國立臺灣大學文學院語言學研究所

博士論文

Graduate Institute of Linguistics

College of Liberal Arts

National Taiwan University

Doctoral Dissertation

中文的常用詞串

Lexical Bundles in Chinese

許展嘉

Chan-Chia Hsu

指導教授：謝舒凱 博士

Advisor: Shu-Kai Hsieh, Ph.D.

中華民國 105 年 6 月

June 2016

doi:10.6342/NTU201600328

國立臺灣大學博士學位論文
口試委員會審定書

中文的常用詞串
Lexical Bundles in Chinese

本論文係許展嘉君（學號 D00142001）在國立臺灣大學語言學研究所完成之博士學位論文，於民國 105 年 1 月 19 日承下列考試委員審查通過及口試及格，特此證明

口試委員：

謝舒凱 （簽名）

（指導教授）

呂佳蓉　　張妙霞

高照明　　李佳霖

doi:10.6342/NTU201600328

謝辭

自從論文口試通過以來，我收到身旁很多人的恭賀與讚美。然而，我自己相當清楚，能夠進入博士班就讀，並且順利畢業，是上天對我的眷顧，此外，也是因為有身旁的各位陪伴著我，才能一路走完這趟旅程。我對大家的關懷無以回報，只能以這段短短的謝辭，表達心中的感激。

首先，我要向我的指導教授謝舒凱老師致上最深的謝意。一開始進入博士班時，我對程式設計一竅不通，但謝老師仍然願意接受我，耐心等待我從零開始學習。對於實驗室裡的計畫，我有時冒出不成熟的想法，謝老師總是親切說明。在指導論文的過程中，謝老師用最開明的態度支持我想鑽研的題目、配合我的研究時程，但是用嚴格的標準檢視我的研究方法，並對分析結果提出許多寶貴的建議。為了能夠掌握我的學習狀況，謝老師建立了一個只有我們兩個能夠進入的線上共享文件，讓我在上面更新研究進度，而謝老師則會不時留下鼓勵的語句，這份文件裡有我們之間的悄悄話，我會永遠珍惜。另外，這幾年求學的過程中，謝老師為我寫過無數封推薦函，協助我申請各項國際會議補助、獎學金、以及求職等，我之所以能夠得到這些榮譽，要歸功於謝老師的指導與大力推薦。

對於這篇論文的完成，我也由衷感激口試委員們，在百忙之中抽空閱讀，並賜予許多有建設性的建議。我要謝謝張妙霞老師，在碩士班時期的篇章分析、學術寫作課程中，為我日後的研究奠定穩固的基礎，並培養出我對篇章分析濃厚的興趣，在論文口試中，張老師批判性的思維模式，引導我探究更深入的議題。我要謝謝高照明老師，從計算語言學的角度，提供這篇論文未來可再繼續發展的方向，我常在各種場合上聽到高老師的演講，也從中獲益良多。我要謝謝呂佳蓉老師，在一年級時上「認知科學基本議題」時，跟我交換許多有趣的想法，後來上「詞彙語意學」時，深入淺出介紹各學派的理論，在論文口試時，謝謝呂老師點出一些可再多加闡釋之處。我要謝謝李佳霖老師，雖然從來沒有修過李老師的課，但李老師仍願意擔任我的口試委員，從心理語言學的角度，提出不少值得關注的面向，讓我的論文內容更加豐富。

此外，我要感謝臺大語言所其他老師們在這五年來對我的諄諄教誨，包括黃宣範老師、蘇以文老師、江文瑜老師、馮怡蓁老師，所有老師們對於博士生時時表達關懷，讓我感受到溫暖。我尤其感謝以下三位老師：謝謝黃宣範老師在課堂上不吝與我們分享畢生所學，此外，黃老師雖未擔任這篇論文的口試委員，仍撥空出席論文大綱審查，當時對我的研究題目表現出很高的期待，黃老師的鼓勵鞭策著我，讓我不敢懈怠。謝謝蘇以文老師，在多年前我尚未進入臺大語言所，就讓我有機會參與臺大寫作教學中心的教材編纂工作，在我入學後繼續提供許多工讀的機會，這些對我來說都是極為珍貴的學習機會。謝謝宋麗梅老師，在我擔任博士班代表期間，指導我處理行政事務，在一同籌辦「第七屆語言、言談、與認知國際研討會」(CLDC)時，謝謝宋老師對我們學生團隊的支持與信任。另外，我也要感謝語言所的三位助教：劉美玲女士、劉嘉蘭女士、楊靜琛女士。在申請各項獎助與處理所務的流程中，三位助教給予許多協助，為我省下寶貴的時間。

