanese learners indicates that there is less variety in the use of multi-word sequences by Taiwanese learners than for native English speakers. This simply reflects the fact that non-native English speakers tend to restrict themselves to a small selection of over-used sequences, and this can be explained in part by the lower lexical diversity and lexical coverage of non-native speakers (Crossley & Salsbury, 2011; Wray, 2002).

Another interesting finding is probably the quantity of three-word units identified in online and spoken data. From Table 2 it is apparent that the percentage figures of the multi-word sequences used in FTF interaction (6.46% by Taiwanese participants and 5.66% by British participants) are lower than those used in CMC. These results seem to contradict the generally accepted estimate that there are many more multi-word sequences in spoken data than in written discourse, as shown in previous studies (e.g., Biber, 2009; Erman & Warren, 2000).



	CMC_Phase 1	CMC_Phase 2	CMC_Phase 3	FTF
Social Interaction	13	15	25	27
Necessary Topics	32	30	21	13
Discourse Devices	5	5	4	10

Figure 6.1

Functional Use of Multi-word Sequences Across Four Phases of Intercultural Exchange

Furthermore, the top 50 highly recurrent three-word units retrieved from different phases were inductively grouped to their functional categories based on their usage in the texts. Figure 6.1 above illustrates the distribution of functional use across the four phases of the intercultural exchange programme. With regard to the sequences for social interaction, a subtle growth can be seen from Phase One

(26%) to Phase Two (30%), and the rate of increase is much more extreme from Phase Two (30%) to Phase Three (50%). On the other hand, there is a clear decreasing trend in the use of three-word units for necessary topics in the percentage of top 50 sequences over time, which began at 64% in Phase One, followed by a slight decline to 60% in Phase Two and then a considerable drop to 42% in Phase Three. This seems to show that the participants frequently talked about specific topics on the discussion board in the first two phases of CMC, and as their relationship was gradually built over time, they used increasingly large numbers of multi-word sequences for social interaction. Particularly in Phase Three, more sequences of personal assertion are found, such as *I think I, I can't wait* and *looking forward to*, which occurred in very few instances in Phases One and Two. Moreover, when entering the FTF mode of communication, even more three-word units were used for social interaction, whereas the amount of necessary topics sequences dropped appreciably. With reference to the use of three-word sequences as discourse devices, it is apparent from Figure 6.1 that no improvement can be seen in terms of the amount of use in the online exchange over the year of the programme, and the sequences of this type used in different phases are also quite similar. For example, the sequences serving linking functions such as *and I am, and I love* and *and we have* occur in the top 50 highly recurrent sequences in all three phases of CMC.

6.3 Discourse functions of recurrent three-word sequences

The previous section has demonstrated the use of three-word units over time in the one-year programme. This section takes a closer look at the discourse functions that recurrent three-word units serve in the intercultural context of CMC and

spoken discourse, and also compares the items frequently used across two modes of communication. I firstly explore the primary discourse functions served by the three-word units commonly used in two different communication modes. I then examine the extent to which the use of three-word units differs in different discourse functions in different registers. Furthermore, I scrutinise the different use of sequences by different groups of participants, namely British and Taiwanese young learners.

Table 6.3 lists the 50 most common three-word units derived from BATTICC-O and BATTICC-F, as well as large reference corpora of online discourse (CANELC) and spoken discourse (CANCODE) for further comparison. The recurrent three-word sequences were inductively grouped into three central categories with regard to the primary discourse function that they served in my data. The three domains are social interactions, necessary topics and discourse devices, which will be discussed further in the following sub-sections. It is, nevertheless, worth noting that a number of sequences do not have a clearly recognisable function, such as *to go to, to be a, to be the, to do it* and *to have a*, which are mainly composed of high-frequency function words. These items appear widely across online and spoken data, and this may simply be due to the highly recurrent nature of these grammatical fragments (Biber, 2009; Ellis et al., 2008). In this analysis, therefore, these five three-word units are excluded since they might not be helpful for this present research. The following subsections will concentrate on the three different functional categories: social interactions, necessary topics and discourse devices, followed by a discussion of the distribution of high-frequency three-word units across functional types.

6.3.1 Social interaction

One common function for which multi-word sequences are often employed is the maintenance of social interaction (see Nattinger & DeCarrico, 1992; Schmitt & Carter, 2004; Wray & Perkins, 2000). In this category, a large amount of conventionalised language is typically associated with different speech acts in social interaction, such as *thanks for the* to express politeness, *it would be* to comply with a request, *I can't wait* to express personal intention, and *would you like* to express an offer. However, a single unit might sometimes serve multiple functions. In this case, for example, the sequence *do you want* derived from BATTICC-F was used by participants for different speech acts. One such example is an offer, which is a speech act in which "the speakers volunteer to do something beneficial for the listener (or a third party) or give something to the listener (or a third party)" (Carter & McCarthy, 2006, p. 699). In the following two examples, although the surface form is a question, it is apparent that offers are being made by both British pupils (i.e., BT06) in (1) and Taiwanese students (i.e., TW10) in (2).

(1) <BT06>: Someone gave it to me. [laughter] ***Do you want*** it?
<TW05>: Yeah. Thanks. (passing the item)

(2) <TW10>: Look at that ... ***Do you want*** to write England in Chinese?
<BT13>: Yeah. Have a go.
(<TW10> is writing on <BT13>'s workbook.)

(BATTICC-F)

In (1), according to Levinson (1983), BT06's utterance is both an offer and a question in that *yeah* is used to respond to the question, while *thanks* responds to the offer. In this particular situation, BT06 passes the item to TW05 following TW05's response. It is obvious that *do you want* in (1) is used as an offer of the physical thing since if BT06's utterance is simply a question, BT06 would have no

obligation to do anything. Similarly, in (2), TW10 offers to do the Chinese writing for BT13, although no *thank you* is included in the response.

Since more instances of the use of *do you want* serve the function of offers in my data, the sequence is categorised in this sub-group. Similar to such use, *do you want* was occasionally used as an invitation, offering an opportunity to do or share something pleasurable with the speaker (Carter & McCarthy, 2006), as in (3), (4) and (5).

(3) <BT01>: ***Do you want*** to come?
(4) <TW15>:***Do you want*** to come to Taiwan again?
(5) <TW10>: Yeah ***do you want*** to try?

(BATTICC-F)

In other cases, *do you want* was used in the form of questions as requests or polite directives, which have the purpose of eliciting information (Tsui, 1994), as in (6) and (7). In such cases, the speaker TW15 wishes the interlocutor to write his/her birthday on a card, and TW07 makes a request for everyone's mobile phone number. *Do you want* was also used simply as a question in an interrogative form in that speakers asked specific information about a particular issue, event or other related topics, as in (8) and (9):

(6) <TW15>: Erm .. ***.do you want*** to write your birthday on a card? I have a pen.
(<BT17> is writing on the card.)
<BT17>: We'll have to send a birthday card.
(7) <TW07>: ***Do you want*** to ask anyone's cell phone number or
(passing a sheet) (Everyone is writing on the sheet.)
<BT09>: I'd just like to say thank you very much.
(8) <BT13>: What ***do you want*** <TW26>?
<TW26>: I want a paper and pen.
(9) <TW10>: So ***do you want*** to see anything in the Temple like ...erm.. what

(BATTICC-F)

The differences between "questions as requests" and "questions as questions" can be seen from the preceding excerpts. According to Tsui (1994), questions simply elicit an obligatory verbal response so that "the interaction between the speaker and the addressee is completed entirely at the verbal level" (p. 80), as in (8) and (9). Requests, however, elicit "an obligatory non-verbal response with perhaps an accompanying verbal response and the interaction is completed at the non-verbal level" (ibid.), which is shown in the excerpts (6) and (7). This can also be applied to the distinction between the "questions as offers" and "questions as questions". In BATTICC-F, since more instances of the use of *do you want* serve the function of offers, the three-word sequence is categorised in this sub-group.

In addition, with respect to multi-word sequences expressing speech acts such as complying, offering, responding to requests and making personal assertions , it can be seen that a number of three-word expressions include the modal verb *would*, such as *would you like, I would love, it would be* and *would be a* . A further look at the users of these sequences indicates that Taiwanese students use relatively few *would* expressions, particularly in CMC. Examples of such use in online exchanges, retrieved from the BATTICC-O, are shown in (10).

(10) Nice to see you. I am <TW28>. ***Can we*** make friends? (from <TW28>)
- yes sure. ***I would love to*** be friends. ***It would be*** really nice. (from <BT30>)
- ***I would very much like*** to be friends with you and ***I*** too ***am hoping to*** make lots of friends through the connecting classrooms experience! (from <BT32>)

(BATTICC-F)

As (10) shows, the Taiwanese learner TW28 expresses his/her desire to maintain the friendship by asking *Can we make friends*? The British participant BT30 then

responds to the request with the lexical phrases *I would love* and *it would be* to demonstrate willingness; similarly, the other British learner BT32 uses *I would very much like to be friends*. Many similar instances can be found in BATTICC-O, indicating the high frequency with which modal *would* is used by British students to respond to requests in online communication. The calculation of the word frequencies of modal verbs indicates that young Taiwanese learners significantly underuse *could* and *would* and overuse *can* and *will* when compared to British English speakers. According to Carter and McCarthy (2006), *could* and *would* are generally perceived as more polite and less forceful than *can* and *will*. In addition, in (10) it can be noted that participant BT32 writes in the progressive, *I am hoping to*, instead of the present, *I hope to*. This seems to indicate a higher level of politeness. Similar use of progressive aspect as a politeness device is found in a number of instances in British students' discourse, as shown in (11)-(14), retrieved from BATTICC-O:

(11) hi ***just wondering*** what kind of electric stuff should i take? like a DS, Ipod, phones stuffs?
(12) I ***was wondering*** about that, are there many ice rinks around Taiwan?
(13) ***I'm hoping*** to make a few friends.
(14) I ***am*** really ***looking forward to*** trying the ones that you reccomend haha :)))

(BATTICC-F)

Biesenbach-Lucas (2007) claims that the past tense (e.g., *could you* instead of *can you*), progressive aspect (e.g., *I was wondering* instead of *I wonder*) and embedding (e.g., *I would appreciate it if you could...*) are three syntactic politeness devices. In this case, the use of politeness devices by young Taiwanese and British learners is slightly different in that Taiwanese learners use relatively fewer past tense and progressive syntactic modification devices to indicate

Table 6.3
Functions of 50 Most Common Three-word Sequences Across Corpora

BATTICC-O	CANELC	BATTICC-F	CANCODE
	SOCIAL INTERACTION:		
	Summoning and greeting		
nice to meet	how are you	nice to meet	
to meet you		to meet you	
		how are you	
	Questioning		
do you have	do you think	do you like	do you think
do you like	what do you	do you have	do you want
what do you		what do you	do you know
		how do you	you want to
		have you ever	what do you
		how is the	have you got
		did you see	
		do you do	
		do you go	
		how about you	
		do you think	
	Offering		
	you want to	do you want	
	if you want	you want to	
	would you like		
	Inviting		
come to Taiwan		come to Taiwan	
	Expressing politeness		
	thanks for the		
	Responding to requests		
	it would be		it would be
	would be a		
	Asserting (personal)		
I have a	looking forward to	I don't know	I don't know
I go to	I have a	I want to	I don't think
I want to	I don't think	we have a	you have to
we have a	be able to	we went to	I think it's
I can play	I don't know	I have a	I think it
I can't wait	I want to	I think it's	I think I
we go to	I'm going to	we have to	you've got to
would like to	you have to	I think I	
I would like	I have to		
I would love	I think I		
	I am not		
	I think it		
	Asserting (impersonal)		
It is very	going to be	it was very	it was a
there are many	it was a	it's very nice	there was a
	this is a		
	there is a		
	it is a		
	this is the		
	Complying		
			yeah you know
			yeah I mean
			yeah I think
	Responding		
			yeah yeah yeah
			no no no

Table 6.3 *Cont.*

BATTICC-O	CANELC	BATTICC-F	CANCODE
	NECESSARY TOPICS:		
		Autobiography	
my name is		my name is	
I am #			
am # years			
My birthday is			
I live in			
birthday is on			
I am not			
I am very			
I come from			
		Time/location	
in your country	in the UK	in the UK	at the moment
	of the year	in your country	the end of
	at the moment		all the time
	the end of		at the end
	of the day		
	the first time		
	at the end		
	out of the		
		Quantity	
a lot of	a lot of	a lot of	a lot of
	a couple of	a lot more	one of the
	a bit of		a bit of
	one of the		a couple of
	some of the		a little bit
	part of the		
	one of my		
	the rest of		
	one of those		
	most of the		
		Likes	
I like to		fish and chips	
my favourite food		what's your	
favourite food is		favourite	
like to play		I don't like	
I also like		I like it	
I love my			
I love to			
pearl milk tea			
		Schools	
junior high school		go to school	
my school is		in your school	
I study in			
go to school			
		Other topics	
with my friends	Happy New Year	Dragon Boat	
Chinese New Year		Festival	
my friend and			
go to bed			

Table 6.3 *Cont.*

BATTICC-O	CANELC	BATTICC-F	CANCODE
		DISCOURSE DEVICES:	
		Linking functions	
and I am	as well as	and it was	but I mean
but I am	but I think	and we have	and it was
and I love	and I have	and we were	and you know
but I don't			and I think
			and I was
		Fluency devices/ Elaboration	
	the fact that	so er erm	I mean I
		you know I	you know I
		I was like	you know and
		it was like	you know what
			you know the
			you know yeah
			you know you
			I mean it's
			you know it's
			what I mean
			I mean you
			mm you know
			that you know
			know what I
		Shifting topics	
by the way		so do you	
		Exemplifiers	
		sort of thing	sort of thing
		sort of like	
		Evaluators	
		to be honest	

politeness. Regarding the embedding of forms, few instances are found in both Taiwanese and British young learners' discourse in this data. This is probably due to the informal nature of online discussion and status-equal communication; embedding may well be commonly used in formal/epistolary settings or when writing to authority figures.

As can be seen in Table 6.3, the majority of three-word sequences used in social interaction are for the expression of assertions, similar to what Biber et al. (2004) call *attitudinal/modality stance bundles*, which express "attitudes toward the actions or event described in the following proposition" (p. 390). Asserting

sequences are divided into personal and impersonal; in interaction, most are overtly personal and express desire (e.g., *I want to*; *I can't wait*), personal opinions (e.g., *I think it*; *I think I*), intention/prediction (e.g., *I'm going to*; *I hope you*), ability (e.g., *I can play*; *be able to*), or obligation (e.g., *you have to*). These sequences are directly attributed to the speaker or writer. However, some sequences of asserting do not explicitly mention the speaker or writer, such as in descriptions of existence (e.g., *there are many*; *there was a*), evaluations of specific things or events (e.g., *it is very*; *it's very nice*), narratives of past events (e.g., *it was a, it was very*), or expressing predictions of future events (e.g., *going to be*).

A comparison of different columns in Table 6.3 illustrates the different uses of three-word sequences in different communication modes. In particular, it can be seen that sequences for asserting are more commonly used in CMC (i.e., BATTICC-O and CANELC), while questioning, complying and responding are considerably more frequent in FTF interaction (i.e., BATTIC-F and CANCODE). These differences can be explained in part by the highly interactive nature of FTF conversation, in which people are consistently asking each other questions, clarifying questions and responding to questions. This, in turn, may facilitate personal relationship building (Belz, 2007). In particular, the number of different questioning three-word sequences in BATTICC-F is extremely high. This suggests that the participants in the intercultural exchange project also demonstrate *skills of discovery and interaction*, namely "the ability to employ a variety of questioning techniques in order to elicit from members of the foreign culture" (Byram, 1997, p. 61), which are some of the most important skills that constitute intercultural competence. Take, for example, the sequence *do you have* in concordances:

How is your country ***in*** er ***the UK***?	***Do you have***	the view... the same view?
Yeah they're really nice.	***Do you have***	any like .. er .. ***local*** sort of like
<TW10>:	***Do you have***	any festival ***in England***?
<BT17>: Mm..	***Do you have***	any beaches ***in Taiwan***?
It's called the Lake District, so ...	***Do you have***	any lakes ***in Taiwan***?
<TW10>: Er	***Do you have***	a Chinese town ***in England***?
<BT21>:	***Do you have***	the same amount ***as we do***?

(BATTICC-F)

As the preceding concordance lines show, the questioning sequence *do you have* often co-occurs with countries, showing that young learners in this project demonstrate a willingness to engage with otherness and a curiosity in discovering different perspectives regarding their own and other cultures. This is the prerequisite attitude of being an intercultural speaker, as has been discussed by a number of scholars (e.g., Belz, 2007; Byram, 1997; Fantini, 2012).

6.3.2 Necessary topics

Another specific function of the use of multi-word sequences is that of introducing or progressing necessary topics (the second section of Table 6.3), that is, topics about which questions are often asked or which are necessary in daily conversation (Nattinger & DeCarrico, 1992). These sequences provide overt signals on specific themes, such as autobiography (e.g., *my name is*), food (e.g., *fish and chips*), time/location (e.g., *in the UK*), school life (e.g., *in your school*), likes (e.g., *what's your favourite*), quantity (e.g., *a lot of*), and some culturally specific topics (e.g., *dragon boat festival*). However, very few high-frequency sequences found in the reference corpora CANELC and CANCODE are grouped in the domains of autobiography, likes/food or school life. For example, the autobiography sequence *my name is* was frequently used by both Taiwanese and British pupils, but only four occurrences of it were identified in the CANELC.

We can also find a number of three-word sequences that are not fixed. These flexible phrases are often called *variable expressions* (Sinclair, 2004) or *phrasal constraints* (Nattinger & DeCarrico, 1992). More precisely, Schmitt (2010) labeled them *formulaic language with open slots*, which "combines a number of words which are frozen, but also allows variety in one or more slots" (p. 132). For example, the slot *I am* # can be filled with various words or phrases, and in this case, it is often completed with a name (e.g., *I am Kim*), with a number referring to age (e.g., *I am 14 years old*) or height (e.g., *I am 5 foot 5 ins*), with an adjective which is used to describe their appearances (e.g., *I am tall*) or emotion (e.g., *I am excited*) and also other different uses. These types of sequence function as "sentence builders", providing the framework for whole sentences (Nattinger & DeCarrico, 1992). As for the other formulaic expression *am # years*, the slot is semantically more restricted, with only a number used to express their age (e.g., *I am 14 years old*). Moreover, some overlaps among the sequences could be found. For example, *my birthday is* and *birthday is on* are likely to both be part of an extended sequence such as *my birthday is on*. As such, the overall frequency figures have to be assessed in the light of this observation, although lists of the use of three-word sequences in each corpus provide useful information about different types of language varieties (Adolphs, 2006).

The multi-word sequences included in the necessary topics are similar to what Biber (2009) calls *referential bundles*, which "identify an entity or single out some particular attribute of an entity as especially important" (p. 285). For example, a number of sequences refer to particular places or locations (e.g., *in the UK*; *in your country*) in BATTICC-O and BATTICC-F, while these are not found with a high frequency in CANELC and CANCODE. This is probably not surprising, as

this project involved learners from different countries and they frequently talked about their own and other cultures during the exchange programme.

With regard to the domain of quantity, most of the sequences, such as *a couple of, a lot of* and *a bit of*, describe amounts or quantities of the subsequent head noun, as in (15)-(18).

(15) <TW07>: Really?
<BT07>: Not all the time – for ***a couple of*** days – and then there's ***a couple of*** months and it's quite warm. (BATTICC-F)
(16) We have ***a lot of*** snow at the moment and I love it! (BATTICC-O)
(17) my worst school memory was when i didnt get a part in the school play in primary school, ***a couple of*** years ago now. (BATTICC-O)
(18) So it was ***a bit of*** a shame that the Waterside had only ordered two casks and they ran out about twenty minutes. (CANELC)

From the extracts above, we see how the participants use vagueness and approximations in both FTF and CMC. This shows that the young learners prefer to describe quantities with vague language to avoid being too precise and pedantic in intercultural exchange. This was claimed by O'Keeffe et al. (2007), especially "in such domains as references to number and quantity, where approximation rather than precision is the norm in conversation" (p. 74). This is sometimes described as "vague additives" (Channell, 1994) or "vague approximators" (Koester, 2007).

Moreover, some of the three-word units in this category precede the adjectives they modify instead of noun phrases, as in (19) to (23):

(19) <BT08>: Here is ***a lot more*** like relaxed.
(20) <BT13>: ...I'm not sure. Just ***a little bit*** difficult. Maybe just that word.

<TW05>: I think the question is boring.
<BT06>: They were ***a bit confusing***.

(BATTICC-F)

(21) she is very lively and happy, she like to do as she is told but sometimes is ***a bit cheeky***. *:P*

(22) ir looks ***a bit strange*** but i looovveee oysters so i think it would be very very nice!!

(BATTICC-O)

What is interesting here is that the sequences *a little bit* and *a bit of* (or *a bit)* occasionally have a more specialised function in our data, frequently being used to downtone an utterance, especially in collocation with negative situations. Such use of vague quantifiers hedges the utterances, which is highly likely to be more appropriate in relation building and considered of more "pragmatic adequacy and integrity" in informal contexts (O'Keeffe et al., 2007, p. 71).

In addition, some three-word units in this category may serve as focus markers, such that new information or the focus of the utterance often follows. Biber et al. (2004) label these *identification/focus bundles*, "focusing on the noun phrase following the bundle as especially important" (p. 394). This can be seen in (24), with *one of my*, and (25), with *one of those*, which identify the food and platforms respectively and are the focus of the utterances.

(23) ***One of my*** favourite food is Tomatoes on sticks. It is very sweet!

(BATTICC-O)

(24) If you have a blog at ***one of those*** platforms, follow us there. If not, just choose one

(CANELC)

It is also worth noting that the time/location and quantity sequences found in CANELC outnumber those in the other three datasets. Some of the sequences, such as *one of my*, *some of the*, *part of the* and *one of those*, do not commonly

occur in FTF spoken discourse. Most of these three-word units incorporate noun phrase and prepositional phrase fragments that have been shown to be one of the typical features of written discourse (Biber, 2009; Carter & McCarthy, 2006; O'Keeffe et al., 2007). The use of three-word units in asynchronous CMC therefore demonstrate a closer approximation to those sequences used in ordinary written discourse as displayed in corpora of native speakers of English.

6.3.3 Discourse devices

The third category of multi-word sequences is discourse devices, which refers to "lexical phrases that connect the meaning and the structure of the discourse" (Nattinger & DeCarrico, 1992, p. 64). As such, they serve an organising function for the flow of information being transmitted, and further improve the fluency of utterances. From Table 6.3, the sequences which serve linking functions, such as *and I am, but I am, and I love, but I don't, and it was* and *and we were,* are common in both online and FTF interaction. Examples of these sequences can be seen in the following excerpts:

(26) Yes there is a lot of snow in England ***but I am*** OK there is not too much around my area which is good I don't like too much snow ***but I like*** some

(27) My brother plays the guitar and is teaching me. he is nine ***and I am*** 13 ***but he is*** so much better than me!!

(BATTICC-O)

(28) <BT17>:... I walked around with Aiden and Katie ***and it was*** very fun.

(29) <BT09>: Yeah, yeah it was straight and like others we walked together. <BT07>: ***And we were*** listening to music and ...

(BATTICC-F)

It is evident that coordinating conjunctions *and* and *but* were frequently used by the participants to express a variety of logical relations between phrases and

sentences in both online and spoken datasets. As shown in Table 6.3, somewhat equal numbers of multi-word sequences that serve linking functions can be found across the four datasets. This result is slightly different from Crossley and Louwerse's (2007) study, which examines two-word sequences (bigrams) and found that the use of coordinating conjunctions collocating with first person pronouns, such as *and I, so I, but I,* and *and we*, is an important feature distinguishing natural dialogues from written discourse. However, their finding can be generated when comparing written discourse and unplanned real-time communication, while in this present study the online discourse also exhibits this feature of unplanned speech.

Nevertheless, the multi-word sequences including coordinating conjunctions found in spoken discourse have a slightly different function to those in the CMC corpora. Many of them are used as a turn-initial resource for speakers, as in (25), and such use is not common in CMC. Evison (2008) defines such units as *flexible instalment openers* because their "lack of specificity means that they can begin an instalment of talk without having to commit to a more complex relationship between upcoming and prior talk from the outset of the turn" (p. 223). As such, the turn is still occupied, and the processing load can further be eased.

The differing use of fluency devices in spoken discourse is also notable between the target and reference corpora. From the data in Table 2 it is apparent that larger numbers of sequences are found in CANCODE as compared with BATTICC, and in particular most of the sequences are centred by the 2-word sequence *you know* or *I mean*, such as *you know I, you know it's* or *I mean I*. The instances of *you know* sequences in BATTICC-F function as interpersonal discourse markers,

marking statements as assumed shared knowledge or experience between speakers and hearers (Fox Tree & Schrock, 2002; Fung & Carter, 2007; Hellermann & Vergun, 2007; House, 2009; Jucker & Smith, 1998; O'Keeffe et al., 2007; Östman, 1981; Schiffrin, 1987). This has been discussed in detail in 4.2.3.5. This discussion demonstrated that *you know* does not simply act as filler or time-buyer; both Taiwanese and British learners use it as a pragmatic marker for interpersonal, attitudinal and organisational purposes, which is broadly consistent with earlier research (e.g., Fung & Carter, 2007; Hellermann & Vergun, 2007; Jucker & Smith, 1998; Schiffrin, 1987).

In the domain of fluency devices, it is also worth considering the sequence *so er erm*, which marks speaker hesitation. As has been discussed in 4.1.2, hesitation markers feature frequently in spontaneous conversations and fulfil an important pragmatic function, and they are pervasive in that *er* and *erm* are ranked sixteenth and twentieth in the most frequent words in BATTICC-F. Other three-word sequences that serve a similar function and occur at least three times include *I er I, er er I, I I I, er I er, I like erm* and *er I think*, which all contain hesitation items and/or repeats. What needs to be emphasised is that these sequences are mainly found in Taiwanese learners' speech; although *er* and *erm* are used slightly more frequently by the British participants, there are very few three-word multi-word sequences containing these two items in their top 50 sequences and the first sequence of such type is *erm I think* (rank 77). This also accords with De Cock's (2004) findings, which indicate that EFL learners use significantly more multi-word sequences that contain repeats and/or hesitation items than native speakers of English. In her analysis, 12 out of the top 20 high-frequency sequences are of this type, and the total numbers of hesitation or repeat sequences in her learner corpus are approximately three to four times larger than those found

in native speakers' discourse.

With regard to the exemplifiers in discourse devices, the sequence *sort of thing* appears in the top 50 three-word units in both BATTICC-F and CANCODE. Such an expression is often referred to as vague language in this thesis. As was discussed in 4.2.2.4, one of the primary functions of being vague is to "indicate assumed or shared knowledge and mark in-group membership" (O'Keeffe et al., 2007, p. 177). In this way it is not necessary for speakers/writers to convey precise and concrete information, and the hearers/readers in most instances know what a vague expression refers to. A number of multi-word sequences that serve a similar function include *something like that* and *that sort of*, which can be found in both online and spoken discourse. Such forms of vague exemplifier have been found to be particularly distinctive features in adolescent speech and informal online messages in terms of use and frequency as compared with adult talk (see Martínez, 2011; Tagliamonte & Denis, 2010). Examples of these three-word units in the context have been illustrated in 4.2.2.3.

6.3.4 Distribution of common three-word sequences across functional types

The previous sections have demonstrated that three-word units are often tied to particular conditions of use, and they can be identified according to Nattinger and DeCarrico's (1992) three functional categories: social interaction, necessary topics and discourse devices. Nevertheless, it can be clearly seen that the use of multi-word sequences in different communication modes differs in relation to the functional types. Table 6.4 and Figure 6.2 present the distribution of functions served by three-word sequences across corpora.

The recurrent sequences for social interaction are extremely common across each dataset, ranging from 38% to 54% of the top 50 high-frequency multi-word sequences in each corpus. On the other hand, a more noticeable distribution difference of multi-word sequences across corpora can be seen in the percentage of necessary topics and discourse devices, which range from 18% to 54% and from 8% to 36% respectively. Necessary topics are overall particularly common in CMC, as presented in BATTICC-O (54%) and CANELC (38%), and this is strikingly higher than the percentage figures of the spoken data presented in BATTICC-F (24%) and CANCODE (18%). The use of multi-word sequences as discourse devices demonstrates the opposite pattern in that a generally higher rate can be found in spoken discourse compared to CMC. As shown in Table 6.4, the sequences of discourse devices can be found in only four instances (8%) out of the first 50 high-frequency multi-word sequences in both BATTICC-O and CANELC, while the percentage figures of BATTICC-F and CANCODE reached 22% and 42% respectively.

Table 6.4

Distribution of Common multi-word sequences across Corpora

	BATTICC-O		BATTICC-F		CANELC		CANCODE	
Social Interaction	18	36%	27	54%	27	54%	23	46%
Necessary Topics	27	54%	12	24%	19	38%	9	18%
Discourse Devices	5	10%	11	22%	4	8%	18	36%

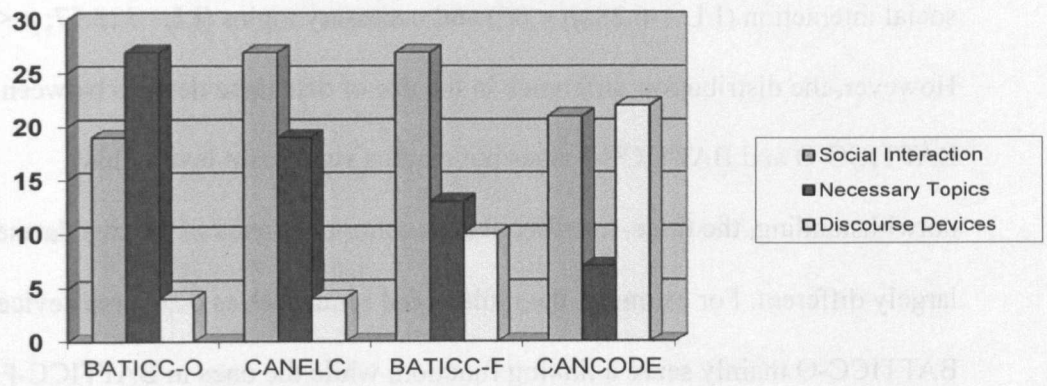


Figure 6.2

Distribution of Top 50 Three-word Sequences across Functional Types

Table 6.5

Accumulative Frequencies of Three-word Sequences and the Statistical Test of Significance

	BATTICC-O	BATTICC-F	CANELC	CANCODE	Significance
Social Interaction	361	278			$p<.05$
			1237	31227	$p<.001$
Necessary Topics	838	118			$p<.001$
			1178	10782	$p<.01$
Discourse Devices	92	78			
			284	18274	$p<.001$

Table 6.5 shows statistically significant differences in the use of three-word sequences with different functions among the corpora using log-likelihood (LL) ratio (Rayson, 2008) based on the accumulative frequencies of sequences. The table indicates significant differences between CANELC and CANCODE in three functional categories: social interaction (LL= -1374.98; p <.001), necessary topics (LL= 8.13; p <.01) and discourse devices (LL= -1907.71; p <.001). The negative values indicate a significantly higher rate of three-word sequences as social interaction and discourse devices in CANCODE. The difference in distribution of

functional categories between BATTICC-O and BATTICC-F is significant in social interaction (LL= -6.28; p <.05) and necessary topics (LL= 328.57; p <.001). However, the distribution difference in the use of discourse devices between BATTICC-O and BATTICC-F does not reach a significant level. This notwithstanding, the three-word sequences commonly used in the two datasets are largely different. For example, the multi-word sequences as discourse devices in BATTICC-O mainly serve a linking function, while the ones in BATTICC-F include four different functional types (see Table 6.2).

The highly frequent use of multi-word sequences for social interaction in BATTICC-F is likely due to the phatic nature of FTF communication in that young learners focused more on social interaction than specific information when they met face-to-face. This may also be because of the fact that the multi-word expressions in CMC are less interactional in nature. Concerning necessary topics, the significantly higher rate in BATTICC-O might be due to the fact that the patterns of language use on electronic discussion boards reflect the particular topics that the participants are interested in, while in FTF interaction, topics are more easily adapted to the immediate environment. The discourse is also more likely to be oriented to topics that can be referred to pronominally (e.g., *it, this, that*) (higher frequencies of these proximal deictic forms can be found in BATTICC-F), while written forms tend to require full lexicalisation and hence show more topic-instantiating multi-word sequences. In addition, the notably greater discourse devices used in BATTICC-F compared with BATTICC-O can possibly be attributed to differences in the nature of spoken and written modes. Nattinger and DeCarrico (1992) explain that:

> Writers are removed from their audience in a way that speakers are not from theirs. Speakers and hearers work jointly, in a rather spontaneous, unplanned manner, to establish meaning inside the immediate context in which the interaction takes place. They can thus rely on shared signals ... to regulate the speed and content of the message. (p. 83)

It emerges that participants in online interaction based on written form may not have such proximate relationships with each other since the discourse is more explicit, with less recurrent discourse devices in the mediation of online discussion. Fewer high-frequency discourse marking multi-word sequences may be due to the fact that the foregoing discourse is preserved in online communication, rather than real-time speech, which needs to be more explicitly organised.

6.3.5 Brief Summary: Discourse functions of multi-word sequences

This section has explored the discourse and pragmatic functions of three-word sequences in intercultural CMC and FTF communication (BATTICC-O and BATTICC-F), as well as two large reference corpora (CANELC and CANCODE). The evidence of the study presents that the high-frequency three-word sequences in the four datasets serve three central discourse functions: social interaction, necessary topics and discourse devices. These findings add to a growing body of literature on the functional use of multi-word expressions, which has been shown in a range of previous studies (e.g., Biber, 2009; Biber et al., 2004; Nattinger & DeCarrico, 1992; Schmitt & Carter, 2004; Wray & Perkins, 2000). It further shows that three-word sequences often perform systematic discourse functions, even though they do not usually constitute complete grammatical or idiomatic

structures. They function as "important building blocks in discourse" (Biber, 2009, p.284), and accord with interlocutors' expectations and preferences, which may facilitate efficient and effective communication for different communicative purposes (Schmitt & Carter, 2004; Wood, 2010).

It is also apparent that three-word sequences employed in CMC and FTF conversation are significantly different. In the category of social interaction, questioning, complying and responding were generally more frequently used in FTF communication, while sequences used for making assertions in both personal and impersonal contexts were found in more instances in CMC. In addition, the three-word units employed in the area of necessary topics were particularly common in CMC, reflecting topics such as autobiography, time/location, likes/interests, quantity and schools, while these sequences were not found with a high frequency in FTF talk. With regard to the three-word sequences functioning as discourse devices, a large number of highly recurrent sequences in BATTICC-F and CANCODE were not commonly used in online discussion. Some examples include fluency devices (e.g., *you know I, I mean I, it was like*), exemplifiers (e.g., *sort of thing*) and evaluators (e.g., *to be honest*). Nevertheless, the three-word sequences that serve linking functions (e.g., *and I am, and it was, but I don't, and we have*) were very common in both CMC and FTF interaction.

The analysis of the use of three-word sequences by different groups of participants also reveals a number of differences between British and Taiwanese participants' discourse. For example, some sequences that were frequently used by British participants can only be found in a very few instances in Taiwanese learners' discourse, such as sequences serving linking functions (e.g. *and I love,*

but I think), expressions with *would* for responding to requests (e.g. *it would be, would love to*), vague exemplifiers (e.g. *sort of like, sort of thing, things like that*), vague quantifiers (e.g. *a couple of*) and hedges for downtoning their utterances (e.g. *a bit of, a little bit*). These findings highlight the need for teachers and materials developers to incorporate multi-word sequences commonly used by native English speakers in learning materials and EFL instruction. Learners will thus be exposed to appropriate expressions in different communicative situations. As Schmitt and Carter (2004) claim, multi-word sequences are "not only helpful for efficient language usage; they are essential for appropriate language use" (p.10). The following sections will examine the extent to which intercultural exposure to native English speakers affects the use of three-word sequences by young Taiwanese learners over one year of contact.

6.4 Development of three-word sequences

As has been discussed in Chapter 2, a burgeoning field of research has looked at language development in intercultural settings, particularly in the development of vocabulary and lexical richness, but there seems to be a paucity in the research area of the development of multi-word patterns of language use, which play a prominent role in language learning and language use (e.g., Biber, 2009; Ellis, Simpson-Vlach, & Maynard, 2008; Schauer & Adolphs, 2006; Schmitt 2010, 2013; Qi & Ding, 2011; Wray, 2002, 2013). As a result, assessing the development of multi-word sequences provides another way of assessing the success of online intercultural contact as a language learning method. Adolphs and Durow (2004) conducted a longitudinal case study of two EFL postgraduates' improvement in their use of three-word sequences in spoken English, showing

that the level of social integration into the native speaker community has a positive impact on the acquisition of formulaic sequences in language use. Li and Schmitt (2010) investigated collocation use in academic writing in a longitudinal learner corpus to identify the learners' improvement over the course of one academic year. Qi and Ding (2011) analysed the use of multi-word sequences by 56 Chinese university English majors in their prepared monologues at the beginning and end of a three-year period and compared the student performance with that of 15 American college students. The most challenging area for them to tackle was the use of multi-word sequences containing prepositions and articles.

Although a number of studies have demonstrated developmental changes in the use of multi-word sequences using a longitudinal approach, the majority has focused on advanced, mature EFL/ESL learners; that is, young learners of English at a beginner-intermediate level do not seem to be a group upon whom focus has been concentrated. This section therefore attempts to answer the following central research question:

> To what extent does intercultural contact with native English speakers facilitate the development of multi-word sequences by young Taiwanese learners in a one-year online exchange?

Three investigations were carried out with a view to examining the extent to which intercultural online exposure to native English speakers affects the use of multi-word sequences by young Taiwanese learners over one year of contact. Li and Schmitt (2010) note that conducting longitudinal studies of the same learners over time is the only truly reliable way to identify patterns of development in the

use of formulaic language by L2 learners (p. 25). This study therefore attempts to do so from three investigations. The first two studies (6.4.1 and 6.4.2) analyse the sequences frequently used by Taiwanese participants during the online interaction in order to examine their approximation to those sequences used by native English-speaking participants. Such evaluation of the "proximity to or distance from real-world discourse" can provide teachers with insights to better assess their own learners' performance and lead to better classroom task design (McCarthy et al., 2010, p. 67). The third study (6.4.3) employs an experimental approach, a pre-test–post-test control group design, to measure the development of multi-word sequences by Taiwanese participants. This will provide empirical evidence to support the first two investigations of the development of multi-word sequences.

6.4.1 Three-word sequences surrounding frequently used lexical items

Corpus studies have shown that highly recurrent multi-word sequences usually incorporate high-frequency lexical items (Adolphs & Durow, 2004; Biber 2009; Schmitt 2010). As such, in this case the most frequent words in each phase of interaction may not only simply elucidate how they are employed over time, but also exhibit the frequent use of multi-word patterns. The first investigation pays particular attention to the most frequent lexical items and their surrounding three-word sequences in the three phases of the intercultural exchange programme.

6.4.1.1 Most frequent items over time

This analysis was begun by identifying the most frequent words from the Taiwanese learner dataset during the different phases using *WordSmith Tools 5.0* (Scott, 2008), as shown in Table 6.6. This initial frequency information provided

an immediate snapshot of how lexical items were employed and the development changes over time, thus it is a good starting point for subsequent analysis.

Table 6.6

Most frequent items over time in the Taiwanese learner dataset

	Phase One		Phase Two		Phase Three	
1	I	302	I	289	I	267
2	IS	202	IS	184	IS	202
3	MY	185	TO	155	THE	190
4	TO	139	MY	154	MY	157
5	THE	109	THE	144	TO	148
6	IN	97	AND	109	AND	144
7	YOU	93	IN	97	IN	113
8	LIKE	92	A	91	A	96
9	A	90	YOU	88	IT	87
10	AND	84	LIKE	87	LIKE	83

From the table it is apparent that most of the high-frequency items are extremely similar in different phases although the order is slightly different. What needs to be stressed is the use of *and* over the course of the programme's three phases, with occurrences (and percentages) of 84, 109 and 144 respectively. In this case, the increase in use of the item *and* by Taiwanese learners over time is marked. As a result, the three-word sequences that include the coordinating conjunctions *and* (e.g., *and I am*) or *but* (e.g., *but I don't*) will be further examined in the following section.

6.4.1.2 Overlap of three-word sequences with coordinating conjunctions

The total numbers and percentages of three-word sequences including *and/but* used over the three phases of the programme are given in Table 3. In Phase One, 17 instances of such sequences out of the total 286 items (5.94%) were identified, and the percentages increase to 6.66% and 9.67% in Phases Two and Three

respectively. Log-likelihood (LL) ratio analysis also indicates a significant difference (LL= 5.48; p <.05) in the frequency of use between Phases One and Three. Such a significant rise from the first to the third phase suggests that Taiwanese learners tended to use more three-word sequences with linking functions over time in the three phases. This can probably be expected since the raw frequency count has revealed the increasing use of *and* (see Table 6.7), yet the differences in the use of such three-word sequences over time by British pupils were not marked in that approximately 15-16% of the sequences were found incorporating *and* or *but* in different phases. In the comparison of these figures, the developmental trend in the use of *and/but* sequences by the Taiwanese learners appears to indicate an increasingly close approximation to the use by their British peers.

Table 6.7
The Three-word Sequences Including and/but and the Overlap between the Use of Taiwanese (TW) and British (BT) Participants

	Phase One			Phase Two			Phase Three		
	TW	BT	overlap	TW	BT	overlap	TW	BT	overlap
and/but sequences	17	66	7	19	52	9	28	54	13
Percentage	5.94%	16.1%		6.66%	15.7%		9.67%	15.4%	

As shown in Table 6.7, the numbers of the overlapping three-word sequences (those used by both British and Taiwanese participants) that included *and* or *but* in Taiwanese and British datasets are seven and nine in the first two phases respectively and are then followed by a substantial increase to 13 occurrences in Phase Three. This ongoing rise shows that the Taiwanese learners increasingly used those three-word units with *and/but* that were also frequently used by the British participants. It was clearly shown that the Taiwanese learners used an

increasing and more varied number of sequences including *and* or *but* over the three phases, and this further shows an increasingly higher level of approximation to the use of three-word sequences by native speakers of English.

Table 6.8

Overlap of Three-word Units with Coordinating Conjunctions

	three-word sequences	British texts Total freq.	Taiwanese texts Phase One	Phase Two	Phase Three
1	AND I AM	20	2	2	4
2	BUT I AM	11	0	0	3
3	AND I GO	7	0	1	3
4	BUT I DON'T	7	2	2	1
5	AND I LIKE	6	1	2	2
6	AND I HAVE	6	1	0	2
7	AND HAVE A	5	0	3	0
8	BUT I THINK	5	0	0	4
9	AND MY FAVOURITE	4	2	1	1
10	AND WE HAVE	4	0	2	2
	Total	79	8	12	22

Table 6.8 lists the overlap of the three-word sequences with coordinating conjunctions by the two groups of participants, ranked by their frequencies in the British participants' discourse. It is apparent that all of the high-frequency items were combined with first person pronouns (i.e., *I* and *we*). In this way the participants frequently used coordinating conjunctions for the linking of their utterances in expressing personal opinions, experiences and desires. From Table 6.8 a comparison of the amount of use of multi-word sequences based on *and/but* by Taiwanese learners over time indicates a progressive development. The first sequence *and I am*, for example, occurs twice in the first two phases and doubles in Phase Three; the frequency of the sequence *but I think* (rank 8) rises from 0 to 4 over the three-phase programme. With regard to the total numbers of frequencies of the top 10 sequences, a slight increase can be seen from the Taiwanese learners'

use of the *and/or* sequences from Phase One to Two, followed by a sharp increase by the end of the programme.

6.4.2 Analysis of key sequences

The second investigation of evaluating the development of multi-word sequences employs a corpus linguistic approach using *WordSmith Tools 5.0* (Scott, 2008) to identify the *key sequences*, which refer to those sequences that occur unusually frequently or unusually infrequently in a text as compared with some kind of reference corpus (see O'Keeffe et al., 2007; Scott, 2010). The procedure works by comparing the actual observed frequency of each multi-word sequence in the target corpus with its equivalent in the reference corpus. For this study, the Taiwanese and British participant datasets were set as the target and reference corpus respectively, working on the basis of the relative frequencies of each three-word unit in the two datasets and further identifying the significant underuse and overuse of multi-word sequences.

6.4.2.1 Key three-word sequences

Using *WordSmith Tools* to analyse the key sequences, working on the frequency wordlists in the Taiwanese and the British datasets, for $p < .01$ with 1 d.f., the cut-off of 6.63 revealed 87 sequences that are either significantly overused or underused by Taiwanese learners in BATTICC-O data. The top 10 three-word sequences (with the largest LL values) that were overused by Taiwanese students, all of which were much less frequently used by the British participants in this analysis, are presented in Table 6.9.

Table 6.9

Key Three-word Sequence List of the Taiwanese Discourse (BATTICC-O)

	Key sequence	Taiwanese texts		British texts		Keyness
		Freq.	%	Freq.	%	(LL)
1	JUNIOR HIGH SCHOOL	35	0.23	0	0	52.0
2	I STUDY IN	21	0.14	0	0	31.2
3	CHINESE NEW YEAR	20	0.13	0	0	29.7
4	BIRTHDAY IS ON	19	0.12	0	0	28.2
5	A LOT OF	43	0.28	11	0.06	23.5
6	I LIKE TO	66	0.43	25	0.15	23.3
7	HIGH SCHOOL IN	15	0.1	0	0	22.3
8	THERE ARE MANY	14	0.09	0	0	20.8
9	I LOVE MY	13	0.08	0	0	19.3
10	PEARL MILK TEA	12	0.08	0	0	17.8
11	MY BIRTHDAY IS	23	0.15	4	0.02	16.7
12	MY SCHOOL IS	23	0.15	4	0.02	16.7
13	IN MY FREE	11	0.07	0	0	16.3
14	MY FREE TIME	11	0.07	0	0	16.3
15	DRAGON BOAT FESTIVAL	11	0.07	0	0	16.3

It can be easily seen that some of the highly recurrent sequences distinctively reflect Taiwanese learners' social and cultural customs and practices, such as *Chinese New Year*, *Pearl milk tea* (milk tea containing small chewy balls made of tapioca starch, called "Pearls" in Chinese) and *Dragon Boat Festival* (a traditional holiday celebrated on the fifth day of the fifth lunar month on which people race dragon-like boats). The full meaning of these sequences may not be easily understood by people not based in Taiwanese, Chinese or East Asian societies. Therefore, this cultural discourse would be helpful for enabling the British students to know more about Taiwanese culture through online intercultural exchange.