在語言學研究這條路上，我還要感謝我的貴人：畢永峨老師。無論是治學的嚴謹程度或是待人接物的態度，畢老師都是我的典範。從碩士班時期開始，畢老師就訓練我做事有條不紊的方式，這對我日後的人生幫助很大。在我取得碩士學位後，畢老師建議我進入計算語言學的領域，鼓勵我勇於突破語言學研究的瓶頸，謝謝畢老師當時提醒我這些，讓我後來有機會接觸到更多有趣的語言現象。我進入博士班之後，謝謝畢老師仍時常關心我的生活，與我分享人生的道理與體悟。而畢老師退休之後，仍不忘提攜我，不厭其煩多次為我撰寫推薦函。畢老師對我的恩情，我將永生難忘。

接著我要感謝謝老師研究團隊(Lab of Ontologies, Language Processing and e-Humanities, LOPE)的成員。我要感謝我的同學施孟賢(Simon)，每次出國參加研討會都

doi:10.6342/NTU201600328

會記得寄明信片給我，上面祝福的話讓我覺得很溫暖，而去年我們兩人一起前往芝加哥短期進修，Simon 當了我的短期室友，謝謝 Simon 當時對我的照顧。我要感謝張瑜芸(Taco)，常常犧牲自己的時間指導我的程式設計，我一發出求救訊息，Taco 幾乎都是秒回，義不容辭放下自己手邊的工作。另外，我要感謝呂珮瑜(Emily)、王伯雅(Amber)、莊茹涵(Yvonne)、林怡馨(Shanon)、劉純睿(阿吉)等，與我交換許多新鮮的研究想法、彼此支援程式設計、分享研究的甘苦、一同慶生等。在謝老師團隊中歡愉的氣氛，是支持我繼續向前的動力之一，我此生都會以身為一名 Loper 為榮。

我也要感謝臺大語言所的以下幾名博士生：蕭書珮(Peggy)、林盈妤(Tiffany)、謝承諭、王國樹、沈文琦、陳萱芳。謝謝他們在我擔任語言所博士班代表的期間，針對所務提出建言並大力相助，也謝謝他們的努力，讓 CLDC 會議、所友會的交接一切順利。其中 Peggy 與 Tiffany 是與我一同籌辦 CLDC 的伙伴(還有 Simon)，在過程中或許我有些不夠周到之處，謝謝他們的包容與支持。

另外，我要感謝臺大教育學程英文科的伙伴們。我要謝謝黃恆綜老師(Danny)，與我們分享許多有趣的教學活動，並花了許多心力為我們安排試教學校、邀請講者等，此外，謝謝 Danny 大方與我分享在大學求職的經驗與技巧，額外撥空觀看我的教學演示，提供了許多相當有用的建議。謝謝班上的所有同學：黃鈺婷(Diana)、王昶皓(Harry)、范榮約(Yvonne)、蕭瑤(Helen)、嚴文臨(Wayne)、謝佳恩(Joanne)、曾國奕(Chander)、陳以恩(Emily)、張心惟(Amy)。我是班上年紀最大的學生，謝謝他們願意接納我，他們的支持讓我在博士班期間順利完成教育學程的課程。我要特別謝謝 Diana 曾經連續兩個學期與我一起完成教學計畫，與她一起合作是很愉快的經驗，我也從她身上學到許多。

在這五年攻讀博士學位的過程中，難免有覺得力不從心的時候，而好朋友們就會在此時為我加油打氣，我要特別感謝以下幾位朋友。我要謝謝張建斌，在我說出一些抱怨、賭氣、任性的話時，忍受我的壞脾氣、當我的垃圾桶、耐心聆聽我、開導我，使我豁然開朗、一掃陰霾，還常在假期時邀我一起旅行，讓我放鬆心情。我要謝謝王炳勻，他同時遠在美國攻讀語言學博士學位，不過就算是時差也不能阻隔我們聊天，從高中時期到現在培養出的默契，讓我們對彼此的處境能夠感同身受，此外，也要謝謝他對我的研究提出許多獨到的見解。我要謝謝林柏仲、黃治瑋、沈正嵐、何信昌(Harvey)，他們不時關心我的近況，鼓勵我不要放棄，我們的人生中都有些難題，我們彼此扶持、一起度過，而我要特別感謝 Harvey 曾無償替我校閱資格論文。另外，我要謝謝紀怡嫻女士(Kelly 姐)把我當成家人一樣，這些年來持續關心我、鼓勵我，對我的生涯規劃表達支持之意。

在我就讀博士班的期間，為了生計在外兼課。我要特別感謝大同大學外語教育中心的謝富惠主任、蘇安德老師、石慧中助教、張聖潔助教，以及補習班裡的 Bernie 主任、Catherine 主任、Phil 老師、Sean 主任、Ariel 教務主任、Jaime 老師、Winston 老師。謝謝他們在排課時考量到我的博士班課業，讓我享有彈性，並提供許多教務上的支援，也要謝謝他們對我的鼓勵。