Table 6.10, on the other hand, illustrates the three-word units that were commonly used by the British participants, and most of these were not used by the Taiwanese learners. In this list it can be noted that the three-word sequences based on the

coordinating conjunction *and*, such as *and I love*, *and go to* and *me and my* occurred in the native speakers' discourse frequently, while no instances of these three sequences can be found in the Taiwanese pupils discourse. This seems to indicate that Taiwanese learners significantly underused the multi-word sequences of this type, as compared with their British counterparts. This notwithstanding, a number of three-word sequences including *and* can still be found in Taiwanese learners' discourse, such as *and I am, and I like, and I have* and *and I go*, which, however, occur to a lesser degree than in the discourse of British participants. Also, the sequences with the modal *would*, such as *I would love*, *would love to* and *it would be,* commonly used by natives are not found in the Taiwanese learners' texts either. These findings further support the previous study on key parts-of-speech analysis, which shows that the multi-word sequences with modal verb *would* and coordinating conjunctions are the two most underused grammatical categories by the Taiwanese participants as compared with the use by the British learners.

Table 6.10

Key Three-word Sequence List of the British Discourse (BATTICC-O)

	Key sequence	British texts		Taiwanese texts		Keyness
		Freq.	%	Freq.	%	(LL)
1	HI MY NAME	15	0.09	0	0	19.4
2	WHEN I GET	12	0.07	0	0	15.5
3	AND I LOVE	12	0.06	0	0	15.5
4	WITH MY FRIENDS	26	0.15	5	0.03	13.7
5	AND GO TO	10	0.06	0	0	12.9
6	I GET HOME	10	0.06	0	0	12.9
7	I WOULD LOVE	10	0.06	0	0	12.9
8	WE GO TO	10	0.06	0	0	12.9
9	WOULD LOVE TO	10	0.06	0	0	12.9
10	ME AND MY	9	0.05	0	0	11.6
11	ON THE COMPUTER	9	0.05	0	0	11.6
12	SOMETHING TO EAT	9	0.05	0	0	11.6
13	TIME WITH MY	9	0.05	0	0	11.6
14	IT WOULD BE	9	0.05	0	0	11.6
15	NOT VERY GOOD	9	0.05	0	0	11.6

Table 6.11

Key Three-word Sequence List of the British Discourse (BATTICC-F)

	Key sequence	British texts Freq.	%	Taiwanese texts Freq.	%	Keyness (LL)
1	IT WAS VERY	9	0.08	0	0	6.8
2	I THINK IT'S	7	0.07	0	0	5.3
3	A LOT MORE	6	0.06	0	0	4.5
4	AND IT WAS	6	0.06	0	0	4.5
5	AND WE WERE	6	0.06	0	0	4.5
6	DO YOU DO	6	0.05	0	0	4.5
7	IT'S VERY NICE	6	0.05	0	0	4.5
8	TO DO IT	6	0.05	0	0	4.5

Applying the keyness method to the analysis of BATTICC-F, for $p < .05$, at 1 d.f., the cut-off of 3.83 generated 21 overused and 8 underused sequences that reach a significant level by Taiwanese learners in face-to-face interaction. Table 6.11 lists the three-word sequences that are frequently used by the British pupils but which can hardly be found in Taiwanese learners' discourse. From the table it can be seen that the top five key sequences with a high LL value contain two sequences comprising the coordinating conjunction *and*, namely *and it was* (rank 4) and *and we were* (rank 5), which were significantly underused by Taiwanese learners (LL > 3.38, $p < .05$), while these can be found with a frequency of six in the British participants' data. This further confirms the earlier results that the coordinating conjunction *and* was considerably underused by Taiwanese learners.

In addition, the analysis of two-word sequences (bigrams) provides us with some interesting insights into the types of sequences used by learners and native speakers, offering access to both paradigmatic and syntagmatic features of language (Crossley & Salsbury, 2011). In this case, the analysis reveals the significantly different use of two-word sequences by British and Taiwanese pupils. Table 6.12 presents the top 15 two-word sequences (with the largest LL values)

that were significantly underused by Taiwanese learners, as compared with British participants (LL > 6.64, p <.01). The most striking result to emerge from the data is that the Taiwanese participants tended to underuse two primary categories: vague language (e.g., *sort of, a bit, and stuff, over here*) and the past-tense units (e.g., *it was, I was, we were, we had, I didn't, that was, was very, was like*). The underuse of past-tense forms in speaking by Taiwanese learners also accords with the earlier observations in the key POS analysis, which has been discussed in 4.7.1.

Table 6.12
Key Two-word Sequence List of the British Discourse (BATTICC-F)

	Key sequence	British texts		Taiwanese texts		Keyness
		Freq.	%	Freq.	%	(LL)
1	IT WAS	48	0.45	0	0	36.4
2	SORT OF	23	0.21	0	0	17.4
3	I WAS	21	0.2	0	0	15.9
4	YOU DO	14	0.13	0	0	10.6
5	I'VE GOT	13	0.12	0	0	9.8
6	WE WERE	13	0.12	0	0	9.8
7	WE HAD	13	0.12	0	0	9.8
8	I DIDN'T	13	0.12	0	0	9.8
9	AND STUFF	12	0.11	0	0	9.1
10	VERY NICE	12	0.11	0	0	9.1
11	A BIT	11	0.1	0	0	8.3
12	THAT WAS	11	0.1	0	0	8.3
13	OVER HERE	11	0.1	0	0	8.3
14	WAS VERY	11	0.1	0	0	8.3
15	WAS LIKE	11	0.1	0	0	8.3

With regard to the use of vague sequences, *sort of* (rank 2), *and stuff* (rank 9), *a bit* (rank 11) and *over here* (rank 13) were frequently used by British participants, whereas the Taiwanese learners did not use these at all. Such results corroborate the findings of a great deal of the previous studies in this field (e.g., Channell, 1994; De Cock, 2004; Ellis et al., 2008; Fernandez & Yuldashev, 2011; O'Keeffe

et al., 2007), which show that a number of vague multi-word sequences pervasively used in native-speaker spoken discourse are sometimes significantly underused by EFL learners. In this case, Taiwanese learners are less likely to use vague expressions to convey the indirectness that is helpful in enhancing interpersonal relationship. As discussed in section 5.5, vagueness or lack of precision is an indicator of intersubjectivity, which is highly likely to be more accepted and preferred in informal interaction. Without using it, the discourse may sound rather bookish and pedantic even it is grammatically and lexically correct.

Since striking differences were found in the use of multi-word sequences between Taiwanese and British participants, this is likely to make the Taiwanese discourse seem unnatural, and consequently it is desirable for EFL learners to acquire the specific sequences that their English-speaking interlocutors expect and prefer (Schmitt & Carter, 2004). As Wray and Perkins (2000) point out, within the group, "formulaic language is better suited to this than novel language is, because a hearer is more likely to understand a message if it is in a form he/she has heard before, and which he/she can process without recourse to full analytic decoding" (p. 18). Despite the difference in use, multi-word sequences did not lead to misunderstandings or conflicts in both online and face-to-face interaction. As House (2009) suggests, learner corpus research can be carried out in order to improve the learners' communicative competence, and, as a result, differences in language use between native and non-native speakers can be a stepping-stone to further analysis and applications. These research findings delineate the pedagogical merit of key sequence analysis and thus help to inform teachers and materials writers in relation to course design for EFL learners.

6.4.2.2 The decline of the key sequences

In the previous section we have seen the key sequences that occurred with an unusual frequency in each dataset as compared with the other as a baseline. We now consider the quantity of key sequences identified in Taiwanese data as compared with the British data over time. Table 6.13 presents the number and percentage of three-word key sequences that were identified based on each phase of Taiwanese discourse in comparison with the sequences used by British participants. For $p < .01$ with 1 d.f., 6.63 as the cut-off of LL value, 31 items were identified as *key* out of the total 112 sequences for Phase One (27.67%). This figure then decreases to 26 (27.65%) and 18 (19.78%) in the following two phases respectively. This continuous decline in the number and percentage of key sequences indicates that the Taiwanese participants gradually use fewer three-word units that were identified as statistically significant in their overuse or underuse relative to the British participants.

Table 6.13
Number and Percentage of Key Sequences in Taiwanese Discourse

	Phase One	Phase Two	Phase Three
Number of three-word *key* sequences	31	26	18
Percentage of three-word *key* sequences	27.67%	27.65%	19.78%
Overlap in the 50 most common three-word sequences	8	9	14

Table 6.13 also shows the numbers of three-word units that overlap between the British and Taiwanese participants' discourse in terms of the 50 most common three-word units retrieved from different phases of discourse. Eight and nine out of 50 highly recurrent sequences in Taiwanese learners' discourse overlapped in

Phases One and Two respectively, and this goes up to 14 by the end of the programme. The increase of overlap in the high-frequency sequences also indicates an increasing convergence between the use of sequences by Taiwanese and British native speakers of English. This may be due to the fact that Taiwanese learners over time were simply imitating native speakers, in which case the user may have control of part of the holistic or the componential meaning (Wray, 2000). A clear example is shown in the following excerpts, and a number of instances were found with highly similar uses of multi-word sequences in Taiwanese and British students' texts.

> Hi *my name is* <BT23> but my friends call me XXX. :D
> *I am 13 years old* and *I live in* Workington and *I like to* watch telly and play out with my friends. *My favourite food is* chocolate. <BT15> and <BT22> are my best friends!
>
> (from a British participant)

> Hi *my name is* <TW36>. *I am 13 years old. I live in* Hualien.
> I study in <SN04> junior high school. *I like to* play volleyball and basketball. *My favourite food is* pizza and fried chicken.
>
> (from a Taiwanese participant)

In the two instances above it can be seen that six three-word sequences are the same: *my name is, I am #, # years old, I live in, I like to, my favourite food* and *favourite food is*. These seem to be extreme examples, but they appear to indicate that Taiwanese participants' choice of language tended to be consciously or unconsciously affected by their British interlocuters. Crystal (2006) also observes that in the online community, although the members come from different backgrounds and write in different styles, they tend to accommodate each other, and as such their contributions progressively develop a shared linguistic style.

From the analysis of recurrent three-word units by the two groups of participants, the increase of overlap in the high-frequency sequences and the decline of key sequences show an increasing convergence between the use of three-word sequences by Taiwanese and British native speakers of English. Although this approach does not measure the appropriateness or correctness of three-word units used by Taiwanese participants, it indicates a level of approximation to the use of sequences by English native speakers. Finally, the third investigation of the multi-word sequences employs an experimental approach, a pre-test–post-test control group design, to measure the development of three-word units by the Taiwanese participants to further support the first two investigations based on a discourse perspective.

6.4.3 Development of single- and multi-word knowledge

6.4.3.1 Experimental approach

To determine whether the intercultural contact facilitated the development of Taiwanese learners' lexical proficiency, two vocabulary tests were used, focusing on both single-word and multi-word units. The test of single-word knowledge is based on the *Word Reading Test* (see Appendix D) designed by Hung et al. (2006), which contains a 100-word meaning proficiency test assessing the number of real printed words that could be accurately identified by the Taiwanese participants. The development of this testing instrument proceeded in several stages and established satisfactory reliability and validity, as well as a reasonable item discrimination index (0.39~1.0) and level of difficulty (0.17~0.79), which was selected after a pilot research and *ITEMAN*[9] analysis. The score of the test was

[9] *ITEMAN* is a software program designed to provide detailed item and test analysis reports using classical test theory (CTT). The program produces summary output regarding the examinee scores, including reliability analysis, analysis of domains (content areas) and frequency distributions.

the number of words from 0-100 that were answered correctly.

Furthermore, I designed a test for three-word sequences to assess recognition knowledge of the form-meaning connection of multi-word sequences (see Appendix E). In this test participants were simply asked to write down the equivalent meaning in their first language, Taiwanese Mandarin. In this way, such a translation assessment tests receptive knowledge because "the learners move from the given multiword unit to meaning" (Nation & Webb, 2011, p. 189). Both the tests for single- and multi-word knowledge adapted the concept of L1 translations since it is a more effective way of conveying word meaning than L2 definitions (ibid.), particularly for language learners at a preliminary and intermediate level, who may not have a large L2 vocabulary.

To develop the test for multi-word sequences, one important issue is to select the target sequences to be tested. I used the frequency lists as the criterion for selection since they avoid sampling issues and can be the best way of deciding which multi-word sequences to include (Nation & Webb, 2011; Read, 2007; Schmitt, 2010). However, in drawing a sample of items for vocabulary tests, Read (2007) notes that "there is no definitive word frequency test, either for English generally or particular uses of English" (p. 109). For the present study I employed the wordlists generated from general e-language corpus CANELC, which consists of language used online; the high-frequency items retrieved from this corpus are therefore more likely to occur during the online exchange programme.

Furthermore, since the corpus tool normally simply produces all the three-word continuous units, careful manual selection was needed. In this case, the 80 most

common three-word items were first selected from the frequency list. However, a number of sequences were excluded. One such example was the items that mainly consist of grammatical units, such as *to be a, to be the* and *to have a,* etc. Moreover, some sequences are likely to be part of an extended sequence. For example, three-word sequences like *would be a* and *it would be* seem to be part of a four-word sequence *it would be a.* Some sequences are also quite similar, such as *am going to* and *I'm going to*, and such duplication may lead to inaccurate estimates. As a result, in the latter two cases only one three-word sequence was chosen to include in the test. Forty items were finally selected in the test.

In addition, in the multi-word sequence test, each sequence was selected with its context, namely concordance lines in the original utterance in the CANELC. Some examples are illustrated in the following:

- I think I already did *a couple of* weeks ago! Haha
- I'm not sure what's going on with it *at the moment*.
- I'm *looking forward to* seeing everyone.

Although whether words should be assessed in context is still debatable (see Read, 2007), in this study the original context of each three-word sequence was given owing to the structurally incomplete nature of multi-word sequences. Regarding the selection of concordance lines for the test, my criterion was to include the easier items because on the one hand, this would not scare the participants and, on the other hand, the Taiwanese learners do not have large L2 vocabulary since they have only been learning English for approximately four to five years. I therefore controlled for the word difficulty by using the online version of Laufer and Nation's (1995) *Lexical Frequency Profile*, which can be found at the *Compleat*

Lexical Tutor website under the heading *Vocabprofile*[10] (http://www.lextutor.ca/). In this way, it could be ensured that most of the items (94%) included in the test were from the first 1,000 words, which most of the participants should be familiar with.

In this study, a pre-test–post-test control group design was carried out by setting two groups of participants, namely the treatment or experimental group that "receives treatment or which is exposed to some special conditions" and the control group, whose role is to "provide a baseline for comparison" (Dörnyei, 2007, p. 116). In the present study, the experimental group (N=35) was involved in a one-year intercultural exchange programme, while the control group (N=35) was not. Both the groups still received classroom-based English instruction. To enable the two groups to be more comparable, as has been suggested, researchers have to try to make the control group as similar to the treatment group as possible (Dörnyei, 2007; Mujis, 2010). In the present study, the participants for the control group were carefully chosen. Firstly, the two groups of learners were all Taiwanese of similar age, that is, they were all 7th graders at the time of the start of the programme. In addition, they received similar English teaching based on the same EFL textbooks during the one-year project. With regard to intercultural experiences, all of the participants from both groups had relatively little experience of interacting with people from English-speaking countries.

To measure the progress of lexical proficiency, pre- and post-tests of both single- and multi-word knowledge were given before and after the intercultural exchange

[10] *VocabProfile* is a web-based program that performs lexical text analysis. It takes any text and divides its words into different categories by frequency, which helps to measure the proportions of low and high frequency vocabulary in a written text.

programme. Data collected from the Taiwanese students' performance on the measures of the tests were quantified with descriptive and inferential statistics by employing the Statistical Package for the Social Sciences (SPSS) version 19.0 for Windows. In particular, an independent-samples t-test was carried out to compare the scores for the experimental and control groups in order to investigate whether participants who have regular intercultural contact with native speakers can improve their word knowledge more than students who simply received classroom instruction and studied by themselves. In addition, the effect size was further computed as it indicates "the magnitude of an observed finding" (Rosenfeld & Penrod, 2011, p. 342). This is also able to tell us whether the difference or relationship we have found is strong or weak (Mujis, 2010, p. 70). Since SPSS does not provide this figure, I use the following formula to compute the eta squared value: $t^2 / t^2 + (N1 + N2 - 2)$[11] (Dörnyei, 2007). This value can be interpreted as "the percentage of the variance in the target variable explained by the grouping variable" (ibid., p. 217).

6.4.3.2 Students' performance on the tests

Table 6.14 summarises the two groups of Taiwanese participants' performance on single- and multi-word knowledge in pre- and post-tests, including mean scores (*M*), standard deviations (*SD*), the degrees of freedom (*d*), the t-value and the effect size. There was a significant difference in scores for experimental and control groups in both the single-word test (p=.027) and multi-word test (p=.031) after the treatment, with an eta squared value of .20 and .36 respectively, indicating that 20% and 36% of the variance in post-test scores could be

[11] *N* refers to the size of the groups; *t* refers to the t-values produced when running a t-test in SPSS.

accounted for by the treatment. These eta squared values, according to Dörnyei (2007), suggest a very large effect size[12], which indicates a substantial impact of the intercultural contact on the acquisition of both single-word and multi-word knowledge.

Table 6.14

Independent-samples T-Tests of the Taiwanese Students' Performance on Single- and Multi- Word Tests (N=35 in each group)

	Pre-test		Post-test				
	M	*SD*	*M*	*SD*	*d*	*t*	Effect size[a]
Single-word test					68	4.22*	.20
Experimental group	68.40	19.51	81.63	13.33			
Control group	60.80	22.38	66.74	21.99			
Multi-word test					68	6.23*	.36
Experimental group	21.71	5.88	31.82	5.14			
Control group	20.57	7.08	25.97	8.22			

* $p < .05$

[a] Eta squared.

We can see here the analysis from the experimental approach has reported the development of Taiwanese learners' lexical proficiency, which was shown in the single- and multi-word tests. From the three investigations presented in 6.4.1, 6.4.2 and 6.4.3, it was clearly shown that intercultural online exposure to native English speakers facilitates the development of three-word sequences for young Taiwanese learners over one year of contact. It appears that online exposure to native English speakers provides language immersion in an authentic socio-cultural context in which the target language was used, and thus students were given ample opportunity to encounter a considerable amount of authentic

[12] The usual interpretation of eta squared is that .01=small effect, .06=moderate effect, and .14=large effect (Dörnyei, 2007, p. 217).

language used by native English-speaking people through being able to read messages, interact with their peers and learn the formulaic multi-word expressions habitually used by British students in such online interactions. An online discussion forum can thus be a useful source of authentic materials for EFL learners, giving learners an opportunity to learn by themselves by observing the linguistically and culturally relevant features in context (Montero et al., 2007). Nevertheless, it is advisable that teachers choose online tools carefully, with a view to suiting their aims and their students' particular learning needs.

6.5 Summary

This chapter focuses on the functional and developmental perspectives of the use of multi-word sequences in intercultural CMC and FTF communication. The discourse functions of three-word sequences in BATTICC-O and BATTICC-F, as well as two large reference corpora (CANELC and CANCODE), have been examined, and it was evident that the high-frequency three-word sequences in the four datasets realise different purposes in online and FTF communication, serving three central discourse functions: social interaction, necessary topics and discourse devices. The results indicate that even though three-word units do not usually constitute complete grammatical or idiomatic structures, they provide important building blocks upon which the speakers or writers can create more extended utterances for different purposes in language use (Biber, 2009; Schmitt, 2010, 2013). The results have been presented in detail in 6.3 and summarised in 6.3.5.

This chapter has also examined the developmental perspectives of multi-word

sequences, analysing the extent to which intercultural contact with native English speakers affects the use of multi-word sequences by young Taiwanese learners in a one-year online intercultural exchange programme, as presented in 6.4. The section first investigated the overlap of three-word sequences used by British and Taiwanese learners, concentrating especially on the sequences including the coordinating conjunctions *and* or *but*. A progressive development in terms of quantity and percentage of the three-word sequences of this type was found in Taiwanese learners' discourse. The study then considered the key sequences used by the Taiwanese and British students over the course of the programme. There was a continuous decline by Taiwanese participants in the number and percentage of sequences identified as significantly over or underused relative to the discourse of the British participants. Both analyses, in turn, showed an increasingly close approximation by Taiwanese learners to the use of three-word sequences by native speakers of English. This confirms previous findings (e.g., Adolphs & Durow, 2004) that intercultural contact with native speakers of English fosters the longitudinal development of the use of multi-word sequences, and this also contributes additional evidence that online interaction can achieve such positive results. The method here, which focuses on naturally occurring language output online, diminishes the effects of the artificial contexts often created in language testing settings.

Moreover, an independent-samples t-test was carried out to compare the scores for the experimental and control groups of participants in order to investigate whether participants who have regular intercultural contact with native speakers could improve their word knowledge more than students who simply receive classroom instruction and study by themselves. There was a significant difference in scores

for the two groups in both the single-word test (p=.027) and the multi-word test (p=.031) after the treatment, with a large effect size, indicating that the intercultural contact had a great impact on the acquisition of both single-word and multi-word knowledge.

With respect to the acquisition of multi-word sequences, it is widely accepted that mastering the formulaic expressions commonly used by English native speakers is difficult for young learners of English as a foreign language (O'Keeffe et al., 2007; Schmitt, 2010; Wray, 2000). Wray (2000) indicates three possible reasons. One is the poor quality of the learner's learning experience, and they are often not taught very well in class. It is also easy for the wrong sequences to be taught. One way around these problems, therefore, is to "provide learners with language experience which offers the exposure to the most useful patterns of the language" and to help them "to notice patternings and to speculate about them" (Willis, 1990, p. 38). As such, when learners are engaged in meaningful activities that involve manipulating language, they learn more and retain that information longer (Reppen, 2010). This is often referred to as "data-driven learning" (DDL), a learning process that "confronts the learners as directly as possible with the data" and tries to "make the learner a linguistic researcher" (Johns, 2002, p. 108). Take, for example, the modal *would*, which causes difficulty in use for many of the Taiwanese young learners. A study of the selected concordance lines below, taken from the British participants' texts, would allow the students to know more about the appropriate use of the multi-word sequences:

hope to learn some soon, I think	***it would be***	very fun to learn another language!
yeh I love ice skating, I think	***it would be***	very good if you tried it because
I would love to be friends haha	***it would be***	really nice to have a penpal to

I always sleep too much!	***it would be***	cool though, I can play badminton
but i looovveee oysters so i think	***it would be***	very very nice!!

play any musical instruments but i	***would love to***	be able to play the acoustic guitar.
hello Enya i	***would love to***	make friends with you
I have never had Taiwanese tea, but	***would love to***	try it when we come over.
they do rock music so i	***would love to***	know what sort of music you and
Yes sure, I	***would love to***	be friends haha, It would be really

The concordance data above is selected to illustrate two of the useful expressions with modal *would* in those instances where *it would be* is used for expressing possibility/complying and *would love to* is commonly used for volition and needs. This information is explored based on the discourse that native English speakers actually produced, rather than by intuition, and such examples can be very useful resources for the development of EFL learning materials. I summarised these concordance lines in a sample teaching material (see Appendix F), demonstrating how authentic data from BATTICC can be used to inform vocabulary and multi-word sequence instruction. The evidence from the current study along with much previous corpus-based research suggest that many multi-word sequences are as frequent as or more frequent than the single-word lexical items. Learners therefore cannot fully understand a word without learning how it is used as part of formulaic multi-word expressions in context.

CHAPTER 7

Textbooks and Authentic Intercultural Communication

7.1 Introduction

From Chapters 4 to 6 this thesis has presented the investigation of British and Taiwanese participants' discourse in authentic intercultural encounters and identified the particular patterns of language use by the two groups of teenagers. It is apparent that a significant number of patterns of language that serve important discourse or interpersonal functions in natural communication could not be found in Taiwanese participants' discourse in both online and spoken communication. To investigate further, this chapter examines the EFL textbooks used by the Taiwanese participants and considers what the role of textbooks might be in this context, addressing the following research questions:

1. To what extent do the EFL textbooks used in Taiwanese junior high schools display the most distinctive linguistic features of naturally occurring discourse in intercultural communication?
2. How can corpus evidence support the development of EFL teaching materials?

This section explores the language use in TETCOC [Taiwanese EFL Textbook Corpus of Conversation] and BATTICC-F, concentrating particularly on the use of three-word sequences (7.3) and spoken grammar (7.4) in the two corpora. Before these results are presented, the following section (7.2) will provide a brief review of textbooks in language learning and teaching.

7.2 Textbooks in language learning and teaching

EFL textbooks have been considered as the basis for much of the language input that English language learners receive (Tomlinson, 2011). They serve as "a facilitative instrument for learning" (Trabelsi, 2010, p. 105), which provide content for the lesson, supplement teachers' instruction, simulate language use and offer learners opportunities to experience language in use, as well as helping learners make discoveries about the language for themselves (McGrath, 2002; Richards, 2005; Tomlinson, 2011). In an EFL context in particular, textbooks may "even constitute the main and perhaps only source of language input that learners receive and the basis for language practice that occurs both inside and outside the classroom" (Nguyen, 2011, p. 18). As a result, it stands to reason that effective pedagogical materials should expose learners to language use in authentic contexts. However, artificiality can still be identified throughout teaching and learning materials used in the EFL classroom. One short example, taken from a corpus of Taiwanese textbooks used in junior high schools, is presented below:

> Peter: Where's Linda?
> Sam: She's practicing badminton at the gym.
> Peter: What time does she practice?
> Sam: She practices from five to six.
> Peter: What day is today?
> Sam: It's Thursday.
>
> (TETCOC)

As can be seen in the script, the turn-taking is neat, tidy, and predictable without any hesitation and the utterances are all complete sentences. In the dialogue Peter constantly asks questions and Sam answers; the dialogue has the appearance of an interview rather than actual casual conversation. This pedagogic artifice is understandable as it may well reflect the textbook writers' perceptions regarding

teachability and learnability at a beginner or intermediate level. Widdowson (1998) also notes that it is actually impossible to use authentic language data in the classroom as "the language cannot be authentic because the classroom cannot provide the contextual conditions for it to be authenticated by the learners" (p. 711). Invented data, then, is "perfectly justified in materials as a stage in the process of becoming a competent user of another language" (Gilmore, 2004, p. 371).

However, to what extent should we deprive our learners of exposure to authentic language use? In addressing this question, previous studies of commercially produced textbooks have criticised them for not offering natural contexts (e.g., Carter et al., 2011; Cullen & Kuo, 2007; Gilmore, 2004; Nguyen, 2011). Gilmore (2004), for example, investigated the discourse features of textbook dialogues and contrasted them with comparable authentic interactions in a corpus. He found that a range of typical features of naturally occurring conversation, such as false starts, repetitions, pauses, latching, terminal overlap, back-channels and hesitation devices, are identified in extremely few instances in textbook conversation. Gilmore also reports that recently published textbooks are beginning to incorporate more of the discourse features found in authentic data. In addition, studies have also analysed the grammatical features of EFL textbooks, namely spoken grammar (Carter et al., 2011; Cullen & Kuo, 2007; Lin, 2012). They found that the EFL textbooks lack core spoken language features, such as discourse markers, vague language, headers, tails, situational ellipsis and loose grammatical agreement (e.g., *There's lots of things left to do*). It seems that there continues to be a substantial missing link between what tends to be presented to learners in classroom experiences of the target language and the actual language used in

natural conversation outside the classroom.

7.3 Multi-word sequences: textbook conversation vs. naturally-occurring communication

Although EFL textbooks have been criticised for not offering natural contexts (Gilmore, 2004, 2007; Römer, 2009; Meunier & Gouverneur, 2009), few have systematically examined the use of multi-word sequences in textbooks. Even fewer investigate how young learners of English use multi-word sequences for intercultural communication and what the role of textbooks might be in this context. This section investigates the most frequently used three-word sequences in the Taiwanese EFL Textbook Corpus of Conversation (TETCOC) to reveal what kind of language exposure the participants receive and whether the textbooks present the patterns that are commonly used in authentic intercultural communication.

Chapter 6 has explored the discourse functions of three-word sequences in intercultural CMC and FTF communication (BATTICC-O and BATTICC-F), as well as two large reference corpora (CANELC and CANCODE). It was evident that the high-frequency sequences in the four datasets serve three central discourse functions: social interaction, necessary topics and discourse devices. This section identifies the most common recurrent sequences in TETCOC, which includes three recent series of EFL textbooks used in junior high schools in Taiwan, and contrasts them with naturally occurring communication among the Taiwanese participants interacting with adolescents based in the UK (i.e., BATTICC-F). The framework used for this analysis is the same as the one used in Chapter 6. In addition, keyness method was applied to reveal the overuse and underuse of

three-word units in the textbook conversation as compared to the authentic discourse in BATTICC-F. The research findings will demonstrate the pedagogical merit of keyness analysis and thus help to inform teachers and materials writers in relation to course design for intercultural interaction of adolescents.

7.3.1 Functional types of three-word sequences in TETCOC

This section pays particular attention to the high-frequency sequences and their discourse functions in TETCOC. In the analysis the 50 most frequent three-word units were first automatically extracted from TETCOC using *WordSmith Tools*. Nevertheless, a number of items do not have a clearly recognisable function, such as *to go to* and *to be a*, which are mainly composed of high-frequency function words. They were therefore excluded from the list of the 50 most common three-word sequences. Each of the three-word units in the list was then examined in its original discourse contexts to identify its primary discourse function, and accordingly all items were inductively grouped into three central categories: social interaction, necessary topics and discourse devices. As was mentioned in Chapter 6, assigning a sequence to a category is sometimes rather difficult owing to the multi-functional nature of multi-word sequences. For example, the sequence *would you like* functions as an offer, an invitation or occasionally a request, as can be seen in (1) and (2).

(1) (Nora is talking to Lisa.)
Nora: My parents are planning to go camping on Yushan. *Would you like* to join us?
Lisa: Sure! It's the highest mountain in Taiwan, right?

(2) (During the break)
Lisa: I don't think so. I don't have a headache, and I don't have a runny nose, either.
Ella: Maybe you're just hungry. *Would you like* to have a donut?

Lisa: No, thanks. My stomach hurts. I can't eat anything now.

(TETCOC)

From (1) and (2) we can see that different textbooks use *would you like* with different discourse functions in different contexts. In (1), *would you like* is used for an invitation to go camping, while (2) presents it as an offer of a donut. However, no explicit information with regard to appropriate use for an invitation or an offer is presented in either textbook. In the two excerpts, moreover, limited information about the relationship between the speakers is presented; for example, excerpt (1) simply states *Nora is talking to Lisa* at the top of the conversation, which seems to be insufficient to understand the relationship between them. In most other cases, there is not even any description about the context. This indicates an inadequate treatment with regards to the presentation of speech acts in TETCOC. This is in line with Nguyen's (2011) investigation of Vietnamese EFL textbooks, showing that little attempt is offered by textbooks to explicitly draw students' attention to the situational context and contextual variables or how they affect the appropriate use of speech acts in different situations. In the whole TETCOC, seven instances of *would you like* are found, and I classified four of them as offers. The sequence *would you like* is therefore grouped in the subcategory of offering.

Table 7.1 presents the functional categories of the 50 most common three-word units retrieved from BATTICC-F and TETCOC. Within each functional category the items are sequenced in descending order of frequency. It can be noted that the vast majority of the high-frequency multi-word sequences in TETCOC are used for social interaction, accounting for 38 (76%) of the 50 most common sequences. This seems to suggest that the textbooks used in Taiwanese junior high schools

Table 7.1

Functional Categories of Three-word Units in TETCOC and BATTICC-F

TETCOC	*BATTICC-F*
SOCIAL INTERACTION	
Greeting to meet you, nice to meet	**Greeting** nice to meet, to meet you, how are you
Questioning what are you, do you want, you want to, do you have, are you doing, did you go, what do you, did you do, how about you, what did you, do you know, do you like, where are you, are you going, can we do, how old are	**Questioning** do you like, do you have, what do you, how do you, do you think, have you ever, how is the, did you see, do you go, how about you, do you want, you want to
Requesting let me see	**Inviting** come to Taiwan
Commanding look at the, you have to	**Asserting: personal** I don't know, I want to, we have a, we went to, I have a, I think it's, we have to, I think I
Offering would you like, you like to	**Asserting: impersonal** it was very, it's very nice
Suggesting maybe we can, why don't we, let's go to	
Asserting: personal I want to, I have to, want to go, I went to, this is my, I was in, want to be, you have a, I'm going to, we have a, we're going to, I need to	
Asserting: impersonal there are many	
NECESSARY TOPICS	
Location go to the, at the park, the living room, in the classroom, at the party, in front of, in the living, on the street	**Autobiography** my name is
	Location in the UK, in your country
Quantity a lot of	**Quantity** a lot of, a lot more, a couple of
Other topics want to buy, I don't like	**Food/likes** fish and chips, what's your favourite, I don't like, I like it
	School go to school, in your school
	Other topics Dragon Boat Festival
DISCOURSE DEVICES	
Linking functions by the way	**Linking functions** and it was, and we have, and we were, so do you
	Fluency devices erm do you, you know I
	Exemplifiers sort of thing, sort of like, I was like, it was like
	Evaluators to be honest

pay much attention to social interaction, with a view to preparing learners for successful daily-life communication. For example, a large amount of conventionalised language typically associated with different speech acts in communication is presented, such as *nice to meet* and *to meet you*, which are both part of an extended sequence *nice to meet you,* used to express a formal greeting; *I have to,* used to express personal intention; and *would you like*, used to express offering.

Multi-word sequences used during questioning are predominant in both corpora. This can be explained in part by the highly interactive nature of spoken conversation, in which people are constantly asking and responding to questions, which, in turn, may facilitate personal relationship building (Belz, 2007). This suggests that the people in authentic intercultural exchange, such as that captured in the BATTICC-F corpus, demonstrate *skills of discovery and interaction*, namely "the ability to employ a variety of questioning techniques in order to elicit from members of the foreign culture" (Byram, 1997, p. 61). Moreover, as can be seen in Table 7.1, a large number of multi-word sequences used in social interaction are for the expression of assertions, similar to what Biber et al. (2004) call *attitudinal/modality stance bundles*, which express "attitudes toward the actions or event described in the following proposition" (p. 390). Asserting sequences are divided into personal and impersonal; in interaction, most are overtly personal and express desire (e.g., *I can't wait*), personal opinions (e.g., *I think it*), intention/prediction (e.g., *I'm going to*; *I hope you*), ability (e.g., *I can play*; *be able to*), or obligation (e.g., *you have to*). These sequences are directly attributed to the speaker. Nevertheless, impersonal asserting sequences do not explicitly mention the speaker, such as in descriptions of existence (e.g., *there are*

many), evaluations of specific things or events (e.g., *it's very nice*), narratives of past events (e.g., *it was very*), or predictions of future events (e.g., *going to be*).

Another major function of the use of multi-word sequences is that of introducing or progressing necessary topics (the second section of Table 7.1). These sequences provide overt signals of specific themes, such as autobiography (e.g., *my name is*), food (e.g., *fish and chips*), location (e.g., *in the UK*), school (e.g., *in your school*), likes (e.g., *what's your favourite*), quantity (e.g., *a lot of*), and some culturally specific topics (e.g., *dragon boat festival*). Within this category it can also be seen that most of the highly recurrent items in TETCOC are related to location (e.g., *at the park, the living room, in the classroom*), while the topics that are commonly talked about in authentic adolescent intercultural communication, such as food, likes and schools, are not presented with a high frequency in TETCOC. With regard to the domain of quantity, more types of quantity markers are found in BATTICC-F, such as *a couple of, a lot of, a lot more* and *a little bit*, indicating purposive vagueness and approximation, which are not commonly found in TETCOC. As was claimed by O'Keeffe et al. (2007), "approximation rather than precision is the norm in conversation" (p. 74).

The third category of multi-word sequences is discourse devices, which connect the meaning and the structure of the discourse. As such, they serve an organising function for the flow of information being transmitted and further improve the fluency of utterances. From Table 7.1, clearly various types of discourse devices are used in BATTICC-F, while only one item is found in TETCOC. In the sequences serving linking functions we can see that the coordinating conjunction *and* is frequently used to express a variety of logical relations between phrases

and sentences in conversation. Examples of these linking sequences have been presented in 7.3, but such use can rarely be found in the whole textbook corpus (in only 4 instances). In the domain of fluency devices it is also worth considering the sequence *erm do you*, which marks the speaker's hesitation and planning in their utterance. Although these hesitation markers have important functions in discourse, TETCOC only presents 22 instances of *er* and *erm*, while in BATTICC-F 249 instances are found (0.99%), a rate that is similar to the large corpus of native-speaker discourse CANCODE (1.09%).

The other important item in the category of fluency devices is *you know I*, which includes a frequently used two-word unit *you know*. While it is common in BATTICC-F, only 2 instances are found in TETCOC, as in the following excerpts:

(3) Sandy: There! Over those flowers! Aren't they beautiful?
Mrs. Beck: Yes, they are. ***You know***, we should come out more often.
Sandy: I think so too. Next time we should come earlier and watch the sun rise.

(4) Nora: Taking such a long trip by bicycle was really difficult. But they made it.
Ella: They're really good at cycling.
Nora: Yes, they're very good. But ***you know***, cycling is just their hobby.

(TETCOC)

In (3) and (4), *you know* functions in this context as an interpersonal discourse marker, indicating speaker attitude, inviting "the addressee to recognise both the relevance and the implications of the utterance marked with you know" (Jucker & Smith, 1998, p. 194). That is, *you know* is used as a device to aid in the joint construction of the representation of the event being described. *You know* is also used to introduce additional information marking its relevance to the current issue, as in (3). In this case it simply displays the speaker as an information provider who depends upon hearer reception of information (Schiffrin, 1987). Although

Table 7.2

Distribution of Functional Types in TETCOC and BATTICC-F

	TETCOC		BATTICC-F		Significance	
	No.	Freq.	No.	Freq.	LL	*p*
SOCIAL INTERACTION	**38**	**411**	**25**	**272**	**+19.42**	
Greeting	2	24	3	27	-0.46	
Questioning	16	228	11	133	+18.96	*
Offering	2	15	0	0	+19.74	**
Requesting	1	8	0	0	+10.53	**
Commanding	2	14	0	0	+18.42	***
Inviting	0	0	1	9	-13.13	
Suggesting	3	20	0	0	+26.32	***
Asserting: personal	11	94	8	84	+0.07	
Asserting: impersonal	1	8	2	19	-5.44	*
NECESSARY TOPICS	**11**	**118**	**12**	**110**	**+0.18**	
Location	8	72	2	17	+32.75	***
Quantity	1	31	3	35	-0.61	
Food/Likes	0	0	4	32	-46.69	***
Schools	0	0	2	18	-26.26	***
Others	2	15	1	8	+1.69	
DISCOURSE DEVICES	**1**	**18**	**11**	**87**	**-54.42**	***
Linking functions	1	18	4	38	-8. 81	***
Fluency devices	0	0	2	13	-18.97	***
Exemplifiers	0	0	4	29	-42.31	***
Evaluators	0	0	1	7	-10.21	***

*p <.05 (LL>3.83), **p <.01 (LL>6.64), ***p <.001 (LL>10.83)

two different functional uses of *you know* are presented in textbook conversation, many other common functions served by *you know* in BATTICC-F are not found in TETCOC, such as a marker of shared knowledge or experience between speakers and hearers, an indicator of a type of thinking process and a device for reformulation, elaboration and hesitation. This indicates that Taiwanese learners are not formally taught the pragmatic uses of *you know* as the textbooks only

present it in 2 instances, and such a minimal language exposure to this particular lexical item is not sufficient for language acquisition.

The distribution of the functional types of the 50 high-frequency thee-word sequences in TETCOC and BATTICC-F are summarised in Table 7.2, with log-likelihood (LL) values, which indicate the statistical significance of the total frequencies of sequences for different functions between the two datasets. In the column of LL, a positive (+) and a negative (-) value indicates an overuse and an underuse respectively in TETCOC (compared with BATTICC-F). From the table it can be seen that among the three different functional domains, the category of discourse devices in the two datasets reaches the most significant difference, followed by necessary topics and social interaction.

Regarding multi-word sequences for social interaction, TETCOC generally presents more instances than BATTICC-F. Within the category, significantly larger numbers of three-word sequences for questioning, requesting, commanding, offering and suggesting are found in TETCOC in comparison with those items used in BATTICC-F, while TETCOC significantly underuses the sequences for asserting (impersonal). This suggests that the speech acts of directives, which are attempts by the speaker to get the addressee to do something (O'Keeffe et al., 2011), are overused in TETCOC. This notwithstanding, TETCOC generally covers a wider range of multi-word sequences indicating different speech acts than BATTICC-F.

In the comparison of three-word sequences for necessary topics, although the total amount of use between the two datasets does not reach a significant level, some

topics within this category differ significantly. For example, sequences indicating locations are significantly overused in TETCOC as compared with BATTICC-F, while the items used for describing food, likes and schools are underused in TETCOC.

With respect to discourse devices, an extremely high LL value is generated (LL=-55.27), which shows a highly significant difference in the use of discourse devices between the two corpora. Within this category all types of discourse devices reach a significant level, including linking functions, fluency devices, exemplifiers and evaluators. In particular, four exemplifiers (i.e., *sort of thing, sort of like, it was like, I was like*) are commonly found in BATTICC-F, but no instances of such units can be seen in TETCOC.

The comparison of three-word sequences and their discourse functions in the TETCOC and BATTICC-F has indicated a gap in language use between textbook conversation and authentic intercultural communication. This further informs textbook writers and classroom teachers of the need to include a broader range of multi-word sequences for different pragmatic functions and topics and present the items that serve discourse devices in natural contexts when teaching conversation.

7.3.2 Analysis of key sequences

The underused and overused three-word sequences of different functional types in the comparison of TETCOC and BATTICC-F have been outlined in the previous section. This section attempts to identify the particular items with an unusual frequency in the comparison of two corpora. Using *WordSmith Tools 5.0* to analyse the key sequences of TETCOC by comparing TETCOC and BATTICC-F

reveals 41 sequences that are significantly overused in the textbook corpus, as compared to BATTICC-F (p <.01 with 1 d.f., log-likelihood >6.64). This suggests that these 41 three-word units commonly presented in Taiwanese textbooks are rarely used in authentic intercultural communication. The top 15 items with the largest keyness (Log-likelihood) values are presented in Table 7.3.

Table 7.3

Key Sequence List: Overuse in TETCOC as Compared with BATTICC-F

	Key sequence	TETCOC		BATTICC-F		Keyness
		Freq.	%	Freq.	%	(LL)
1	WHAT ARE YOU	28	0.11	0	0	38.82
2	ARE YOU DOING	19	0.07	0	0	26.34
3	I HAVE TO	16	0.06	0	0	22.55
4	DID YOU DO	14	0.06	0	0	19.14
5	WHAT DID YOU	13	0.05	0	0	18.02
6	THIS IS MY	13	0.05	0	0	18.02
7	THE LIVING ROOM	13	0.05	0	0	18.02
8	I WENT TO	12	0.05	0	0	16.17
9	IN THE CLASSROOM	12	0.05	0	0	16.17
10	AT THE PARK	11	0.04	0	0	14.33
11	AT THE PARTY	10	0.04	0	0	13.86
12	LET ME SEE	10	0.04	0	0	13.86
13	WHERE ARE YOU	10	0.04	0	0	13.86
14	ON THE STREET	20	0.08	2	0.01	12.02
15	BY THE WAY	8	0.03	0	0	11.09

As can be seen from the table, the first two sequences in many cases in TETCOC are both part of an extended sequence *what are you doing?*, which appears frequently (approximately 20 times) in TETCOC. However, no instances of such a sentence are found in BATTICC-F since in face-to-face communication it might be redundant to ask the interlocutors *what are you doing?* With regard to the top 15 overused sequences of TETCOC presented in the table below, 5 items can be categorised as questioning and 5 sequences clearly mark locations, which confirms the results in the previous section, showing that questioning and location

markers are significantly overused in TETCOC, as compared with BATTICC-F. On the other hand, in the analysis of underused items in TETCOC, for $p < .01$ with 1 d.f., cut-off of 6.63, 64 sequences that reach a significant level are revealed. This indicates that a considerable number of sequences frequently used in authentic intercultural communication are not presented in TETCOC. Table 7.4 presents the 15 most underused sequences (those with the largest keyness values). This also accords with the earlier results, which shows that TETCOC underuses the multi-word sequences for asserting: impersonal (e.g., *it was very*), the topic of food (e.g., *fish and chips*), linking functions (e.g., *and we have*, *and it was*) and fluency devices (e.g., *erm do you*). Along with the previous investigation, this result shows that many of the sequences that commonly appear in natural adolescent intercultural discourse are not presented at a high frequency in EFL textbook conversation, and the input that learners receive is therefore impoverished in this regard.

Table 7.4

Key Sequence List: Underuse in TETCOC as Compared with BATTICC-F

	Key sequence	BATTICC-F		TETCOC		Keyness
		Freq.	%	Freq.	%	(LL)
1	DO YOU LIKE	35	0.13	8	0.03	36.11
2	FISH AND CHIPS	12	0.05	0	0	20.02
3	IT WAS VERY	11	0.04	0	0	18.02
4	ERM DO YOU	11	0.04	0	0	18.02
5	I DON'T KNOW	22	0.08	5	0.02	16.54
6	SO DO YOU	10	0.04	0	0	16.02
7	AND WE HAVE	10	0.04	0	0	16.02
8	AND IT WAS	10	0.04	0	0	16.02
9	IN THE UK	10	0.04	0	0	16.02
10	HOW IS THE	10	0.04	0	0	16.02
11	I THINK IT'S	10	0.04	0	0	16.02
12	ERM I LIKE	10	0.04	0	0	16.02
13	AND WE WERE	8	0.03	0	0	14.01
14	I LIKE IT	8	0.03	0	0	14.01
15	IT'S VERY NICE	8	0.03	0	0	14.01

A closer look at the use of these units also reveals some important pragmatic aspects that are rarely presented in TETCOC. For example, the high-frequency units *do you think* and *I don't know* are commonly used in BATTICC-F as indirectness markers, which are important for polite and non-threatening expressions of attitude, opinion and stance (Carter & McCarthy, 2006; O'Keeffe et al., 2007), as in *do you think it would be ok to do that?* and *I don't know if it's right* (from BATTICC-F). In fact, these utterances can be formed in a more direct way (i.e., *I want to do that* and *it's wrong*), but instead the speakers employ the multi-word sequences that have pragmatic integrity in order to soften the utterances for mutual protection of face. Moreover, the recurrent three-word sequences *I think it's* and *I think I* indicate the pervasive use of *I think* as a hedge modifying evaluation of situations or assertions to make them less assertive and less open to challenge or refutation (Carter & McCarthy, 2006). It appears that some of the common multi-word sequences in BATTICC-F exhibit pragmatic adequacy and integrity and play a significant role in the polite progression of the talk, but they can rarely be found in TETCOC.