最後，我要對我的家人表達深深的謝意。這些年來，由於課業、工作繁忙，犧牲了無數與家人相處的時光，我要謝謝他們的包容與無條件的支持。我要特別感謝我的媽媽，從小到大無微不至地呵護著我，為我營造一個無後顧之憂的學習環境，我進入博士班之後，仍時時掛心我的健康狀況。我也感謝上天讓弟弟的孩子宸熙在兩年多前誕生在我們家，他可愛的笑臉為我消除心理上的疲勞。

在這段期間以來，所有曾對我說過一句加油的人，都是我在這裡要感謝的對象。前方的路或許有更艱難的挑戰，然而，有大家的陪伴，我會努力往前走，不輕言放棄。

doi:10.6342/NTU201600328

摘要

本論文旨在抽取中文口語與新聞裡的常用詞串(lexical bundle)，並分析其在篇章中的使用。本研究為中文常用詞串的語言結構建立一套分類架構，也仔細審視這些詞串的功能。

本研究從中央研究院現代漢語平衡語料庫第四版中，抽取三字、四字常用詞串。一開始，先自動抽取出在每百萬詞中出現至少二十次、並出現於至少五個檔案的詞串。這種幾乎是純以頻率為本的取向有些研究方法上的議題仍待解決，因而須加採更為敏感的離散指標(dispersion measure)、詞彙搭配力指標(word association measure)，所得結果也需要再經過人工分析。

探索性資料分析(data exploratory analysis)顯示，口語對話中出現的詞串類型較新聞多；此外，關於詞串在語料庫中所佔的比例，口語對話也較新聞高。同樣的傾向早已在英文中觀察到，這些發現意味著，在自然口語中，說話者面對即時的壓力，因此更依賴像是詞串這類預製語塊(prefabricated chunk)。

本研究接著深入探討中文對話裡常用詞串的使用，為先前英文裡的發現提供跨語言的支持。第一，中文對話裡大部分的詞串在結構上並不完整，且跨越傳統的語法結構，但我們仍可根據這些詞串的結構特徵為其做系統化的分類。第二，這些詞串在篇章中具有三大主要的功能類型，可以促進人際溝通，例如表達立場，亦可組織篇章，例如引介話題，還有各種指涉的用法。第三，常用詞串的結構與功能之間存有明顯的關聯性：用於表達立場的詞串大多以子句、動詞片語的形式出現，而用於指涉的詞串則大多以名詞短語的形式出現。另一方面，中文與英文有一項顯著的差異，即中文裡名詞詞串的數量相當多，這點可歸因於中文特殊的語言結構特徵。

此外，本研究亦深入探討中文新聞裡常用詞串的使用。結果發現，新聞寫作的傳統與原則，例如貼近事實、避免模糊不清、使新聞事件與讀者產生關聯、精簡等，會影響新聞中詞串的分布。例如，相對於口語對話來說，用於表達不精準、

doi:10.6342/NTU201600328

立場不確定的詞串在新聞中顯然較少出現，而有些用於強調新聞價值的篇章組織詞串則時常在新聞中出現。儘管有這些差異，用於處理對話詞串的分類架構仍適用於新聞裡的詞串，常用詞串結構與功能之間的關聯性亦存在於新聞語體中。

我們期望本論文對中文常用詞串的研究成果，能夠從以語言使用為本(usage-based)的觀點，闡明多字組合(multi-word unit)何以浮現(emerge)出來，並說明語言結構與不同語體溝通需求之間複雜的關係。本研究所抽取出的詞串可用於擴增中文的語言資源，亦可作為語言教師、學生重要的參考資料、以及心理語言學實驗的素材。

關鍵詞：

篇章組織詞串；頻率；人際溝通詞串；常用詞串；指涉詞串；語體；使用基礎模型

doi:10.6342/NTU201600328

Abstract

The present dissertation aims to identify lexical bundles (i.e., recurrent word sequences) in Chinese conversation and news and investigate their use in discourse. A structural taxonomy is created for lexical bundles in Chinese, and their functions are also closely examined.

In the present study, three-word and four-word lexical bundles are identified from the Academia Sinica Balanced Corpus of Mandarin Chinese (the fourth edition). An initial list includes word sequences occurring at least twenty times per million words and in at least five corpus texts. To deal with vital methodological issues concerning this almost purely frequency-based approach, another more sensitive dispersion measure, a word association measure, and a manual analysis are needed.

A data exploratory analysis of lexical bundles in Chinese shows that conversations feature a much wider range of lexical bundles than newswire texts; as for the proportion of corpus data covered by lexical bundles, conversation is also higher than news. The same tendency has been observed in English, suggesting that in spontaneous speech, speakers are under real-time pressure and thus rely more heavily on prefabricated chunks such as lexical bundles.