7.4 Spoken Grammar: textbook conversation vs. naturally-occurring communication

The previous section has revealed the gap between BATTICC-F and TETCOC regarding the discourse and pragmatic functions of multi-word sequences. This section reports on an investigation into the spoken grammar in TETCOC and contrasts it with BATTICC-F, with a view to examining the extent to which spoken grammar is reflected in contemporary textbooks for EFL learners in Taiwan. The same framework adapted from work by Carter and McCarthy (2006) and Cutting (2011) and used for the analysis of BATTICC in Chapter 5 is applied

to this analysis, comprising lexical features (e.g., vague expressions, approximations, and hedging), syntactical features (e.g., ellipsis, headers and tails) and discourse features (e.g., discourse markers, hesitation and turn-taking patterns). This study also examines the spoken grammar used by Taiwanese learners in intercultural communication and discusses what the role of textbooks might be in this context. The research findings identify gaps between textbook conversation and naturally occurring intercultural discourse, and I suggest opportunities for how teachers might bridge these gaps and support learners to achieve better spoken communication.

In this section I pay particular attention to the analysis of TETCOC in the five following aspects: (1) situational ellipsis, (2) vagueness and approximation, (3) headers and tails, (4) pausing, repeating and recasting and (5) discourse marking. These five distinctive features of naturally occurring discourse have been previously identified in section 5.4. Table 7.5 illustrates the total frequencies and percentages of different features of spoken discourse found in TETCOC and BATTICC-F. It is apparent that most of the categories reach a highly significant difference between the two datasets. This suggests that these distinctive features of spoken discourse are significantly underused in TETCOC as compared with data collected in authentic communication. In particular, headers and tails and repeating and recasting are not found in TETCOC. The following subsections will focus on the four aspects of spoken grammar: situational ellipsis (7.4.1), vagueness and approximation (7.4.2), pausing, repeating and recasting (7.4.3) and discourse marking (7.4.4). The different uses concerning frequencies and discourse functions in TETCOC and BATTICC-F will be discussed.

Table 7.5

Spoken Grammar in TETCOC and BATTICC-F

Vague expressions	TETCOC		BATTICC-F		Sig.
	Number	per 1000 words	Number	per 1000 words	(LL)
Vague expressions	5	0.20	137	8.11	207.72
Approximation	37	1.51	68	4.02	24.49
Situational ellipsis	7	0.29	89	5.27	116.86
Headers and tails	0	0.00	14	0.83	25.12
Pauses	67	2.73	1647	97.46	2459.0
Repeating and recasting	0	0.00	38	2.25	68.17
Discourse marking	660	26.89	1400	82.85	619.18

7.4.1 Situational ellipsis

As discussed in 5.4.2, situational ellipsis involves the deliberate omission of items such as subject pronouns and verb complements that may not be necessary in the utterances since they contain enough information for the purpose of the conversation (Carter & McCarthy, 2006; Thornbury & Slade, 2006). The BATTICC-F contains 89 instances of situational ellipsis in the Taiwanese and British participants' discourse. The frequent elliptical elements include initial *I* plus copular verb *be* in declaratives (e.g., *I'm*), subject pronoun *it* or other demonstrative pronouns plus *be* (e.g., *it's*), interrogatives (e.g., *what's*), subject pronouns (e.g., *we, he*), existential *there* (and its accompanying verb *be*), and copular verb *be* and prepositions (e.g., *in, at*). However, in TETCOC, situational ellipses occur in only seven instances. Some examples are shown in the following excerpts:

(5) Nora: Are we going to exchange gifts?
Eric: (***That***) Sounds great!

(6) Harry: Mom, how long will it take us to get to Singapore?
Mom: (***It will take***) About four hours.

(7) Sam: Yeah, many boys do. They spend most of their free time on sports. But one of us is strange - he wants to be a nurse!

Hank: (***That's***) Not strange at all.

(TETCOC)

Table 7.5 shows that situational ellipses occur significantly more in BATTICC-F than in TETCOC. This may be understandable when we bear in mind that traditionally ellipsis is considered as an incorrect form of written grammar. However, this study, as well as previous research (e.g., Carter & McCarthy, 2006; Cullen & Kuo, 2007), has shown that ellipsis appears frequently in natural native-speaker English conversation. Consequently, it seems reasonable for EFL textbooks to include ellipsis more often in their content owing to the fact that in real-time informal communication, ellipsis is actually more appropriate than full grammatical forms (Carter & McCarthy, 2006; Mumford, 2009), and hence is an important aspect of spoken competency.

7.4.2 Vagueness and approximation

Vague language is another pervasive feature of spoken English. Section 5.4.1 found a large amount of vagueness in BATTICC-F, including vague categories, approximations and hedging. In BATTICC-F, 104 instances of expressions indicating vague categories are found, which typically include words and phrases such as *thing, stuff, like, (or/and) something, (or/and) anything, kind of* and *sort of*. However, in TETCOC, only five instances of this type of vague language are found, and the lexical choices are quite limited, with three instances of *something* and two instances of *anything*. The most highly frequent items in BATTICC-F, such as *stuff, thing, sort of* and *kind of*, are therefore not included in TETCOC. The following excerpts present the use of *something* and *anything* in textbook dialogues:

(8) Clerk: May I help you?
Emma: Yes. We want to buy ***something*** for our American friend.

(9) (Richard tries on an orange cap.)

Clerk: It looks good on you.
Richard: Thanks. I'll take it.
Clerk: That's one hundred fifty dollars.
Richard: Here you go.
Clerk: Here's your change. Have a nice day!
Richard: Emma, did you find ***anything***?
Emma: No. Let's go to another shop.

(TETCOC)

Another type of vague language is the use of approximations, which are particularly used to modify numbers, quantities or some other measurable units. These are frequently introduced by speakers in BATTICC-F in order to downtone what might otherwise sound overly precise. In TETCOC a total of 37 instances are found, including a range of different lexical choices, such as *lots of/a lot* (15 instances), *a little* (12 instances), *about* (7 instance) and *a couple of* (3 instances), as in:

(10) Mr. Yang: I'll spend ***a couple of*** nights in a hotel in New York.
(11) Harry: I know it's boring. But give me ***a couple of*** minutes, and you'll be surprised to see what happens.
(12) Kevin: Yes. How many eggs do you need?
Tina: Let me see ... I need ***about*** twenty.
(13) Ella: Wow! Were you close to him?
Lisa: Yes. I was ***about*** ten feet away. He was so cool on the stage.
(14) Linda: Do young men in Canada play dodge ball, too?
Peter: Yes, we do. It's ***a little*** different.
(15) Peter: We put ***a little*** butter on the person's nose for good luck.

(TETCOC)

As discussed in 5.4.1, vagueness is relational language that is frequently used by speakers to convey information that is softened in some way so the utterances do not appear overly direct or unduly authoritative and assertive. Its absence, therefore, may result in language that sounds more domineering than the speaker may intend (Carter & McCarthy, 2006; Mumford, 2009; O'Keeffe et al., 2007). In

light of this, it is concerning to note that this feature is all but absent from textbook conversation. In this case, the learners may risk being perceived as overbearing or pedantic if they apply the interaction model learned from the textbooks in authentic intercultural communication.

7.4.3 Pausing, repeating and recasting

As discussed in 5.4.4, pausing, repeating and recasting are typical features of naturally occurring discourse, and they allow speakers to buy time for speech planning and keep the floor; meanwhile listeners are also provided with time to figure out what is going on and what will come next, which can further aid the comprehension of communication. In the dataset 67 pauses are found, including 22 instances of filled pauses (i.e., *er, erm*) and 55 entries of unfilled ones (i.e., ...). For example:

(16) Tina: Well ... My cousin, David, has big eyes, too.
(17) Tina: Well ... I only need three bags.
(18) Sarah: Erm ... I play tennis from Monday to Thursday.
(19) Emma: Er ... I don't like the color.
(20) Father: Well, I....
(21) Tony: I hate spring. There's too much rain! Look, all my books are wet. And my clothes, my shoes....

(TETCOC)

As can be seen in the excerpts from TETCOC, unfilled pauses (i.e., ...) mainly appear following the cognitive discourse marker *well*, as in (16) and (17). It can also be noted that pauses frequently collocate with *er* or *erm*, indicating a speaker's lack of certainty, as in (18) and (19). They also occur as an ellipsis in an utterance, showing that the speaker has something more to say, as in (21), or simply as an unfilled pause that marks the end of a turn, as in (20). Although pauses are presented in TETCOC with different discourse functions, they are

relatively scarce. Research shows that pauses have very important discourse functions, allowing speakers to buy time for speech planning and maintaining the floor if the pause is filled with *er* or *erm*. At the same time listeners are also provided with time to anticipate what will come next, which can further aid the comprehension of communication (O' Keeffe et al., 2007). As Gilmore (2004) claims, pauses add little to the cognitive load of the learners, and may actually aid the task of comprehension by breaking up utterances into smaller meaning chunks (p. 369). These natural features of spoken language can therefore be introduced in EFL textbooks more often even from a very early stage without affecting the difficulty of the texts. In addition, as mentioned in 5.4.4, some instances of pauses found in the Taiwanese participants' discourse seem to be their L1 equivalents, such as *a ya, ei*, etc. These may sound very unnatural when they are embedded in English conversation. It seems that if textbook dialogues deal with the natural discourse features more often, learners would probably adopt the pauses that sound natural in English instead of their L1 equivalents.

7.4.4 Discourse marking

Discourse markers (DMs) investigated here fall into four categories: interpersonal, referential, structural and cognitive DMs. Table 7.6 presents the four most common items in each type of DM in TETCOC and BATTICC-F respectively. In the comparison of interpersonal DMs in the two datasets, the most common DMs identified are quite different. For example, *yeah* (rank 1 in BATTICC-F) is predominant in naturally occurring discourse (16.86 per 1000 words), while it only occurs with a frequency of 0.98 per 1000 words in TETCOC. Also, *sort of* and *you know* are very common in BATTICC-F but they cannot be found at all in

TETCOC.

Table 7.6

Discourse Markers in TETCOC and BATTICC-F

DMs	TETCOC		DMs	BATTICC-F	
	Number	per 1000 words		Number	per 1000 words
Interpersonal DMs	**132**	**5.38**		**392**	23.20
oh	67	2.73	*yeah*	285	16.86
great	27	1.10	*oh*	68	4.02
yeah	24	0.98	*sort of*	23	1.36
all right	14	0.57	*you know*	16	0.95
Referential DMs*	**319**	**13.0.**		**522**	**30.89**
But	150	6.11	*and*	237	14.02
And	134	5.46	*but*	110	6.51
because	25	1.02	*so*	108	6.39
So	10	0.41	*coz/because*	67	3.96
Structural DMs	**163**	**6.64**		**208**	**12.31**
okay/OK	69	2.81	*so*	92	5.44
then	49	2.00	*okay/OK*	69	4.08
how about	31	1.26	*then*	34	2.01
So	14	0.57	*right*	13	0.77
Cognitive DMs	**46**	**1.87**		**278**	**16.45**
well	34	1.39	*like*	213	12.60
I think	12	0.49	*I think*	37	2.19
			well	18	1.07
			you know	10	0.59
Total	**660**	**26.89**		**1400**	**82.85**

* The item which has a higher frequency in BATTICC-O than in BATTICC-F

In the use of referential DMs, the most frequent items in TETCOC and BATTICC-F are the same, namely *and, but, so* and *because*. In these four items, only *but* occurs with a similar frequency in the two datasets, while the others are presented at an extremely low percentage in TETCOC. In addition, *and*, which has been shown to be one of the most frequent DMs in informal speech, only occurs at a rate of 5.46 per 1000 words in TETCOC, while in BATTICC-F 14.02 instances per 1000 words can be found.

In structural DMs, *so, okay* and *then* are commonly used in both TETCOC and BATTICC-F. However, the numbers of *so* in the two datasets differ largely with frequencies of 0.57 and 5.44 per 1000 words respectively. This indicates that the high-frequency structural DM *so* is not widely employed in textbook conversation. With regard to cognitive DMs, it is apparent from the table that only two items serving this function are found (i.e., *well, I think*) in TETCOC. Also, the two common DMs in BATTICC-F *like* and *you know* are not used at all in TETCOC.

7.5 Mind the gap: TETCOC vs. BATTICC-F

In 7.3 and 7.4 the most striking result to emerge from the study is that a number of marked differences in the use of multi-word sequences and spoken grammar in textbook dialogues (TETCOC) and naturally occurring discourse (BATTICC-F) were identified. Such a gap between the two corpora may be due to a number of different factors. Firstly, the development of teaching/learning materials often depends largely on materials writers' experience and intuitions, rather than actual evidence of language use (Carter et al., 2011; O'Keeffe et al., 2007; Reppen, 2010; Tono, 2011). EFL textbooks therefore would not be able to present the most important and frequent linguistic items used in authentic communication. As a result, corpus linguists have proposed that the authentic data in a corpus can provide an empirical basis for language description by showing how language is actually used in natural contexts. By bringing to light features of language use that have eluded intuition, corpus data and its interrogation can also help illustrate that syllabus and teaching materials design in English language teaching could be dramatically improved (ibid.). Accordingly, an increasing number of corpus-informed ELT products have been published. For example, the *COBUILD*

English dictionary developed in 1987 was the first ELT material based on a corpus, illustrating how words are used in authentic context by providing corpus evidence from attested language use data. Another famous example is *Touchstone* (McCarthy, McCarten, & Sandiford, 2006a and b), which utilised the Cambridge International Corpus of North American English as a *touchstone* to ensure that the language presented in each lesson was authentic and useful. It appears that a corpus-informed approach has gradually changed foreign language learning and materials design. It might be argued, however, that this use of corpora in ELT should be more finely tuned to specific learners' needs, and for this further corpus evidence should be exploited.

Another reason that the Taiwanese textbooks do not present the multi-word sequences of discourse devices can be explained by the fact that some sequences including hesitation devices (e.g., *erm I think*) or discourse makers *like* or *sort of* (e.g., *sort of like, sort of thing, like you know*) may make the textbook dialogues look slightly messy. However, as has been discussed, these markers have very important discourse functions, such as organising the utterances, which may actually aid comprehension (Gilmore, 2004, p. 369), indicating turn-taking (Carter & McCarthy, 2006), helping speakers keep the floor while formulating their next utterance, or in some cases indicating that they are ready to relinquish the floor. Moreover, they have important interpersonal functions (O'Keeffe et al., 2007), which are highly relevant to successful interaction in an informal communication setting. As a result, it seems reasonable for pedagogical materials to include these important multi-word sequences that commonly occur in authentic data. Although including them may make textbook dialogues not as neat and tidy as most of the published textbooks, it presents the actual use of the language and authentic

situations of intercultural communication. This would not only enrich the description of the target language use, but would also increase novices' awareness of the patterns of use in an authentic communication context (Wood, 2010); as Gilmore (2004) suggests, "if our learners' goal is to be able to operate independently in the L2 outside the classroom, then at some point they have to be shown the true nature of conversation" (p. 371).

On the other hand, many of the natural features identified in this study are often considered as representative of one particular group of native speaker, and consequently there are some criticisms from the perspective of World Englishes or English as a Lingua Franca (ELF) (e.g. Jenkins, 2009; Rajagopalan, 2004), which denies any need for specifically native-speaker norms as learners who use them may well risk adopting a false identity. Although there has been considerable debate on whether to use native-speaker models in the EFL classroom, learners still need models of some kind as a point of reference (Cullen & Kuo, 2007). Research on students' perspectives (e.g., Timmis, 2002) shows that learners across a diverse range of countries and contexts of language use have a strong desire to conform to native speaker norms of English. This notwithstanding, I would argue that teaching and learning the natural features of authentic communication does not exactly mean training our learners to speak like a native speaker of English. McCarthy (2012) notes that some of these language features relate to the shared knowledge of all mature, aware human beings, and therefore the aim of learning them is to make learners aware of how human beings interact with each other and how they create and sustain successful interactions. To that end, these spoken features of natural occurring conversation are crucial.

7.6 Summary

This chapter has demonstrated substantial differences in the use of three-word sequences and spoken grammar between textbook dialogues (TETCOC) and authentic intercultural discourse (BATTICC-F). In section 7.2 it was clearly shown that the high-frequency sequences in the two datasets serve three central discourse functions: social interaction, necessary topics and discourse devices. The key sequence analysis has also revealed the underused or overused sequences identified in TETCOC when compared with BATTICC-F. In the category of social interaction, both TETCOC and BATTICC-F cover a good variety of speech acts, while the speech acts of directives (e.g., questioning, requesting, commanding) are significantly overused in comparison with BATTICC-F. In terms of the multi-word sequences indicating necessary topics, location markers are significantly overused in TETCOC, while topics such as food, likes and schools, which are commonly discussed in the intercultural exchange, are comparatively underused. The most appreciable difference, however, is between the use of multi-word sequences as discourse devices. All types of items within this category reach a highly significant difference between the two corpora, including linking devices, fluency devices, exemplifiers and evaluators.

Section 7.3 has reported on an investigation into the spoken grammar used in TETCOC and BATTICC-F. The features examined include five aspects: (1) situational ellipsis, (2) vagueness and approximation, (3) headers and tails, (4) pausing, repeating and recasting and (5) discourse marking. Highly significant differences between the two datasets have been found in all aspects of linguistic features examined. Such a gap between them may well reflect the textbook

writers' perceptions regarding teachability and learnability, and the presentation of the scripts in the textbook. However, this comes at the expense of exemplifying interactionally salient features of veridical discourse to Taiwanese language learners. Excluding these important patterns of spoken discourse may make textbook conversation and, as a consequence, students sound unnatural and perhaps pedantic and bookish. O'Keeffe et al. (2007) note that most of the spoken language features that have also been discussed here have important interpersonal functions that serve to "create and maintain a good relationship between the speaker and hearer" (p. 159). For EFL learners with an intention to maintain a good relationship in face-to-face conversation, it would therefore be very helpful for students to be aware of and learn these features. As such, it is suggested that EFL pedagogical materials expose learners to authentic language use to some extent, including the important formulaic patterns of spoken English in the syllabus. Carter et al. (2011) suggest that "not to provide opportunities for exposure to language use is to take away choices from both teachers and learners" (p. 90). However, the elements of spoken grammar need to be carefully selected according to pedagogic judgments of learners' needs and abilities. For example, some features, such as high-frequency discourse markers (e.g., *right, great, like, I think*) and planning devices (e.g., *er, well*), can be usefully included in a syllabus without dramatically increasing the difficulty of the texts. Other features, such as hedging or vagueness, can be introduced when learners have some basic knowledge of spoken grammar. Since the Taiwanese junior high school students have received four to six years of formal EFL instruction in elementary schools, most of them should have learned about basic English grammar. Junior high school is therefore a good point in time at which to introduce and expose learners

to authentic language use and at the same time raise their consciousness of particular multi-word sequences that feature different registers of language use.

Corpora and results of research based on naturally occurring intercultural discourse have not yet exerted a strong influence on EFL textbooks, and syllabus and teaching/learning materials design could be dramatically improved by a corpus-informed approach accordingly (Carter et al., 2011; O'Keeffe et al., 2007; Reppen, 2010; Tono, 2011). Based on the findings of this thesis, I have developed three sample materials (see Appendices D, E and F) demonstrating how authentic data from BATTICC (e.g., concordance lines) can be used to inform EFL instruction and materials development for intercultural communication, which may further help to bridge the gap between classroom English and naturally occurring discourse. The next chapter will draw conclusions from the thesis and illustrate some limitations and a number of future directions for future applied linguistic research into intercultural discourse.

CHAPTER 8

Conclusion

This thesis sheds light on linguistic patterns in adolescent intercultural communication via a case study of online and spoken interaction between a group of British and Taiwanese participants. Corpus-based studies of naturally occurring discourse in a specific context such as this one reveal the particular patterns of language use in different modes of communication. Based on a newly developed specialised corpus, BATTICC, overall the thesis has provided a detailed description of lexical, grammatical, discourse and pragmatic features of adolescent online and spoken discourse. Specifically, this thesis set out to answer six research questions from three perspectives: keyness approach, a discourse analytical perspective and a multi-word-unit perspective, and each of the research questions will now be revisited in turn:

1. What topics are young people mainly concerned with in online and face-to-face intercultural communication?
2. What are the statistically significant differences in the use of lexical and grammatical categories between Taiwanese and British participants?

Chapter 4 addressed the first two research questions on the basis of keyness approach, which works on the statistical comparisons of frequency information and further highlights the primary elements that are characteristic of a specific collection of texts. The analysis brings together three levels of keyness techniques: keywords, key semantic domains and parts-of-speech. Extending keyword analysis to the levels of POS and semantic domains reduces the number of key items that

the researcher needs to examine. The semantic domain analysis revealed the themes that young people discussed commonly on the online discussion board and face-to-face interaction. For example, the categories of Education in general (P1), Food (F1) and Like (E2+) are found with a high frequency in both BATTICC-O and BATTICC-F, as compared with reference corpora of general communication. This further indicates that these three topics are popular in online communication and face-to-face interaction. Other topics, such as Entertainment (K1), Personal relationship (S3.1) and Music and related activities (K2), are particularly popular in the CMC, while the topics of Happy (E4.1+), Geographical names (Z2) and Weather (W4) were frequently discussed when the participants met face-to-face. In addition, the lexical choices within each category reflect a great number of cultural and social differences. The participants can therefore be encouraged to observe these culturally relevant lexical features used by the other groups and further develop their intercultural awareness and the skills of online and spoken communication.

Keyness at a POS level demonstrated the significantly overused and underused grammatical categories by Taiwanese learners as compared to the discourse of British participants. The analysis of BATTICC-O and BATTICC-F revealed that the appropriate use of modal verbs and tenses are the most problematic areas for the Taiwanese learners in CMC and spoken communication respectively. In this way, the keyness approach identifies those linguistic features that deserve further attention, providing a reasoned basis for drawing learners' awareness to linguistic features specific to the target text (Rayson, 2008; Tribble, 2000). Nevertheless, although quantitative data derived from statistically robust frequency and keyness measures is considered more objective in the sense that decisions on which

linguistic features to study are made on the basis of information mechanically extracted from the data itself (Adolphs, 2006; Rayson, 2008; Scott, 2010), the application of frequency and keyword lists nonetheless entails a degree of subjectivity and selectivity (Harvey, 2008, p. 266-67). Stubbs (2005) also notes the subjective selection and interpretation, although the data is automatically extracted from corpus analysis tools. In addition, computer automatic annotation is not 100% accurate. In this case, using the *WMatrix* web-based tool, for example, accuracy rates quoted for the POS tagger are 96–97% (Leech & Smith, 2000) and 91% for the semantic tagger (Rayson et al., 2004). Careful manual checking of concordances and interpretation of results are therefore required.

While the keyness technique has proven to be useful in identifying statistically significant differences in language use between different groups of participants, a discourse analytical point of view adds greater detail and depth of description of language used in different communication modes, with analysis pursuing the following question:

3. What are the distinctive linguistic features of online and spoken discourse by adolescents? To what extent do the British and Taiwanese participants employ them in intercultural communication? To what extent does spoken grammar exist in online discourse?

Chapter 5 has demonstrated that considerable insight can be gained using the discourse analytical approach, showing how particular linguistic features are employed in CMC and spoken communication. It first concentrated quantitatively on the primary linguistic features of online and spoken discourse by the two

groups of participants. The linguistic patterns were then examined in context to identify the pragmatic and discourse functions. Such findings that pertain to discourse and pragmatic functions in context are not likely to be revealed when only keyness is examined. In the analysis of BATTICC-O, the language usage contains a wealth of cues in CMC in the form of capitalisation/minusculisation, nonconventional spelling, emoticons and punctuation omission and repetition, with a great deal of variation within each category. The examination of these remarkable features shows that they are not simply employed indiscriminately, demonstrating different preferences by participants for different purposes. It was also evident that Taiwanese learners increasingly employed CMC features during the exchange programme in that they noticed how these features were employed by their international peers and gradually adapted their behaviour to match the online community. These features have been shown to exhibit important emotional and interpersonal functions, which facilitate online communication. As noted by Herring (2013), language change is being affected by the Internet and technology, and if anything, these CMC features enrich rather than impoverish language users and languages themselves.

Chapter 5 has also investigated the language use in BATTICC-F and demonstrated a number of distinctive features of spoken discourse in both Taiwanese and British data. It demonstrated that British participants generally produce more instances of vague categories, approximation, hedging and discourse marking, especially hedging and vague categories, which reach a highly significant level. On the other hand, situational ellipsis, headers and tails, pauses, repeating and recasting were more commonly used by Taiwanese students. As has been discussed, some features of spoken grammar that the Taiwanese students never used have very

important discourse and relational functions, which may actually aid comprehension and are highly relevant to successful interaction in an informal communication setting (O'Keeffe et al., 2007, 2011). In addition, the examination of these remarkable features in CMC and spoken discourse displays a high level of informal interaction. While many of these distinctive features do not contribute any specific content or propositions, they have important interpersonal functions and particularly appeal to young people.

One important finding of the discourse analysis in Chapter 5 is that many multi-word patterns of language use are as frequent as or more frequent than the single-word lexical items. The third approach of this thesis therefore focused particularly on the recurrent sequences used in two different communication modes, addressing the following research questions, which have been presented in Chapter 6:

4. What are the high-frequency recurrent multi-word sequences in intercultural communication? Do they serve certain pragmatic functions in the context?
5. To what extent does the use of multi-word sequences by Taiwanese learners develop over time in the one-year intercultural exchange?

With regard to the functional use of multi-word sequences, the thesis specifically examined three-word units, and it was evident that the high-frequency three-word sequences in BATTICC-O, BATTICC-F and reference corpora of online and spoken discourse serve three central discourse functions: social interaction (e.g., questioning, complying, responding), necessary topics (e.g., autobiography, time,

location, quantity) and discourse devices (e.g., fluency devices, exemplifiers, evaluators). This therefore shows that three-word sequences often perform systematic discourse functions, even though they do not usually constitute complete grammatical or idiomatic structures. They function as "important building blocks in discourse" (Biber, 2009, p.284), and accord with interlocutors' expectations and preferences, which may facilitate efficient and effective communication (Schmitt, 2010, 2013; Wray, 2013; Wood, 2010). However, since only the 50 most common three-word sequences retrieved from corpora were examined in detail, a number of sequences that may be unique to this particular intercultural setting but which have a lower frequency may therefore be neglected. Future research can also consider analysing these context-specific, lower-frequency three-word items or those used by other second language learners/speakers. A larger unit, such as four or more words in a sequence, could also be considered in future studies if larger sizes of corpora are compiled.

In addition, it was found that three-word sequences employed in CMC and FTF conversation were significantly different. The category of social interaction was generally very frequently used in both BATTICC-O and BATTICC-F, while the three-word units employed in the area of necessary topics were particularly common in CMC and the sequences functioning as discourse devices were extremely common in spoken communication. It was also apparent that the analysis of highly recurrent three-word sequences demonstrates their important discourse, pragmatic and interpersonal functions, while a large number of them used by British participants cannot be found in Taiwanese learner discourse. For example, in BATTICC-O, three-word units including modal verb *would* (e.g., *it would be*) and coordinating conjunctions (e.g., *and it was*) are significantly

underused by Taiwanese participants. In BATTICC-F, the recurrent three-word sequences *I think it's* and *I think I*, indicating the pervasive use of *I think* as a hedge modifying evaluation of situations or assertions to make them less assertive, cannot be found in Taiwanese learner discourse. Moreover, vague exemplifiers (e.g., *sort of thing, things like that*) have been found to be particularly distinctive features of spoken discourse by British participants, while the Taiwanese learners rarely use them.

The fifth research question concentrates on the developmental aspect of multi-word sequences, examining the extent to which intercultural exposure to native English speakers affects the longitudinal development of the use of multi-word sequences by young Taiwanese learners over one year of contact. Pre- and post-tests show a significant difference in scores for experimental and control groups in both a single-word test (p=.027) and a multi-word test (p=.031) after the treatment, with an eta squared value of .20 and .36 respectively. In addition, the increase of overlap in the high-frequency sequences and the decline of key sequences show an increasingly higher level of approximation between the use of three-word sequences by Taiwanese and British native speakers of English. Assessing the development of sequences therefore provides another way of evaluating the success of intercultural contact as a language learning method. These findings contribute additional evidence that intercultural interaction with native speakers of English can achieve positive learning results for young EFL learners. The method here, which focused on naturally occurring language output, diminishes the effects of the artificial contexts often created in language testing settings.

The last research question considers the EFL learning materials that the Taiwanese participants use in school due to the potential significance of this research to EFL teaching and learning. Since textbooks constitute the main and perhaps only source of language input that the Taiwanese learners receive, I pursued the following question:

6. To what extent do the EFL textbooks used in Taiwanese junior high schools display the distinctive linguistic features of authentic intercultural discourse? How can corpus evidence support EFL teaching/learning materials development?

Three approaches employed in Chapters 4-6 were applied in the analysis of the EFL textbooks, which was presented in Chapter 7. Substantial differences in the use of multi-word sequences and spoken grammar between the textbook dialogues (TETCOC) and naturally occurring discourse (BATTICC-F) were identified. It appears that the Taiwanese EFL textbooks do not present the most important and frequent linguistic items used in authentic communication, and learners as a result would not be able to learn these interactionally salient features of veridical discourse, which serve important discourse and interpersonal functions that facilitate the creation and maintenance of good relationships among interlocutors (O'Keeffe et al., 2007, p. 159). For EFL learners wanting to maintain good relationships in face-to-face conversation, it would therefore be very helpful for students to be aware of and learn these features. On the basis of the findings of the thesis, Appendices D-F demonstrate how authentic data from BATTICC and the three analytical approaches employed in this thesis can be used to inform EFL instruction and materials development for intercultural communication. Three

sample materials for teaching multi-word sequences, e-grammar and spoken grammar, which can be introduced in EFL classrooms in Taiwanese junior high schools, were developed by me.

In conclusion, this thesis has provided a considerable insight into adolescent online and spoken discourse in an intercultural setting. However, there remain a number of caveats to be noted regarding the present study, most notably that, due to the small size of the specialised corpus BATTICC, the present results are not necessarily generalisable to other intercultural discourse. Additionally, the participants recruited in the present study were teenagers from Taiwan and England, and consequently the findings may not be transferable to the language use by native and non-native speakers of English in general. This notwithstanding, the size and composition of the specialised corpus makes it more manageable for qualitative studies and permits a closer link between the corpus and the contexts of its data in order to understand the discourse and functional features of particular linguistic patterns in CMC and FTF interaction.

Another possible limitation of this study is that it concentrates simply on texts or transcripts of interactions in CMC and FTF intercultural communication; therefore, these written representations of language use might be limited as they only have the provision for presenting data in a single format and provide little opportunity for exploring non-verbal, gestural features of discourse, which are important aspects of understanding intercultural communication. Future research can consider including such multi-modal features in discourse analysis; any study that includes extralinguistic features would allow us to get a fuller picture of the complexities of the particular linguistic patterns in intercultural discourse. As

Adolphs and Knight (2010) suggest, "spoken interaction is essentially multi-modal in nature, featuring a careful interplay between textual, prosodic, gestural and environmental elements in the construction of meaning" (p. 44).

While some possible limitations are recognised, this study has nevertheless illuminated important lexical, grammatical and pragmatic aspects of linguistic patterns by British and Taiwanese adolescents in an intercultural exchange project, and identified the gap between naturally occurring discourse and EFL learning materials. The findings as a result have further pedagogical implications in relation to EFL course design for online and spoken intercultural communication, supporting EFL learners to become successful users of English (SUEs) (Prodromou, 2005) so that they can communicate effectively and appropriately with people who have a different cultural or linguistic background.

REFERENCES

Abrams, Z. I. (2003). The effect of synchronous and asynchronous CMC on oral performance in German. *The Modern Language Journal, 87*(2), 157-167.

Adolphs, S. (2006). *Introducing electronic text analysis. A practical guide for language and literary studies*. New York: Routledge.

Adolphs, S. (2010). Using a corpus to study spoken language. In S. Hunston & D. Oakey (Eds.), *Introducing Applied Linguistics: Concepts and Skills* (pp. 180-187). Oxon: Routledge.

Adolphs, S., Atkins, S., & Harvey, K. (2007). Caught between professional requirements and interpersonal needs: vague language in health care contexts. In J. Cutting (Ed.) *Vague language explored* (pp. 62-78). Basingstoke: Palgrave Macmillan.

Adolphs, S., Brown, B., Carter, R., Crawford, P., & Sahota, O. (2004). Applied clinical linguistics: Corpus linguistics in health care settings. *Journal of Applied Linguistics, 1*, 9-28.

Adolphs, S., & Durow, V. (2004). Social-cultural integration and the development of formulaic sequences. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 107-126). Amsterdam: John Benjamins Publishing Company.

Adolphs, S., & Knight, D. (2010). Building a spoken corpus: what are the basics?. In M. McCarthy & A. O'Keeffe (Eds.), *The Routledge handbook of corpus linguistics* (pp. 38-52). London: Routledge.

Adolphs, S., & Lin, P. (2010). Corpus Linguistics. In J. Simpson (Ed.), *The Routledge handbook of applied linguistics* (pp. 148-56). London: Routledge.

Aijmer, K. (2011). *Well I'm not sure I think...* The use of well by non-native speakers. *International Journal of Corpus Linguistics, 16*(2), 231–254.

Alptekin, C. (2002). Toward intercultural communicative competence in ELT. *ELT Journal, 56*(1), 57-64.

Andersen, G. (1998). The pragmatic marker like from a relevance-theoretic perspective. In A.H. Jucker & Y. Ziv (Eds.), *Discourse markers:*

Descriptions and theory (pp. 147–170). Amsterdam: John Benjamins.

Andersen, G. (2000). *Pragmatic markers and sociolinguistic variation.* Amsterdam: John Benjamins.

Averianova, I. (2012). The language of electronic communication and its implications for TEFL. *Procedia - Social and Behavioral Sciences, 34,* 14-19.

Bachmann, I. (2011). Civil partnership – "gay marriage in all but name": a corpus-driven analysis of discourses of same-sex relationships in the UK Parliament. *Corpora, 6*(1), 77-105.

Baker, P. (2004). Querying keywords: Questions of difference, frequency and sense in keywords analysis. *Journal of English Linguistics, 32*(4), 346-359.

Baker, P. (2006). *Using corpora in discourse analysis.* London: Continuum.

Baron, M. S. (2009). Review of the book *Txtng: The gr8 db8. Language, 85*(4), 914-916.

Belz, J. A. (2003). Linguistic perspectives on the development of intercultural competence in telecollaboration. *Language Learning & Technology, 7*(2), 68–117. Retrieved from http://llt.msu.edu/vol7num2/pdf/belz.pdf

Belz, J. A. (2007). The Development of intercultural competence in online interaction. In R. O'Dowd (Ed.), *Online intercultural exchange: An introduction for foreign language teachers* (pp. 127-166). UTP, USA: Multilingual Matters Ltd.

Berber Sardinha, T. (1999). *Using keywords in text analysis: Practical aspects.* DIRECT Working Papers 42, Pontifical Catholic University of São Paulo: São Paulo. Retrieved from http://www2.lael.pucsp.br/direct/DirectPapers42.pdf

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

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing, 8,* 243-257.

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

Biber, D. (2009). A corpus-driven approach to formulaic language in English.

International Journal of Corpus Linguistics, *14*(3), 275–311.

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

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

Biesenbach-Lucas, S. (2007). Students writing emails to faculty: an examination of e-politeness among native and non-native speakers of English. *Language Learning & Technology*, *11*(2), 59-81. Retrieved from http://llt.msu.edu/vol11num2/biesenbachlucas/

Bolden, G. B. (2009). Implementing incipient actions: The discourse marker 'so' in English conversation. *Journal of Pragmatics*, *41*, 974–998.

Boxer, D. (2002). Discourse issues in cross-cultural pragmatics. *Annual Review of Applied Linguistics*, *22*, 150-167.

British Council (n.d.). *What is Connecting Classroom?* Retrieved from http://www.britishcouncil.org/learning-connecting-classrooms.htm

Brown, H. D. (2007). *Principles of language learning and teaching (5th ed.)*. White Plains, NY: Pearson Longman.

Burnard, L. (2005). Developing linguistic corpora: Metadata for corpus work. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 30-46). Oxford: Oxbow Books.

Byram, M. (1997). *Teaching and Assessing Intercultural Communicative Competence*. Clevedon, UK: Multilingual Matters.

Byram, M., Gribkova, B., & Starkey, H. (2002). *Developing the intercultural dimension in language teaching: A practical introduction for teachers*. Strasbourg: Council of Europe.

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

Canale, M., & Swain, M. (1980). Theoretical bases of communicative approaches to second language teaching and testing. *Applied Linguistics, 1*, 1-47.

Carter, R. (1998). Orders of reality: CANCODE, communication and culture. *ELT Journal*, *52*, 43-56.

Carter, R. (2004). *Language and creativity: the art of common talk*. London: Routledge.

Carter, R. (2008). Right, well, OK, so, it's like, you know, isn't it, I suppose: spoken words, written words and why speaking is different. In C. Hudson (Ed.), *The sound and the silence: key perspectives on speaking and listening and skills for life* (pp. 11-23). Coventry: Quality Improvement Agency.

Carter, R., Hughes, R., & McCarthy, M. (2011). Telling tails: grammar, the spoken language and materials development. In B. Tomlinson (Ed.), *Materials development in language teaching* (2nd edition) (pp.78-100). Cambridge: Cambridge University Press.

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

Celce-Murcia, M., & Larsen-Freeman, D. (1999). *The grammar book: An ESL/EFL teacher's course*. Boston: Heinle and Heinle.

Chang, T. H. (2003). *An analysis of the use of modal verbs in senior high school students' compositions.* Unpublished master's dissertation, NCCU, Taipei.

Chang, L. Y. (2010). *Developing intercultural communicative competence through weblogs: the case of three elementary school students.* Unpublished master's dissertation, NCCU, Taipei.

Channell, J. (1994) *Vague language*. Oxford: Oxford University Press.

Chen, L. (2010a). An Investigation of lexical bundles in ESP textbooks and electrical engineering introductory textbooks. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 107-128). London and New York: Continuum.

Chen, X. (2010b). *Discourse-grammatical Features in L2 Speech: A Corpus-based contrastive study of Chinese advanced learners and native speakers of English* (Unpublished doctoral thesis). City University of Hong Kong, Hong Kong.

Chen, Y. H., & Baker, P. (2010). Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology, 14*(2), 30-49. Retrieve from

Cheng, W. (2012). Speech acts, facework, and politeness: Relationship-building across cultures. In J. Jackson (Ed.), *Routledge handbook of language and intercultural communication* (pp. 148-163). London: Routledge.

Cheng, W., & Warren, M. (2003). Indirectness, inexplicitness and vagueness made clearer. *Pragmatics, 13*(3), 381-400.

Cho, T. (2010). Linguistic features of electronic mail in the workplace: A comparison with memoranda. *Language@Internet, 7*, article 3. Retrieve from http://www.languageatinternet.de/articles/2010/2728/index_html/

Chou, C. T. (Ed.) (2010). *Kang-Xuan Junior High School English Book 1-6.* Taipei: Kang-Xuan.

Clark, C., & Dugdale, G. (2009). *Young people's writing: attitudes, behaviour and the role of technology*. London: National Literacy Trust in collaboration with Booktrust.

Conklin, K., & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics, 29*(1), 72–89.

Coulmas, F. (1979). On the sociolinguistic relevance of routine formulae. *Journal of Pragmatics, 3*, 239–66.

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, teaching, assessment*. Cambridge: Cambridge University Press.

Council of Europe. (2008). *White Paper on Intercultural Dialogue*. Strassbourg: Council of Europe.

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

Crossley, S., & Salsbury, T. L. (2011). The development of lexical bundle accuracy and production in English second language speakers. *International Review of Applied Linguistics in Language Teaching, 49*(1), 1-26.

Crystal, D. (2006). *Language and the Internet* (2nd ed.). Cambridge, England: Cambridge University Press.

Crystal, D. (2010). Internet language. In Cummings, L. (Ed.), *The Pragmatics Encyclopaedia* (pp. 234-6). London: Routledge.

Crystal, D. (2011). *Internet linguistics: a student guide*. Oxen and New York: Routledge.

Cullen, R., & Kuo, I. (2007). Spoken grammar and ELT course materials: A

missing link? *TESOL Quarterly, 41*(2), 361-386.

Culpeper, J. (2009). Keyness: Words, parts-of-speech and semantic categories in the character-talk of Shakespeare's Romeo and Juliet. *International Journal of Corpus Linguistics, 14*(1), 29–59.

Cutting, J. (2011). Spoken discourse. In K. Hyland & B. Paltridge (Eds.), *Continuum companion to discourse analysis* (pp. 155-170). London and New York: Continuum.

Davis, B. H., & Brewer, J. P. (1997). *Electronic discourse: linguistic individuals in virtual space*. Albany, NY: State University of New York Press.

Davis, B., & Thiede, R. (2000). Writing into change: Style-shifting in asynchronous electronic discourse. In M. Warschauer & R. Kern (Eds.), *Network-based language teaching: Concepts and practice* (pp. 87–120). Cambridge: Cambridge University Press.

De Cock, S. (2004). Preferred sequences of words in NS and NNS speech. *Belgium Journal of English and Literatures (BELL)*, New Series 2, 225–246.

Department for Education and Skills. (2004). *Putting the world into world-class education*. London: DfES Publications.

Department for Education and Skills. (2005). *Developing a global dimension in the school curriculum*. London: DfES Publications.

Department for Education. (November 2012). *What is the research evidence on writing?* (Ref: DFE-RR238). Retrieved from https://www.education.gov.uk/publications/eOrderingDownload/DFE-RR238.pdf

Dietz-Uhler, B., & Bishop-Clark, C. (2001). The use of computer-mediated communication to enhance subsequent face-to-face discussions. *Computers in Human Behavior, 17*, 269–283.

Dickinson, M. (2005). Error detection and Correction in annotated corpora. Retrieved from http://jones.ling.indiana.edu/~mdickinson/papers/diss/dickinson05-alt.pdf.gz

Dresner, E., & Herring, S. C. (2010). Functions of the non-verbal in CMC: Emoticons and illocutionary force. *Communication Theory, 20*, 249-268.

Dresner, E., & Herring, S. C. (2012). Emoticons and illocutionary force. In D. Riesenfel & G. Scarafile (Eds.), *Philosophical dialogue: Writings in honor*

of Marcelo Dascal (pp. 59-70). London: College Publication.

Dooly, M., & O'Dowd, R. (2012). *Researching online foreign language interaction and exchange: Theories, methods and challenges.* Pieterlen: Peter Lang.

Dörnyei, Z. (2007). *Research methods in applied linguistics.* Oxford: Oxford University Press.

Dubois, S. (1992). Extension particles etc. *Language Variation and Change, 4*(2), 179–205.

Dunning, T. (1993). Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistic, 19*(1), 61-74.

Ellis, R. (2008). *The study of second language acquisition (2nd ed.).* Oxford: Oxford University Press.

Ellis, N., Simpson-Vlach, R., & Maynard, C. (2008). Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly, 42*(3), 375-396.

Elmer-Dewitt, P. (1994). Bards of the Internet. Time, 4 July, 66-67.

Erman, B., & Warren, B. (2000). The idiom principle and the open-choice principle. *Text, 20*, 29–62.

Evert, S. (2008). Corpora and collocations. In A. Lüdeling & M. Kytö (Eds.), *Corpus linguistics: An international handbook (Volume 2)* (pp. 1212-1248). Berlin: Mouton de Gruyter.

Evison, J. (2008). *Turn-openers in academic talk: An exploration of discourse responsibility* (Unpublished doctoral thesis). University of Nottingham, UK.

Evison, J., McCarthy, M., & O'Keeffe, A. (2007). Looking out for love and all the rest of it: vague category markers as shared social space. In J. Cutting (Ed.), *Vague Language Explored* (pp. 138-157). Basingstoke: Palgrave Macmillan.

Fantini, A. E. (2000). A central concern: Developing intercultural competence. Retrieved from http://www.sit.edu/publications/docs/competence.pdf

Fernandez, J., & Yuldashev, A. (2011). Variation in the use of general extenders and stuff in instant messaging interactions. *Journal of Pragmatics, 43*, 2610-2626.

Flowerdew, L. (2004). The argument for using English specialized corpora to

understand academic and professional language. In U. Connor & T. A. Upton (Eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*. Amsterdam, NLD: John Benjamins.

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Language tasks: Teaching, learning and testing* (pp. 74–93). Harlow, England: Longman.

Fox Tree, J. E., & Schrock, J. C. (2002). Basic meanings of *you know* and *I mean*. *Journal of Pragmatics, 34*, 727–747.

Fraser, B. (1996). Pragmatic markers. *Pragmatics, 6*(2), 167–190.

Fraser, B. (1999). What are discourse markers? *Journal of Pragmatics*, 31, 931–952.

Frehner, C. (2008). *Email, SMS, MMS : the linguistic creativity of asynchronous discourse in the new media age*. Bern: Peter Lang.

Fuller, J. M. (2003). Use of the discourse marker *like* in interviews. *Journal of sociolinguistics, 7*, 365–377.

Fung, L., & Carter, R. (2007). Discourse Markers and Spoken English: Native and Learner Use in Pedagogic Settings. *Applied Linguistics, 28*(3), 410–439.

Garrison, A., Remley, D., Thomas, P., & Wierszewski, E. (2011). Conventional Faces: Emoticons in Instant Messaging Discourse. *Computers and Composition, 28*, 112–125.

Garside, R., & Smith, N. (1997). A hybrid grammatical tagger: CLAWS4. In R. Garside, G. Leech & A. McEnery (Eds.), *Corpus annotation: Linguistic information from computer text corpora* (pp. 102–121). London: Longman.

Gavioli, L. (2005). *Exploring corpora for ESP learning*. The Netherlands: John Benjamins Publishing.

Greaves, C., & Warren, M. (2010). What can a corpus tell us about multi-word units? In M. McCarthy & A. O'Keeffe (Eds.), *The Routledge Handbook of Corpus Linguistics* (pp. 212-226). London: Routledge.

Grinter, R. E., Palen, L., & Eldridge, M. (2006). Chatting with teenagers: Considering the place of chat technologies in teen life. *ACM Transactions on Computer-Human Interaction, 13*, 423–447.

Gilmore, A. (2004). A comparison of textbook and authentic interactions. *ELT Journal*, *58*(4), 363-374.

Guilherme, M. (2000). Intercultural competence. In Byram, M. *Routledge Encyclopedia of language teaching and learning* (pp. 298-300). London and New York: Routledge Taylor and Francis Group.

Hall, E. T. (1959). *The Silent Language*. New York: Doubleday.

Hall, E. T., & Trager, G. L. (1953). *The analysis of culture*. Washington, DC: Foreign Service Institute/American Council of Learned Societies.

Hakuta, K. (1974). Prefabricated patterns and the emergence of structure in second language acquisition. *Language Learning*, *24*, 287-298.

Han, Z., Park, E. S., & Combs, C. (2008). Textual enhancement of input: issues and possibilities. *Applied Linguistics*, *29*(4), 597–618.

Hanna, B., & de Nooy, J. (2003). A funny thing happened on the way to the forum: electronic discussion and foreign language learning. *Language Learning & Technology*, *7*(1), 71–85. Retrieved from http://llt.msu.edu/vol7num1/pdf/hanna.pdf

Harris, R. & Paradice, D. (2007). An investigation of the computer-mediated communication of emotions. *Journal of Applied Sciences Research*, *3*(12), 2081-2090.