A comprehensive investigation of lexical bundles in Chinese conversation provides more cross-linguistic support for previous findings in English. First, most lexical bundles in Chinese conversation are not structurally complete and run across traditional grammatical structures, but they can be systematically categorized according to their structural characteristics. Second, these bundles serve important functions in discourse, facilitating interpersonal communication (e.g., expressing stances), organizing discourse (e.g., introducing a topic), and having a variety of referential uses. Third, there is a strong relationship between the structure and the

doi:10.6342/NTU201600328

function of lexical bundles: stance bundles are closely associated with clausal and VP-based categories, whereas referential expressions are closely associated with NP-based categories. On the other hand, a striking difference between Chinese and English is that NP-based bundles are much more dominant in Chinese, and this is attributed to structural characteristics specific to Chinese.

Furthermore, a detailed examination of lexical bundles in Chinese news suggests that conventions and principles in journalistic writing (e.g., sticking to facts, avoiding ambiguities, relating news event to readers, using shorter forms) influence the distribution of news bundles. For example, imprecision bundles and personal epistemic bundles that express an uncertain stance occur less frequently in news than in conversation, while some discourse organizers that are used to identify something newsworthy to the reader occur more frequently in news than in conversation. Despite these differences, news bundles fit comfortably in with the classification frameworks of conversation bundles, and the relationship between the structure and the function of lexical bundles is reconfirmed.

It is hoped that the findings of the present dissertation on lexical bundles in Chinese can elucidate the emergent nature of multi-word units from a usage-based perspective and illustrate complex interactions between language-specific structural properties and genre-specific communicative needs. The lexical bundles identified in the present study can be used to enrich existing language resources in Chinese, and they may also serve as important references for language teachers/learners and psycholinguistic experiments.

Keywords: discourse organizer; frequency; interpersonal bundle; lexical bundle; referential expression; text type; usage-based model

doi:10.6342/NTU201600328

Table of Contents

doi:10.6342/NTU201600328

doi:10.6342/NTU201600328

List of Figures

doi:10.6342/NTU201600328

List of Tables

doi:10.6342/NTU201600328

doi:10.6342/NTU201600328

doi:10.6342/NTU201600328

Chapter 1

Introduction

This chapter opens with how corpus linguistics brings formulaic language to the fore and zooms in on the Biberian approach to lexical bundles. Then a research gap will be highlighted, and the research questions in the present study will be posed.

1.1 General Background

Generally, a corpus is defined as "a collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting point of linguistic description or as a means of verifying hypotheses about a language" (Crystal 2008: 117), and the compilation of corpora is compared to "the invention of telescopes in the history of astronomy" (Stubbs 1996: 231). Our easy access to a wealth of corpus data has revolutionized linguistic studies, and it is suggested that corpus linguistics and discourse analysis have become "the twin pillars of language research" (Sinclair 2004: 10). Nowadays corpora are growing at a pace faster than ever before, and the web is serving as a mega-corpus (Flowerdew 2012: 39). With corpus data, a great many surprising regularities underlying our language use can be uncovered. Gries (2014) suggests that linguistics is generally a distributional science, and summarizes three major issues corpus linguists are exploring (Gries 2014: 365):

(i) the frequencies of occurrence of linguistic elements in corpora, for example, frequency lists;

(ii) the dispersion of linguistic elements in corpora as in, for example, measures of dispersion;

(iii) the frequencies of co-occurrence of linguistic elements in corpora as in, for example, collocation, collocational frameworks, *n*-grams,

doi:10.6342/NTU201600328

colligations/collostructions, etc.

Since the distributional pattern of co-occurrences is a central issue in corpus linguistics, it comes as no surprise that formulaic expressions, or multi-word combinations, have assumed a central place in linguistic studies nowadays.[1] However, some of the earliest research on formulaic expressions is actually traced back to the clinical domain of aphasia studies in the nineteenth century (Wray 2013: 316). Then, the Chomskian approach (i.e., the generative approach) attempts to draw a neat distinction between the lexis and the grammar in a speaker's linguistic competence, and multi-word combinations are regarded as peripheral for a while. Meanwhile, despite the dominance of the generative approach, some have provided fresh insights into the routinized nature of our language use (e.g., Bolinger 1961, Chafe 1968, Makkai 1972, Coulmas 1981, Pawley and Syder 1983, Lambrecht 1984, Tannen 1987, Nattinger 1988). It is argued that humans share a universal drive to imitate and repeat, so the grammar component is considered to be a memory repertoire of innumerable utterances that have been previously heard. Our communication is regarded as "basically a 'compositional' process, one of 'stitching together' preassembled phrases" from that repertoire (Nattinger 1988: 76), and the use of prefabricated patterns is a cognitive strategy language users adopt to avoid too many disfluencies (e.g., hesitations, pauses) and gain more time for their almost spontaneous speech. This view thus blurs the clear-cut dichotomy between the lexis and the grammar within the generative paradigm.