Harvey, K. (2008). *Adolescent health communication: A corpus linguistics approach* (Unpublished doctoral thesis). University of Nottingham, UK.

Hauck, M. (2007). Critical success factors in a TRIDEM exchange. *ReCALL*, *19*(2), 202–223.

Hellermann, J., & Vergun A. (2007). Language which is not taught: The discourse marker use of beginning adult learners of English. *Journal of Pragmatics*, *39*, 157–179.

Heritage, J. (1984). A change-of-state token and aspects of its sequential placement. In A. J. Maxwell, & J. Heritage (Eds.), *Structures of social action. Studies in conversation analysis* (pp. 199–345). Cambridge: Cambridge University Press.

Herring, S. C. (2003). Computer-mediated discourse. In D. Schiffrin, D. Tannen, & H. Hamilton (Eds.), *Handbook of discourse analysis* (pp. 612-634). Oxford: Blackwell.

Herring, S. C. (2004). Computer-mediated discourse analysis: An approach to researching online communities. In S. A. Barab, R. Kling, & J. H. Gray (Eds.), *Designing for virtual communities in the service of learning* (pp. 338-376). Cambridge and New York: Cambridge University Press.

Herring, S. C. (2012). Grammar and electronic communication. In C. A. Chapelle (Ed.) *The Encyclopedia of applied linguistics* (pp. 2338-2346). Oxford: Wiley-Blackwell.

Hill, J. (2001). Revising priorities: From grammatical failure to collocational success. In M. Lewis (Ed.): *Teaching collocation: Further development in the lexical approach* (pp. 47–69). Hove, East Sussex: LTP.

House, J. (2009). Subjectivity in English as Lingua Franca discourse: The case of *you know*. *Intercultural Pragmatics, 2*, 171–193.

Huang, Y. P. (Ed.) (2010). *Han-Lin Junior High School English Book 1-6.* Taipei: Han-Lin.

Hung, Y., Huang, H., Chou, Y., Liu, Y., Lin, C., & Hsieh, L. (2006). *English word reading test.* Taipei: Psycho Press.

Hunston, S. (2002). *Corpora in applied linguistics.* Cambridge: Cambridge University Press.

Jeng, S. N. (2010). *The transferability from synchronous computer-mediated communication to oral discussion* (Unpublished master dissertation). National Tsing Hua University, Taiwan.

Jackson, J. (Ed.) (2012). *The Routledge handbook of language and intercultural communication.* London: Routledge.

Jenkins, J. (2009). *World Englishes: a resource book for students (2nd edition).* London: Routledge.

Jia, G. (2003). The acquisition of the English plural morpheme by native Mandarin Chinese-speaking children. *Journal of Speech, Language, and Hearing Research, 46*, 1297-1311.

Johns, T. (2002). Data-driven learning: the perpetual challenge. In B. Kettemann & G. Marko (Eds.), *Teaching and learning by doing corpus linguistics* (pp. 107-117). Amsterdam: Rodopi.

Jucker, A. H., & Smith, S. W. (1998). And people just you know like 'wow': discourse markers as negotiating strategies. In A. H. Jucker & Y. Ziv (Eds.),

Discourse markers. Descriptions and theory (pp. 171–201). Amsterdam: Benjamins.

Kalman, Y. M., & Gergle, D. (November 2009). *Letter and punctuation mark repeats as cues in computer-mediated communication.* Paper presented at the 95th annual meeting of the National Communication Association in Chicago, IL.

Kalman, Y. M., & Gergle, D. (September 2010). *CMC cues enrich lean online communication: The case of letter and punctuation mark repetitions.* Paper presented at the 5th Mediterranean Conference on Information Systems in Tel-Aviv, Israel.

Kjellmer, G. (1994). *A dictionary of English collocations.* 3 Vols. Oxford: Clarendon Press.

Kjellmer, G. (2003). Hesitation. In defence of *er* and *erm*. *English Studies, 84*(2), 170–198.

Kabata, K., & Edasawa, Y. (2011). Tandem language learning through a cross-cultural keypal project. *Language Learning & Technology, 15*(1), 104-121. Retrieved from http://llt.msu.edu/issues/february2011/kabataedasawa.pdf

Kennedy, G. (2002). Variation in the distribution of modal verbs in the British National Corpus. In R. Reppen, S. Fitzmaurice & D. Biber (Eds.), *Using corpora to explore linguistic variation* (pp. 73-90). Philadelphia: John Benjamins.

Koester, A. (2006). *Investigating Workplace Discourse*. London: Routledge.

Koester, A. (2007). "About twelve thousand or so": Vagueness in north American and UK offices. In Cutting, J. (Ed.), *Vague language explored* (pp. 40-61). Basingstoke: Palgrave Macmillan.

Koester, A. (2010). Building small specialised corpora. In M. McCarthy & A. O'Keeffe (Eds.), *The Routledge handbook of corpus linguistics* (pp. 66-79). London: Routledge.

Krashen, S. (1981). The "fundamental pedagogical principle" in second language teaching. *Studia Linguistica, 35*(1), 50-70.

Krashen, S. (1982). *Principles and Practice in Second Language Acquisition.* Oxford: Pergamo.

Kumaravadivelu, B. (2008). *Cultural globalization in language education.* Yale: Yale University Press.

Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology, 5*(3), 37-72. Retrieved from http://llt.msu.edu/vol5num3/pdf/lee.pdf

Leech, G. (2000). Grammars of spoken English: New outcomes of corpus-oriented research. *Language Learning, 50*, 675–724.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English*. Harlow: Longman.

Levinson, S. (1983). *Pragmatics*. Cambridge: Cambridge University Press.

Lewis, C., & Fabos, B. (2005). Instant messaging, literacies, and social identities. *Reading Research Quarterly, 40*, 470–501.

Liaw, M. L. (2006). E-learning and the development of intercultural competence. *Language Learning & Technology, 10*(3), 49-64. Retrieved from http://llt.msu.edu/vol10num3/pdf/liaw.pdf

Liaw, M-L. (2007). Constructing a 'third space' for EFL learners: Where language and cultures meet. *ReCALL, 19*(2), 224–241.

Liaw, M. L., & Master, S. B. (2010). Understanding telecollaboration through an analysis of intercultural discourse. *Computer Assisted Language Learning, 23*(1), 21-40.

Lin, Y. L. (2012). Mind the Gap! Textbook Conversation vs. Authentic Intercultural Interaction. In Y. Leung, K. Cheung, W. Dai, C. Hsiao, & J. Katchen (Eds.), *Selected Papers from the 21st International Symposium on English Teaching* (pp. 42-54). Taipei: Crane Publishing.

Liu, C. G. (Ed.) (2010). *Nan-I Junior High School English Book 1-6.* Taipei: Nan-I.

Lo, S. (2008). The nonverbal communication functions of emoticons in computer-mediated communication. *CyberPsychology & Behavior, 11*, 595-597.

Lu, Q. X. (2003). *The effects of using cross-cultural e-mail exchange projects on developing Taiwanese junior high school students' linguistic skills and cultural awareness: A case study of seventeen junior high school students*

(Unpublished master dissertation). NTNU, Taiwan.

Markey, A. (2007). Using the Moodle platform in online exchanges. In R. O'Dowd (Ed.), *Online intercultural exchange: An introduction for foreign language teachers* (pp. 264-268). UTP, USA: Multilingual Matters.

Martínez, I. M. P. (2011). "I might, I might go I mean it depends on money things and stuff". A preliminary analysis of general extenders in British teenagers' discourse. *Journal of Pragmatics, 43*, 2452-2470.

Matsumoto, D., Yoo, S. H., & LeRoux, J. A. (2009). Emotion and intercultural adjustment. In H. Kottho & H. Spencer-Oatey (Eds.), *Handbook of applied linguistics, Vol. 7 Intercultural Communication* (pp. 77-98). Mouton: de Gruyter Mouton Publishers.

McArthur, T. (Ed.). (1992). *The Oxford companion to the English language.* Oxford: Oxford University Press.

McCarthy, M. (2006). *Explorations in Corpus Linguistics*. Cambridge: Cambridge University Press.

McCarthy, M. (2012). Corpora and spoken language: Ways forward for TESOL. Paper presented at the 2012 TESOL Convention, Boston.

McCarthy, M., & Handford, M. (2004). "Invisible to us": A preliminary corpus-based study of spoken business English. In U. Connor & T. A. Upton (Eds.), *Discourse in the Professions: Perspectives from corpus linguistics* (pp. 167-201). Amsterdam: Benjamins.

McCarthy, M., McCarten, J., & Sandiford, H. (2006a). *Touchstone student's book 3*. Cambridge: Cambridge University Press.

McCarthy, M., McCarten, J., & Sandiford, H. (2006b). *Touchstone student's book 4*. Cambridge: Cambridge University Press.

McCarthy, M., Matthiessen, C. & Slade, D. (2010). What is discourse analysis? In N. Schmitt (Ed.), *An introduction to applied linguistics (2nd Edition)* (pp. 53-69). Oxon: Hodder & Stoughton Ltd.

McEnery, T., Xiao, R., & Tono Y. (2006). *Corpus-based language studies.* London: Routledge.

Meunier, F., & Gouverneur, C. (2009). New types of corpora for new educational challenges: collecting, annotating and exploiting a corpus of textbook material. In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 179-201).

Amsterdam: John Benjamins.

Ministry of Education (2010). *The general guidelines of grades 1-9 curriculum for elementary and junior high school education (New Edition)*. Taipei: Ministry of Education.

Ministry of Education (2011). *White paper on international education for primary and junior high schools (draft)*. Taipei: Ministry of Education.

Miskovic-Lukovic, M. (2009). Is there a chance that I might kinda sort of take you out to dinner?: The role of the pragmatic particles kind of and sort of in utterance interpretation. *Journal of Pragmatics, 41*, 602–625.

Montero, B., Watts, F., & Garcia-Carbonell, A. (2007). Discussion forum interactions: Text and context. *System, 35*, 566-582.

Mujis, D. (2010). *Doing quantitative research in education with SPSS* (2nd ed.). London: Sage.

Mumford, S. (2009). An analysis of spoken grammar: the case for production. *ELT Journal, 63*(2), 137-144.

NAFSA (2007). An international education policy: For U.S. leadership, competitiveness, and security. Retrieved from http://www.nafsa.org/_/File/_/neip_rev.pdf

Nation, I.S.P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Boston, MA: Heinle.

Nattinger, J. R., & DeCarrico, J. S. (1992). *Lexical phrases and language teaching*. Oxford: Oxford University Press.

Nguyen, M. T. (2011). Learning to communicate in a globalized world: To what extent do school textbooks facilitate the development of intercultural pragmatic competence? *RELC Journal, 42*(1), 17–30.

O'Dowd, R. (2007). Evaluating the outcomes of online intercultural exchange. *ELT Journal, 61*(4), 144-152.

O'Dowd, R., & Ritter, M. (2006). Understanding and working with "failed communication" in telecollaborative exchanges. *CALICO Journal, 61*(2), 623–642.

O'Keeffe, A. (2006). *Investigating Media Discourse*. London: Routledge.

O'Keeffe, A. & Adolphs, S. (2008). Using a corpus to look at variational

pragmatics: response tokens in British and Irish discourse. In K. P. Schneider & A. Barron (Eds.), *Variational Pragmatics* (pp. 69-98). Amsterdam: John Benjamins.

O'Keeffe, A., Clancy, B., & Adolphs, S. (2011). *Introducing pragmatics in use.* New York and Abingdon: Routledge.

O'Keeffe, A., McCarthy, M., & Carter, R. (2007). *From Corpus to Classroom: language use and language teaching.* Cambridge: Cambridge University Press.

Ooi, V., Tan, P., & Chiang, A. (2007). Analyzing personal weblogs in Singapore English: the WMatrix approach. *Studies in Variation, Contacts and Change in English*, 2. Retrieved from http://www.helsinki.fi/varieng/journal/volumes/02/ooi_et_al/

Oppenheim, N. (2000). The importance of recurrent sequences for nonnative speaker fluency and cognition. In H. Riggenbach (Ed.), *Perspectives on fluency.* Ann Arbor: The University of Michigan Press.

Östman, J. O. (1981). *You know: A discourse functional approach, Pragmatics and Beyond II: 7.* Amsterdam: Benjamins.

Overstreet, M., & Yule, G. (2002). The metapragmatics of *and everything. Journal of Pragmatics, 34*(6), 785–794.

Oxford Learning. (2006). *Texting VS writing: The problem with instant messaging.* Retrieved from http://www.oxfordlearning.com/letstalk/2006/10/26/texting-vs-writing-the-problem-with-instant-messag/

Perez, L. C. (2003). Foreign language productivity in synchronous versus asynchronous computer-mediated communication. *CALICO Journal, 21*(1), 89-104.

Pew Internet and American Life Project (2009). *Generations online in 2009.* Retrieved from http://www.pewinternet.org/Reports/2009/Generations8Online8in82009.aspx

Plester, B., Wood, C. & Joshi, P. (2009). Exploring the relationship between children's knowledge of text message abbreviations and school literacy outcomes. *British Journal of Developmental Psychology, 27*, 145-161.

Prodromou, L. (2005). *'You see, it's sort of tricky for the L2-user' the puzzle of idiomaticity in English as a Lingua Franca* (Unpublished doctoral thesis). University of Nottingham, UK.

Provine, R. R., Spencer, R., & Mandell, D. (2007). Emotional expression online: Emoticons punctuate website text messages. *Journal of Language and Social Psychology, 26*(3), 299-307.

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

Rajagopalan, K. (2004). The concept of World English and its implications for ELT. *ELT Journal, 58*(2), 111-117.

Rayson, P., Archer, D., Piao, S. L., & McEnery, T. (2004). The UCREL semantic analysis system. In *Proceedings of the workshop on Beyond Named Entity Recognition Semantic labelling for NLP tasks in association with 4th International Conference on Language Resources and Evaluation (LREC 2004)*, 25th May 2004, Lisbon, Portugal (pp. 7–12). Paris: European Language Resources Association.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics, 13*(4), 519-549.

Reber, E. (2010). Interjections in the EFL classroom: teaching sounds and sequences. *ELT Journal, 65*(4), 365-375.

Read, J. (2007). Second language vocabulary assessment: current practices and new directions. *International Journal of English Studies, 7*(2), 105-125

Reppen, R. (2010). *Using corpora in the language classroom.* Cambridge: Cambridge University Press.

Richards, J. (2005). The role of textbooks in a language program. Retrieved from http://www.professorjackrichards.com/work.htm

Richards, J. C., Platt, J., & Platt, H. (1992). *Longman Dictionary of Language Teaching and Applied Linguistics.* Essex: Longman.

Riordan, M. A., & Kreuz, R. J. (2010). Cues in computer-mediated communication: A corpus analysis. *Computers in Human Behavior, 26,* 1806-1817.

Rogers, E., Hart W., & Miike, Y. (2002). Edward T. Hall and The History of Intercultural Communication: The United States and Japan. *Keio*

Communication Review, *24*, 3-26.

Römer, U. (2004). A corpus-driven approach to modal auxiliaries and their didactics. In J. Sinclair (Ed.), *How to Use Corpora in Language Teaching* (pp. 185-199). Amsterdam: John Benjamins.

Römer, U. (2009). Corpus research and practice: What help do teachers need and what can we offer? In K. Aijmer (Ed.), *Corpora and Language Teaching* (pp. 83-98). Amsterdam: John Benjamins.

Rúa, P. L. (2007). Teaching l2 vocabulary through SMS language: Some didactic guidelines. *Estudios de lingüística inglesa aplicada*, *7*, 165-188.

Rosenfeld, B., & Penrod, S.D. (2011). *Research methods in forensic psychology*. NY: John Wiley and Sons.

Sasaki, A. (2010). EFL students' vocabulary learning in NS-NNS e-mail interactions: Do they learn new words by imitation? *ReCALL Journal*, *22*(1), 70–82.

Schiffrin, D. (1987). *Discourse Markers*. Cambridge: Cambridge University Press.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary manual*. London: Palgrave Macmillan.

Schmitt, N. (2013). *Formulaic language and collocation*. In C. A. Chapelle (Ed.), *The Encyclopedia of applied linguistics* (pp. 2190-2200). Oxford: Wiley-Blackwell.

Schmitt, N., & Carter, R. (2004). Formulaic Sequences in Action. In N. Schmitt (Ed.), *Formulaic sequences: Acquisition, processing and use* (pp. 1-22). Amsterdam: John Benjamins.

Schwienhorst, K. (2003). Learner autonomy and tandem learning: Putting principles into practice in synchronous and asynchronous telecommunications environments. *Computer Assisted Language Learning*, *16*, 427-443.

Scott, M. R. (2007). *WordSmith Tools Help Manual*. Version 4.0. Liverpool: Lexical Analysis Software.

Scott, M. R. (2008). *WordSmith Tools Version 5.0*. Liverpool: Lexical Analysis Software.

Scott, M. R. (2010). *WordSmith Tools Help Manual.* Version 5.0. Liverpool: Lexical Analysis Software. Retrieved from www.lexically.net/downloads/version5/HTML/?type_token_ratio_proc.htm

Sercu, L. (2005). *Foreign Language Teachers and Intercultural Competence: An International Investigation.* NY: Multilingual Matters Limited.

Sharwood Smith, M. (1993). Input enhancement in instructed SLA. *Studies in Second Language Acquisition, 15,* 165–79.

Simpson-Vlach, R., & Ellis, N. C. (2010). An academic formulas list: New methods in phraseology research. *Applied Linguistics, 31*(4), 487-512.

Sinclair, J. (1991). *Corpus, concordance, collocation.* Oxford: Oxford University Press.

Sinclair, J. (2001). Preface to small corpus studies and ELT. In M. Ghadessy, A. Henry, & R. Roseberry (Eds.), *Small corpus studies and ELT* (pp. vii– xv). Amsterdam: John Benjamins.

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse.* London and New York: Routledge.

Sinclair, J. (2005). Corpus and text-basic Principles. In M. Wynne (Ed.), *Developing linguistic corpora: A guide to good practice* (pp. 1-16). Oxford: Oxbow Books.

Skopinskaja, L. (2009). Assessing intercultural communicative competence: test construction issues. *Synergies Pays Riverains de la Baltique, 6,* 135-144.

Sternberg, B. J., Kaplan, K. A., & Borck, J. E. (2007). Enhancing adolescent literacy achievement through integration of technology in the classroom. *Reading Research Quarterly, 42,* 416–420.

Stickler, U., & Emke, M. (2011). LITERALIA: Towards Developing Intercultural Maturity Online. *Language Learning & Technology, 15*(1), 147-168. Retrieved from http://llt.msu.edu/issues/february2011/stickleremke.pdf

Stubbs, M. (2005). Conrad in the computer: examples of quantitative stylistic methods. *Language and Literature, 14*(1), 5-24.

Spencer-Oatey, H., & Franklin, P. (2009). *Intercultural interaction: A multidisciplinary approach to intercultural communication.* Basingstoke, UK: Palgrave Macmillan.

Tagliamonte, S. (2005). So who? Like how? Just what? Discourse markers in the conversations of young Canadians. *Journal of Pragmatics, 37*, 1896–1915.

Tagliamonte, S. &, Denis, D. (2010). The 'stuff' of change: general extenders in Toronto, Canada. *Journal of English Linguistics, 38*, 335–368.

Teubert, W. (2005). My version of corpus linguistics. *International Journal of Corpus Linguistics, 10*(1), 1-13.

Thornbury, S., & Slade, D. (2006). *Conversation: From Description to Pedagogy.* Cambridge: Cambridge University Press.

Thurlow, C. (2001). Language and the Internet. In R. Mesthrie & R. E. Asher (Eds.), *The concise encyclopedia of sociolinguistics* (pp. 287-289). Oxford: Elsevier Science Ltd.

Tomlinson, B. (2011). Introduction: principles and procedures of materials development. In B. Tomlinson (Ed.), *Materials development in language teaching (2nd Edition)* (pp. 1-34). Cambridge: Cambridge University Press.

Tono, Y. (2011). TaLC in action: recent innovations in corpus-based English language teaching in Japan. In A. Frankenberg-Garcia, L. Flowerdew, & G. Aston (Eds.), *New trends in corpora and language learning* (pp. 3-25). London and New York: Continuum.

Tottie, G. (2011). *Uh* and *Um* as sociolinguistic markers in British English. *International Journal of Corpus Linguistics, 16*(2), 173–197.

Trabelsi, S. (2010). Developing and trialling authentic Materials for Business English Students at a Tunisian University. In B. Tomlinson (Ed.), *Research for Materials Development in Language Learning* (pp. 103-120). London: Continuum.

Tremblay, A., & Baayen, H. (2010). Holistic processing of regular four-word sequences: a behavioural and ERP study of the effects of structure, frequency, and probability on immediate free recall. In D. Wood (Ed.), *Perspectives on formulaic language: Acquisition and communication* (pp. 151-173). London and New York: Continuum.

Tribble, C. (2002). Corpora and Corpus Analysis: New Windows on Academic Writing. In J. Flowerdrew (Ed.) *Academic Discourse* (pp. 131-149). London: Longman.

Troncoso, C. R. (2010) The effect of language materials on the development of intercultural competence. In B. Timlinson (Ed.), *Research for materials development in language learning: evidence for best practice* (pp. 83-102).

London: Continuum.

Tsui, A. B. M. (1994). *English Conversation*. Oxford: Oxford University Press.

Underhill, R. (1988) *Like* is, like, focus. *American Speech, 63*(3), 234– 246.

Varnhagen, C., McFall, G., Pugh, N., Routledge, L., Sumida-MacDonald, H., & Kwong, T. (2010). lol: new language and spelling in instant messaging. *Reading and Writing, 23*, 719-733.

VOICE Project. (2007). VOICE transcription conventions [2.1]. Retrieved from http://www.univie.ac.at/voice/voice.php?page=transcription_general_information

Vygotsky, L. S. (1962). *Thought and language*. Cambridge: MIT Press.

Walsh, S., Morton, T., & O'Keeffe, A. (2011). Analysing university spoken interaction. *International Journal of Corpus Linguistics, 16*(3), 325-344.

Warschauer, M. (1997). Computer-Mediated collaborative learning: theory and practice. *The Modern Language Journal, 81*(4), 470-481.

Warschauer, M., & Kern, R. (Eds.) (2000). *Network-based language teaching: concepts and practice*. Cambridge: Cambridge University Press.

Werry, C. (1996). Linguistic and interactional features of Internet Relay Chat. In S. C. Herring (Ed.), *Computer-mediated communication: Linguistic, social and cross-cultural perspectives* (pp. 47–63). Amsterdam/Philadelphia: John Benjamins.

Widdowson, H. G. (1998). Context, community, and authentic language. *TESOL Quarterly, 32*(4), 705–16.

Wierzbicka, A. (1991). *Cross-cultural Pragmatics: The semantics of Human Interaction*. Berlin: Mouton de Gruyter.

Willis, D. (1990). *The Lexical Syllabus*. London: Harper Collins.

Wilson, A., & Rayson, P. (1993). The automatic content analysis of spoken discourse. In C. Souter & E. S. Atwell (Eds.), *Corpus-based Computational Linguistics*. Amsterdam: Rodopi, 215–216.

Wood, D. (2010). Lexical clusters in an EAP textbook corpus. In D. Wood (Ed.) *Perspectives on formulaic language: Acquisition and communication* (pp. 88-106). London and New York: Continuum.

Wood, L., & Kroger, R. (2000). *Doing discourse analysis: Methods for studying action in talk and text.* London: Sage.