Over the past few decades, rapid technological advances have led to the advent of computer-readable corpora and sophisticated tools for text analyses, and corpus-based observations have not only supported the view that highlights the

[1] The terms *formulaic expressions* and *multi-word combination* are used as umbrella terms in the present study, and they are exchangeable.

doi:10.6342/NTU201600328

conventional side of our language use, but also brought multi-word combinations to the fore (e.g., Stubbs 1996, Hunston 2002, Sinclair 2004). A more empirically solid critique of the generative approach is then provided. For example, syntactic categories are not so solidly established as they might appear to be. Sometimes it is difficult to determine the syntactic class of a word when it is used in an idiomatic phrase. Additionally, when co-occurring with other words, a word may have specific behaviors or restrictions that syntactic rules alone cannot accommodate. Therefore, more and more discussion has centered around how "words enter into meaningful relations with other words around them" (Sinclair 2004: 25). The lexis and the grammar are thought to be interdependent.

Studies on formulaic expressions differ to a substantial degree, though.[2] First, methods for identifying multi-word combinations differ significantly (Biber et al. 2003: 72). Some rely simply on frequency criteria, while others rely heavily on more complex quantitative/statistical methods (e.g., Baroni and Evert 2008, Wiechmann 2008). Second, studies on multi-word combinations can also be classified according to their themes. Wray (2013) identifies six themes that most phraseological studies center around, suggesting that a study can fall into more than one theme category, and that each theme has its research timeline which can be pursued more or less independently.[3] No single approach is sufficient to provide the whole picture of the role multi-word combinations play in our language use, so empirical findings from different perspectives need to be synthesized (Biber et al. 2003). Corpus linguists need to constantly move between quantitative data and qualitative interpretations

[2] Wray (2002: 9) provides a long list of terms used to describe multi-word combinations of various kinds, such as chunks, clichés, collocations, lexical phrases, and so on. Some of the terms actually refer to the same kind of multi-word combinations.

[3] The six themes are as follows: (i) theory: processing, lexis and grammar; (ii) clinical: language disorders; (iii) development: first language acquisition; (iv) learning and teaching: second language acquisition; (v) culture: oral traditions, social roles, and cultural variation; (vi) text: corpora. See Wray (2013: 318-319) for the major issues in each theme.

doi:10.6342/NTU201600328

because the former alert us to points of potential interest and the latter provide meaningful explanations for emerging patterns (Hunston 2002).

It has been widely recognized that formulaic expressions are ubiquitous in our language use, as illustrated in the following table. It goes without saying that the considerable fluctuation shown in the table is due to different methods adopted for identifying multi-word combinations. Many studies have also revealed that second language learners, like native speakers, rely on formulaic expressions (e.g., Bolander 1989, Oppenheim 2000, Spöttl and McCarthy 2004). Studies on formulaic expressions have had implications and applications in many related fields, such as language pedagogy, translation research, stylistics research, lexicography, communication in business and healthcare contexts, forensic linguistics, and so forth (Partington 1998, Hunston 2002, Flowerdew 2012).

Table 1.1. Proportion of formulaic expressions in the corpus.
(The following studies are sorted in chronological order.)

Study	Register	Proportion
Sorhus (1977)	spontaneous speech	20%
Howarth (1998)	academic writing	31%-40%
Altenberg (1998)	spoken language	80%
Biber et al. (1999)	conversation	28%
	academic writing	20%
Erman and Warren (2000)	spoken language	58.6%
	written language	52.3%
Foster (2001)	unplanned speech	32.3%
Van Lancker-Sidtis and Rallon (2004)	the screenplay of a film	25%

After frequency data have brought to the fore phraseological patterns that used to go unnoticed by linguists in the pre-corpus generation, some explanations for the ubiquity of multi-word combinations are provided (see Conklin and Schmitt 2008). First, from a sociofunctional perspective, multi-word combinations serve important

doi:10.6342/NTU201600328

discourse functions. For instance, *what happens if/when* is used to introduce rhetorical questions, and *as so often happens* is used to introduce the discussion of a certain phenomenon (Partington 1998: 104). Second, from a psycholinguistic perspective, the use of multi-word combinations can enhance our processing efficiency (Pawley and Syder 1983). A large number of prefabricated chunks may have been stored in the mental lexicon, readily available to language users to effortlessly and fluently handle online interactions. Every language user is likely to have a "phrasicon" (Fillmore 1978: 149), or a "phrasal lexicon" (Conklin and Schmitt 2008: 75), which is compared to "a phrase book with grammar notes" (Pawley and Syder 1983: 220), storing conventionalized ways of encoding certain meanings and thus offering a processing advantage.

One of the most influential studies on formulaic expressions is Biber et al. (1999), in which a large number of multi-word combinations referred to as *lexical bundles* are identified, i.e., "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber et al. 1999: 990). Common four-word bundles include *I don't know what*, *I don't want to*, *I was going to*, *do you want to*, and *are you going to*. The method for identifying lexical bundles is mainly frequency-based: i.e., a lexical bundle is operationally defined as a word sequence occurring at least ten times per million words as well as occurring in at least five texts. Such a frequency-based approach to formulaic expressions is considered valid since frequency effects have been found to be correlated significantly with many aspects of our cognitive functions (e.g., Bybee 1985, Saffran et al. 1996, Brooks and Tomasello 1999, Bybee and Scheibman 1999, Saffran and Wilson 2003, Goldberg et al. 2004). A wide variety of corpora, such as learner corpora (e.g., Cortes 2004), textbook corpora (e.g., Chen 2010), newswire corpora (e.g., Partington and Morley 2004), and historical corpora (e.g., Culpeper and Kytö 2002), have been used to identify lexical bundles in various registers.

Though the method in Biber et al. (1999) appears to be straightforward, the findings have shed new light on our understanding of formulaic expressions. We are not supposed to regard lexical bundles as mere artifacts of a computational extraction; instead, lexical bundles "represent a psychological association between words and reflect a very real part of users' communicative experiences" (Hyland 2008: 5). A lexical bundle (e.g., *the fact that*, *I was going to*) often functions as a "pragmatic head" (Biber and Barbieri 2007: 270) that expresses stances and/or textual meanings and frames following propositions. Furthermore, some empirical studies have found that lexical bundles may be psychologically real units, stored and processed holistically (e.g., Jiang and Nekrasova 2007, Tremblay et al. 2009). To sum up, instead of employing linguistic criteria to identify multi-word combinations, the Biberian approach conducts an exploratory corpus analysis to extract lexical bundles that would go unnoticed otherwise. Interestingly enough, the frequency-based list of recurrent word sequences turns out to adequately reflect how the phraseological resources in our mental repertoire fulfill communicative needs faced by language users in their daily interactions (Biber et al. 2004).

Previous studies have tended to focus on lexical bundles in English, with only a few examining lexical bundles in Spanish (Tracy-Ventura et al. 2007, Cortes 2008) and Korean (Kim 2009). As Aijmer and Altenberg (1996) suggest, cross-linguistic comparisons can provide important insights that often pass unnoticed in monolingual corpus analyses. More cross-linguistic investigations are needed to "understand the role of typology in different languages on the occurrence of lexical bundles" (Kim 2009: 158). However, few studies, if any, have been conducted to identify lexical bundles in Chinese, which is typologically distinct from English and can definitely provide valuable insights. Most phraseological studies on the full range of multi-word

doi:10.6342/NTU201600328

combinations in Chinese have been concerned with idiomatic expressions (e.g., Zhou 1994, Zhou 1997, Cui 2006, Wang 2009).

1.2 Research Questions

The present study adopts the Biberian approach (Biber et al. 1999) for Chinese, with the general aim of investigating the use of lexical bundles in Chinese. We will identify lexical bundles in two text types, i.e., conversation and news. Specific research questions are posed below.

Although the Biberian approach has been widely adopted to identify lexical bundles, some problems remain to be solved. First, though a dispersion threshold of occurring in at least five texts is set to guard against individual idiosyncrasies, most recurrent word sequences exceed such a low threshold (Conrad and Biber 2004). The question of “what if a cluster is very frequent in not just one but several separate texts yet absent in all others in a corpus” (Partington and Morley 2004: 8) has not been adequately answered. Second, some recurrent word sequences composed purely of high-frequency function words do not have identifiable functions, and a word association threshold may be needed to filter them out (Salazar 2014). To address the above methodological issues, the present study will adopt a more sensitive dispersion measure (Gries 2008b) and a newly developed word association measure (Wei and Li 2013). To what extent do these two additional quantitative measures help to minimize manual interventions that are needed to exclude individual idiosyncrasies and semantically/pragmatically incoherent word sequences? Additionally, an exploratory data analysis will be conducted to examine the distributional patterns of the quantitative measures of lexical bundles in Chinese.

After lexical bundles in conversation and news have been identified, the present study will address the following questions.

doi:10.6342/NTU201600328

(i) Do lexical bundles in Chinese have strong structural correlates?

(ii) What communicative functions do lexical bundles in Chinese serve?

(iii) How are lexical bundles in Chinese distributed in terms of their forms and functions?

(iv) Is there any relationship between the structural and functional categories of lexical bundles in Chinese?

If the findings in English are replicated in the present study, the results in Chinese will lend cross-linguistic support for the theoretical status of lexical bundles. Meanwhile, given that Chinese is structurally distinct from English, some differences between Chinese and English in the use of lexical bundles are expected. Furthermore, a comparison between spoken bundles and news bundles in Chinese will be made. The results may reflect essential differences in communicative needs different text types fulfill.