Wray, A. (2000). Formulaic sequences in second language teaching: Principle and practice. *Applied Linguistics, 21*(4), 463–489.

Wray, A. (2002). *Formulaic language and the lexicon.* Cambridge: Cambridge University Press.

Wray, A. (2013). *Formulaic sequences.* In C. A. Chapelle (Ed.), *The Encyclopedia of applied linguistics* (pp. 2201-2206). Oxford: Wiley-Blackwell.

Wray, A., & Perkins, M. (2000). The functions of formulaic language: An integrated model. *Language and Communication, 20*(1), 1–28.

Xiao, Z., & McEnery, T. (2005). Two approaches to genre analysis: Three genres in modern American English. *Journal of English Linguistics, 33*(1), 62–82.

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

Yates, S. J. (1996). Oral and written linguistic aspects of computer conferencing: a corpus-based study. In S. Herring (Ed.), *CMC: Linguistics, social and cross-cultural perspectives* (pp. 29-46). Amsterdam: Benjamins.

Yang, Y. F. (2011). Engaging students in an online situated language learning environment. *Computer Assisted Language Learning, 24*(2), 181-198.

Appendix A: Transcription codes

Transcription convention	Symbol	Explanation
speaker codes	<TW01>, <TW02>, <BT01>, <BT02>, etc.	TW and BT refer to Taiwanese and British participants respectively, and each speaker is numbered.
Extralinguistic information	a square bracket '[]'	[laughter], [coughing], [inaudible speech], etc.
interrupted sentences	a plus '+'	<BT18>:and water melon is a lot fresher+ <TW16>:Yeah. <BT18>: +than we have here.
unfinished words	an equal sign '='	<TW11>:In in typhoon, it's very.. very bad, you know, it's it's wet be= <BT15>: Yeah. <TW11>: because it's raining and it's cold.
lengthened sounds	colons ':' or '::'	exceptionally long sounds (i.e., approximating 2 seconds or more) are marked with a double colon '::'
brief break	a sequence of two dots (..)	a longer pause is marked as a sequence of three dots (...)
punctuation	. ? ,	A full stop or question mark is used to mark the end of a sentence. A common indicates that the speaker has re-cast what he/she was saying.

Appendix B: ***The Composition of BATTICC***

	BATTICC-O	BATTICC-F
Medium	written text	spoken text
Messages/turns	1,307	1,823
Total numbers of words	32,442	20,099
Types	2,983	1,701
Participants	N=70	N=70
Natinationality	35 Taiwanese	35 Taiwanese
	35 British	35 British
Age	13-14	13-14
Gender	M: 32; F: 38	M: 32; F: 38
Numbers of years learning English (in the case of Taiwanese participants)	4-5 years	4-5 years
Proficiency level of Endlish (in the case of Taiwanese participants)	low-intermediate	low-intermediate

Appendix C

The University of Nottingham

Mr Eric Yen-liang Lin
School of English
University of Nottingham
Nottingham NG7 2RD
January 10, 2011

Study on Adolescent Intercultural communication
Informed Consent

1. This study is being conducted by researchers at the University of Nottingham, UK. We wish to analyse data from the Connecting Classrooms Project, which is a global programme administered by British Council to create partnerships between clusters of schools in the UK and others around the world. Our goal is to determine the pedagogical merit of this project and to identify some of key factors for success in intercultural communication.

2. The participating Schools from Cumbria are: Ullswater Community College, William Howard School, Stainburn School & Science College, Appleby Grammar School, Nelson Thomlinson School, Queen Elizabeth Grammar School, Cockermouth School. The participating Schools from Taiwan are: Chi-An, Rui-Sui, Yi-Chang, Fu-Yuan, Tzu-Chiang, Hua-Ren and Guang-Fu Junior High School. There are currently around 45 British school children and around 50 Taiwanese school children involved in the project.

3. In this study the children will be requested to introduce themselves to this online community, and to write about school life, culture-based experiences and other related topics on the Moodle website (http://vle.connectingclassrooms.cleo.net.uk/). In addition, online synchronous meetings for further communication will also be arranged. These are for the participants to directly interact with their international peers and learn different cultures from them. It is hoped to deepen their understanding of other societies and cultures – as well as their own.

4. A face-to-face meeting will be held in Taiwan in May-June of 2011, and some of the discussion between Taiwanese participants will be recorded for research purpose. After the project, your child will be given a questionnaire

to outline what he/she has learnt from the project, communication difficulties they encountered and so on.

5. We do not anticipate any risk or harm involved in participating. Because the Moodle can only be accessed by participants in this project, the children and their identities will not be available to outsiders. The goal is to provide a safe and educational online learning community for all the students. The online environment is monitored by Ms Alison Phillips, from Stainburn School, Cumbria.

6. The data being collected is solely for the purpose of research. When writing up the children's data, names will not be used. The child will be referred to by a pseudonym. At any time you and/or the child has a right to withdraw his/her participation. Also, at the conclusion of the study you and/or the child may request to see the results of the research.

7. Should you have any questions or desire further information, please feel free to contact the researcher or the supervisor.

Mr. Eric Yen-liang Lin	Professor Svenja Adolphs
Teacher in Hua Ren School, Taiwan	Director of the Centre for Research in Applied Linguistics
PhD Student	School of English
University of Nottingham	University of Nottingham
+447412818006	+441158467219
aexyl@nottingham.ac.uk	Svenja.Adolphs@nottingham.ac.uk

Parental Consent Form

I have read and understand the above information and agree to allow _____ (child's name) to participate.

Adult's name _____ Adult's relationship to child _____

Adult's signature _____

Appendix D: Word Reading Test

1	cat	26	talk	51	worker	76	hurt
2	dog	27	tiger	52	super	77	mail
3	he	28	twelve	53	need	78	crazy
4	cake	29	or	54	any	79	change
5	rabbit	30	strong	55	warm	80	feeling
6	jump	31	pie	56	card	81	traffic
7	teacher	32	fourteen	57	hundred	82	convenient
8	mother	33	thirteen	58	husband	83	famous
9	watch	34	many	59	street	84	excited
10	under	35	train	60	fifth	85	weight
11	clock	36	picture	61	different	86	dead
12	very	37	bad	62	hard	87	joke
13	open	38	beach	63	slender	88	price
14	little	39	leg	64	when	89	note
15	cookies	40	miss	65	size	90	final
16	can	41	really	66	engineer	91	clerk
17	cool	42	how	67	afraid	92	skirt
18	park	43	it	68	sixteenth	93	somewhere
19	ice	44	soon	69	advertisement	94	safety
20	here	45	hate	70	son	95	break
21	frog	46	kid	71	parent	96	action
22	ox	47	tennis	72	less	97	gas
23	close	48	more	73	smile	98	mad
24	seventeen	49	waiter	74	mind	99	camp
25	movie	50	driver	75	grade	100	whenever

Appendix E: Test for Three-word Sequences

Please write the meaning of the underlined word sequence in each sentence:
請寫出畫線部分之中文意思：

1. I am great just really busy, ***how are you***?
2. I think I already did ***a couple of*** weeks ago! haha
3. I've been up since 5:30 and ***I don't know*** why!
4. Today ***I'm going to*** start writing a blog. would anyone read it if I did?
5. What ***would you like*** to see at the meeting?
6. ***Thanks for the*** link. For some bizarre reason I've been invited.
7. Tickets are free ***if you want*** to come.
8. I'm ***looking forward to*** seeing everyone.
9. i can do the ordinary bowling and the tennis, but ***I don't think*** I'd be much good at Wii Fit.
10. excellent, I always wanted to ***be able to*** fly.
11. Maybe ***it would be*** good for a dinner or something, but I don't go to many of those.
12. Oh, ***I have to*** tell them to take it...
13. ***There is a*** small boy terrorizing the coffee drinkers of Ealing.
14. Are you ***looking for a*** job?
15. Congrats!! Hope to see your work ***in the UK*** soon :)
16. I'd read it, ***but I think*** I'm too lazy to do ads now...
17. I'm not sure what's going on with it ***at the moment***.
18. This is a huge ***part of the*** reason
19. I think this week is ***going to be*** a good one...
20. At least you're guaranteed a pay check at ***the end of*** every month!
21. in fact this is ***the first time*** I've sat down to watch a programme in a very long time.

22. I think we will make it work, but it won't be easy, and there isn't ***a lot of*** time to do it.
23. Never being on twitter sucks. ***A bit of*** a break was fine, but not at all? Crap.
24. I have been lying down for ***most of the*** weekend (sniff).
25. it offers free downloadable materials for English language classes, ***as well as*** further information.
26. ***Have a look*** here for all details: http://www.channel4.com/
27. I remember my Gran and Grandad coming away with us ***when I was*** little.
28. oh dear - it ***seems to be*** everywhere at the moment!
29. It's just ***one of those*** things that I do really love but...
30. If the man doesn't like singing on a stage ***in front of*** children, then he should get another job.
31. Best of luck! ***Let me know*** how it goes.
32. Hopefully ***it will be*** a great result for us.
33. But, when I got ***back to the*** office, I looked them up and spent ages on their site
34. Slate is hiring a photo researcher in New York. ***If you are*** someone--or know someone--who would be good, contact us!
35. ***I have been*** playing this clip all evening, just because I know I shouldn't...
36. There's lots of things ***I used to*** do that I don't anymore!
37. Thank you - and ***Happy New Year*** to both of you! Xxx
38. See the examples at ***the top of*** this page.
39. And the security are doing their nuts trying ***to find out*** who the secret smoker is...
40. Have a great day everyone, wherever you are ***in the world***.

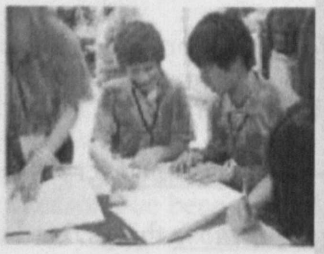
以上文字出自於網路用語，因此書寫方式為非正式用法，部分亦不符合寫作的文法規則。

Would

A. Read and Highlight

*Read the online dialogue between Peter and Sula, and underline the formulaic expressions surrounding **would**.*





Peter: You might want to try some Taiwanese local food.

Sula: **I would love to**. **It would be** really nice to have some special dishes.

Peter: Bubble milk tea is my favourite. **Would you like** to try it?

Sula: Yeah, **I'd like to** have some when I come to Taiwan.

Peter: I think I can make some for you. That's easy.

Sula: Really, wow. **That would be** great.

Peter: Maybe I can also teach you how to make it.

Sula: Oh, **that would be** fun. I've never made this sort of drink.

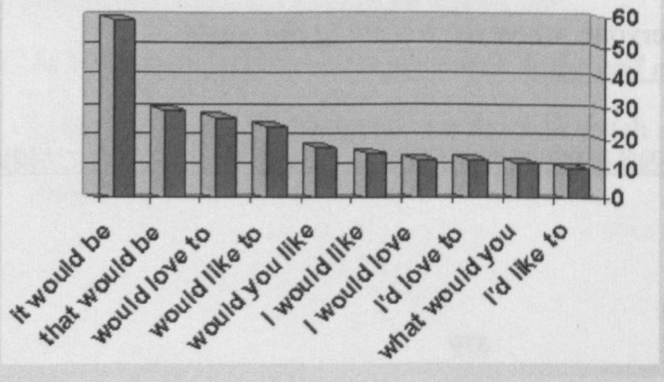
❖ Sorting activities:

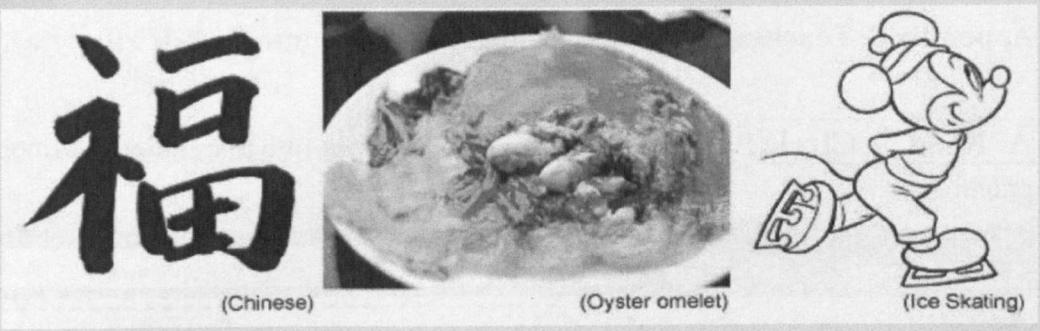
Categorise the expressions according to their functions.

- Volition/Need: _____
- Complying/Possibility: _____
- Offer/Requests: _____

❖ Useful expressions

Most frequent 3-word formulaic sequences in the corpus of e-language





B. Language Building

Complete the sentences with the sequences above. Then practice with your partner.

❖ Practice: Complying/Possibility

1. A: Do you learn Chinese?
   B: I don't learn Chinese, but I hope to learn some soon. I think _____ fun to learn another language.
2. A: This is the picture of an oyster omelet. This dish is very popular in Taiwan.
   B: It looks a bit strange, but I looovveee oyster so I think _____ very nice.
3. A: Maybe we can become pen pals.
   B: _____ really nice to have a pen pal to e-mail or write to.
4. A: Do you like ice skating?
   B: Yeah, I like it so much. I think _____ if you tried it because it's very fun.

❖ Practice: Volition/Need

1. I've never tried Taiwanese tea, but I _____ try it when we come over.
2. So everyone, do you have any New Year's resolutions you _____ share.
3. I _____ know how to make mochi so that I can show my friends.
4. haha, I _____ be friends with you. I can't wait to come over and meet everyone.

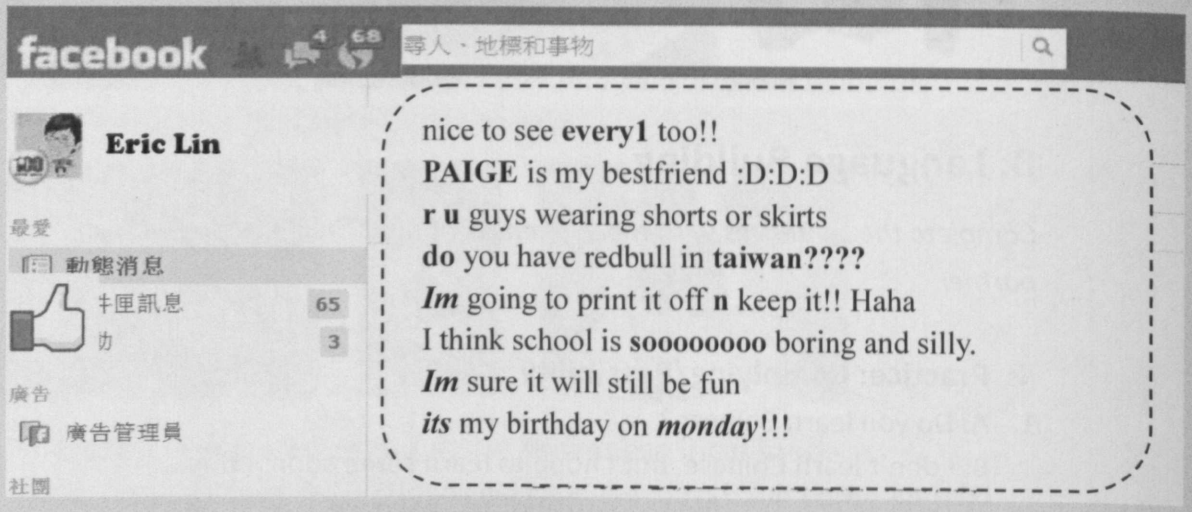
❖ Practice: Offer

What would you like?	Would you like a/some...?	Would you like to...?

1. _____ to do tonight? Are you going for dinner outside?
2. _____ to come with me? They would be very happy if you come.
3. _____ to try the dish I made for you? I hope you like it.
4. _____ some Amis mochi? It's very delicious.

Appendix G: Teaching Materials for E-grammar

A. Read & Circle Please circle the words that do ***not*** fit into traditional written grammar.



- ***Nonstandard upper and lower cases*** _____
- ***Abbreviation*** _____
- ***Acronyms*** _____
- ***Substitution*** _____
- ***Repeating words for stress*** _____
- ***Nonstandard Punctuation*** _____

B. Decoding

its my birthday on **monday!!!** _____
thank you **cindy**, and **i LOVE** the picture!! _____
I ***dont*** actually know how short that is _____
Wat do **u** like to do after school? _____
Plz respond i **wud** love to **here wat u hav** to say _____
i **agreeeeeee** with **u m8** _____
Goin to bed or I won't wake up **In da morning** _____

Hi, _____
How are you doing? _____
Good night, everyone. _____
You look good on this picture. _____
I hope you have fabulous time mate. _____
I'm going to print it off and keep it. _____
_____ _____

Appendix H: Teaching Materials for Spoken-grammar

❖ Conversation Strategy 1: Vague categories

What do you think the underlined expression means?

We have a lot of festivals <u>and things like that</u> in Taiwan.

a. people *b. festival* *c. Taiwan*

❖ *Building Language*

✓ *Now, listen to the CD and add vague expressions to the comment.*

1. The girls have to have their hair out of their faces and they can't wear make-up ***(and stuff).***
2. Every year we Amis people celebrate our harvest festival ***(and things).***
3. We might not have chance to come again because of like money ***(and stuff).***
4. What sort of different things have you been noticing in our culture and traditions ***(and stuff)***?

Useful expressions

and things (like that), and stuff (like that), and everything, or whatever, and that kind of thing, and that sort of stuff

❖ Conversation Strategy 2: Softening comments

Which comment in each pair sounds "softer"?

a. It looks strange. *a. They are shy.*

b. It looks kind of strange. *b. They are just a little shy.*

❖ *Building Language*

✓ *Now, listen to the CD and add 'something' to soften the comment.*

1. Don't mind these two, they're ***(a bit)*** weird.
2. In a way it was ***(a bit)*** boring because we had to ...
3. ***(I think)*** English is ***(a little bit)*** difficult so I can't speak well.
4. ***(I think)*** I'll go for the easy stuff.
5. Yeah, your food generally is a lot more ***(sort of)*** traditional and special than ours. Ours is just ***(sort of)*** simple.

Useful expressions

I guess / I think, ***a little / a (little) bit***

kind of / sort of, ***in a way / just***

Notes on the materials development:

A number of issues needed to be addressed at the outset of the development of the materials. Firstly, choosing appropriate topics to be included in EFL materials for an intercultural programme is one of the most important issues. The key semantic domain analysis (4.4 and 4.5) has revealed the themes that young people commonly discussed online and in FTF interaction among Taiwanese and British adolescents, including food, school life, sports and hobbies. The multi-word sequences analysis (6.3.2) has also illustrated the sequences that clearly mark a number of "necessary topics", which strongly support the findings of keyword and semantic domain analysis. It is suggested that EFL teachers, therefore, start the course for an intercultural programme with those topics identified in this study, and encourage their learners to observe the culturally relevant lexical features and further develop intercultural awareness.

Another important issue involved in the EFL materials development is to meet learners' needs. The thesis has presented a number of significantly overused and underused lexical and grammatical items by Taiwanese learners, as compared with native speaker discourse. The gaps between EFL textbooks and authentic data from corpora are also identified (see Chapter 7). It is therefore very helpful to select these linguistic elements and put them in learning materials, such as modals, vague language, discourse markers and tenses. In this case, what should be included in the learning materials are informed by the three approaches employed in this thesis (i.e., keyness, discourse analysis, multi-word sequence analysis).

Moreover, concordance lines provide information about the context of use for particular words or phrases (e.g., Reppen, 2010; Römer, 2011). Johns (2002) suggests to "confront the learner as directly as possible with the data, and to make the learner a linguistic researcher" (p. 108). However, bringing raw concordance lines extracted from a corpus into the classroom might scare learners, especially those who are just at beginning or intermediate level, as corpus data might contain many unknown words and complex sentence constructions that may be too challenging for them. Tomlinson (2011) notes that "materials should help learners to feel at ease" (p. 8). In this case, the concordance lines are carefully selected, in which I excluded the lines which contain more than two words that are not included in the essential 1200 words for Taiwanese EFL learners listed in *The general guidelines of grades 1-9 curriculum for elementary and junior high school education.* In addition, the format of concordances might also scare some learners. I therefore retain the familiar textbook appearance that learners and

teachers are accustomed to and still maintain the authentic nature of BATTICC. It is to the three sample materials for teaching multi-word sequences, e-grammar and spoken grammar that we now turn.

Sample material 1: multi-word sequences

In the first section of the material, texts from the online discussion board (BATTICC-O) are extracted and amended slightly to suit the purpose of this lesson. Learners are then encouraged to highlight the multi-word patterns, helping them to raise their awareness of useful expressions of *would*. Such textual enhancement of input is important as noticing is a prerequisite for intake, which is based on Sharwood Smith's (1993) input enhancement hypothesis. She notes that L2 learners in general lack sensitivity to grammatical features of target language input, so even if a large amount of authentic data is provided, learners may not be able to benefit from it (see also Han, Park & Combs, 2008). Tomlinson (2011) also emphasises that "the learners' attention should be drawn to linguistic features of the input" (p. 14). These highlighted multi-word expressions as a result could be easily recorded as chunks of data that can be used by learners with a potentially large number of words, phrases, and sentences (O'Keeffe et al., 2007; Wray, 2000).

Apart from stimulating input processing for language form, following the dialogue in the material, the sorting activities encourage learners to read the context more carefully and identify the functions of different multi-word sequences. For example, *I'd like to...* would be categorised as Volition/Need; *Would you like...* expresses an offer; or *it would be...* indicates complying/possibility. In addition, information about sequences based on a body of online language is displayed, letting learners know which expressions are most frequently used in online communication. This information is explored based on the discourse that native English speakers actually produced, rather than by intuition. Section B offers more practice via fill-in-the-gap exercises drawn from concordance data, with multi-word sequences that, when used in context, can be particularly valuable in the illustration of different meanings of lexical items.

Sample material 2: E-grammar

The material first encourages learners to find out the words or linguistic elements that do not fit into traditional written grammar, and these items found are then codified and categorised into different features of CMC. The second part of the material requires learners to practice the codification between the CMC language and traditional writing. For example, the online language "Goin to bed or I won't

wake up In da mornin" is given, so the learners may write "I'm going to bed or I won't wake up in the morning" to show their understanding of CMC language. This practice can enable learners to recognise the common features of online discourse. On the other hand, learners have opportunities to express their thoughts using some CMC features, as can be seen in the last section of the material. Nevertheless, Rúa (2007) notes that the rules of shortening have a certain degree of freedom in their application and the code may be modified to satisfy in-group needs (p. 183).

Sample material 3: Spoken grammar

This thesis has demonstrated that many features of spoken grammar have important interpersonal functions which serve to create and maintain good relationships between the speaker and hearer, as noted by O'Keeffe et al. (2007, p. 159). Teaching spoken grammar would therefore be very helpful for students who are involved in an intercultural exchange to make them aware and learn these important linguistic features. The sample material presents an example of teaching vague language as it is significantly underused by Taiwanese learners as compared with their English-speaking peers, and it cannot be found in the EFL textbooks used in Taiwan either. This worksheet includes two parts of vague expressions: vague categories and softening comments. In Part A, learners are asked to add a vague expression, which refers to vague use of categories of items. Common vague expressions such as *and stuff* and *and things like that* are provided for learners to choose from. Part B asks learners to add a vague expression to soften the comment in some way. As has been discussed in Chapter 5, such a device is frequently used by the British participants in order to hedge the commitment of the speaker to whatever he or she asserts. This material, therefore, would raise the Taiwanese learners' awareness of how their British peers employ vague language, so that their comments do not appear overly direct.