1.3 Organization of the Dissertation

The dissertation is organized in the following way. Chapter 2 first gives a brief overview of phraseological approaches to language and then zooms in to review studies on lexical bundles as well as phraseological studies in Chinese. Chapter 3 introduces both the method for identifying lexical bundles in Chinese and the analysis procedure that follows. Chapter 4 describes how we arrive at the decisions on the quantitative thresholds, evaluates the bundle identification method, and conducts an exploratory data analysis of lexical bundles in Chinese. Chapter 5 presents the structural and functional distributions of spoken bundles in Chinese and makes a cross-linguistic comparison between Chinese and English in the use of lexical bundles. Chapter 6 presents the structural and functional distributions of news bundles in Chinese and contrasts the use of lexical bundles in conversation and news. Chapter 7

doi:10.6342/NTU201600328

is the conclusion, summarizing the main findings, highlighting theoretical contributions, suggesting some practical applications of the present study, pinpointing a number of limitations, and offering some suggestions for further research on lexical bundles in Chinese.

doi:10.6342/NTU201600328

doi:10.6342/NTU201600328

Chapter 2

Literature Review

A research timeline (Wray 2013) for the umbrella concept *formulaic language* has been reviewed in the introduction chapter. In the current chapter, three main approaches to phraseology/phraseologisms will be reviewed (Section 2.1). Then, I will zoom in to review previous studies on lexical bundles, most of which are concerned with lexical bundles in English (Section 2.2). Finally, phraseological studies in Chinese will be reviewed (Section 2.3).

2.1 Phraseological Approaches to Language

One of the most often cited definitions for the term *phraseology*, i.e., “the study of the structure, meaning and use of word combinations” (Cowie 1994: 3168), reflects a syntagmatic perspective of lexical patterns. The idea that the regularities of a word or a phrase can be observed in its contexts actually emerges long before the era of modern corpora (e.g., Firth 1957). Leńko-Szymańska (2014) summarizes three main approaches to multi-word combinations. The first approach, a traditional one, is considered to be top-down, adopting a set of linguistic criteria to distinguish different kinds of multi-word combinations. Then, the combination of large-scale corpora, concordancing programs, and statistical methods has made it possible for corpus linguists to uncover how words are combined with each other. Nowadays a substantial number of corpus-based studies are taking the second approach, a bottom-up one, aiming to identify multi-word combinations based on their distributions in the corpus. The third approach adopts a psycholinguistic perspective, exploring the way multi-word combinations of different kinds are stored and processed in the mind. These approaches have complemented each other and enriched our understanding of

doi:10.6342/NTU201600328

the nature of multi-word combinations.

Traditionally, the first approach focuses on the distinction of multi-word combinations of various kinds (see Granger and Paquot 2008). For example, Cowie (1981) provides a taxonomy for five types of multi-word combinations, as presented in the following table. An important implication of Cowie's (1981) taxonomy is that multi-word combinations in the category COMPOSITES form a phraseological continuum, from the most transparent and variable (i.e., restricted collocations) to the most opaque and fixed (i.e., pure idioms). Mel'čuk (1998) also proposes a classification of multi-word combinations. In this tradition, identifying the most idiomatic units is regarded as identifying the core of phraseological units and establishing a research discipline in its own right. However, the term *phraseology* is no longer strictly restricted to idiomatic expressions nowadays.

doi:10.6342/NTU201600328

Table 2.1. Cowie's (1981) taxonomy of multi-word combinations.

Category	Subcategory	Definition	Example
composites	restricted collocations	(i) have only syntactic and semantic restrictions (ii) have a figurative meaning in one of the elements (iii) verb-noun combinations with a delexical verb	*blow a trumpet*, *heavy rain*, *make a comment*
	figurative idioms	have a figurative meaning but also preserve a literal interpretation	*blow your own trumpet*, *do a U-turn*
	pure idioms	semantically non-compositional to a great extent	*blow the gaff*, *spill the beans*
formulae	routine formulae	perform speech acts	*good morning*, *see you soon*
	speech formulae	(i) organize messages (ii) indicate speaker/writer attitudes	*are you with me*, *you know what I mean*

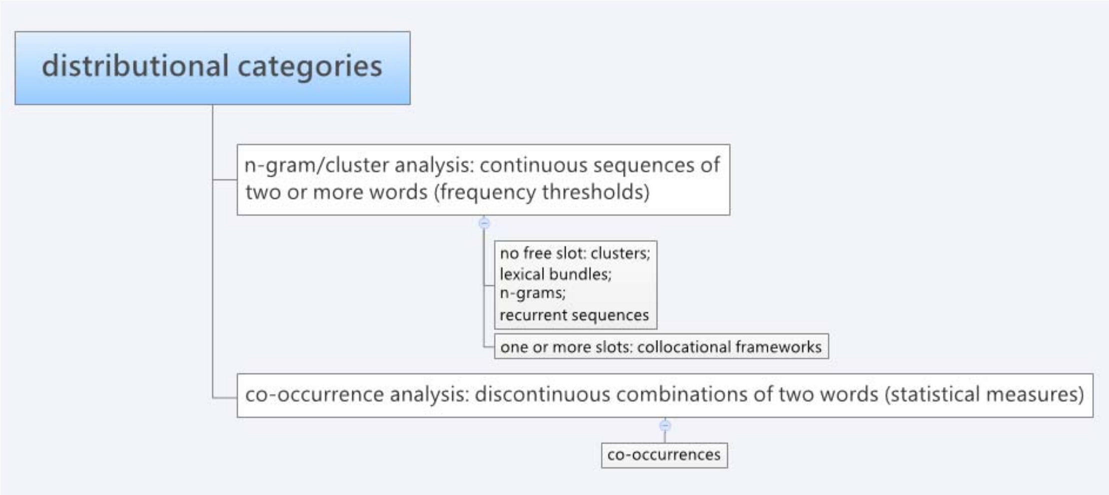
The second approach is generally corpus-based, but studies taking this approach differ considerably. Biber et al. (2003) suggest that corpus-based studies on multi-word combinations differ in at least five ways: (i) the research goals: some studies aim to identify and describe the full range of multi-word combinations in a given register, while others adequately describe selected multi-word combinations; (ii) the methodologies used to identify multi-word combinations: various methods (e.g., salience criteria, frequency thresholds) have been adopted to identify multi-word combinations; (iii) the particular kinds of multi-word combinations considered in the study: some focus on two-word collocations (e.g., *extremely rare*, *greatly appreciated*), others on continuous sequences (e.g., *there is no doubt that*), still others on discontinuous frames (e.g., *a* + __________ + *of*), and so forth; (iv) the representativeness of the corpus used for the analysis: some studies are based on a relatively small corpus (approximately 100,000 words), while others are based on a very large corpus (more than 100,000,000 words); (v) the inclusion of register comparisons: some studies do not consider register differences, while others explicitly compare multi-word combinations in different registers.

Gries (2008a) also presents a list of six dimensions that corpus-based phraseological studies need to take on in their methodologies. These dimensions include: (i) the nature of the elements observed (e.g., lexical items or more general categories); (ii) the number of collocates *l* that make up the collocation (e.g., two-word collocations, lexical bundles); (iii) the number of times *n* an expression must be observed before it counts as a collocation (e.g., raw frequencies or other statistics); (iv) the distance and/or (un)interruptability of the collocates; (v) the degree of lexical and syntactic flexibility of the collocates involved (e.g., word forms, lemmas); (vi) the role that semantic unity and semantic non-compositionality/non-predictability play in the definition.

doi:10.6342/NTU201600328

As can be seen, some dimensions in Gries (2008a) have an approximate equivalent in Biber et al. (2003). For example, a phraseological study always needs to decide on the length of multi-word combinations to be identified. Granger and Paquot (2008) also create a typology for multi-word combinations extracted in different ways, as presented in the following figure, and their typology echoes the above two studies.

Figure 2.1. Granger and Paquot's (2008) typology of multi-word combinations extracted in different ways.


distributional categories
n-gram/cluster analysis: continuous sequences of two or more words (frequency thresholds)
no free slot: clusters; lexical bundles; n-grams; recurrent sequences
one or more slots: collocational frameworks
co-occurrence analysis: discontinuous combinations of two words (statistical measures)
co-occurrences

The third approach to multi-word combinations adopts a psycholinguistic perspective, addressing the issue of how multi-word combinations are stored and processed in the mind. Various techniques have been used, including eye-tracking experiments (e.g., Schmitt et al. 2004), grammaticality judgments (e.g., Jiang and Nekrasova 2007), gap-filling activities (e.g., Nekrasova 2009), to name just a few. In many of these studies, not only native speakers but also non-native speakers are observed. It has been suggested that formulaic sequences are stored and processed holistically, but the evidence for this has not been regarded as conclusive. Moreover, though the frequency of a word sequence has been found to be a crucial factor that determines its formulaic status in the mental lexicon, other factors such as the discourse function of a word sequence have also been found to be influential. There

doi:10.6342/NTU201600328

are also a large number of acquisition studies investigating non-native speakers' use of multi-word sequences in different registers (e.g., academic essays, classroom interactions). Since target sequences used in these studies are mostly lexical bundles, a more detailed review will be presented in the next section.

In the current section, three approaches to multi-word combinations are reviewed. Even studies taking the same approach can adopt significantly different methodologies. No single approach or methodology can put together the whole picture of multi-word combinations, and empirical findings from diverse perspectives need to be synthesized to advance our understanding of the importance of multi-word combinations in our language use (Biber et al. 2003: 72).

2.2 Lexical Bundles

Before the term *lexical bundle* is coined in Biber et al. (1999), there already exist some corpus-based studies on recurrent patterns or word sequences in our discourse. Altenberg and Eeg-Olofsson (1990) and Altenberg (1998) are the first large-scale investigations that have been made of recurrent word sequences in spoken English.[4] The data is from the London-Lund Corpus, which consists of nearly 500,000 words, and word sequences considered to be phraseologically uninteresting based on structural criteria (e.g., mere repetitions such as *the the*, fragments of larger structures such as *in a*) are eliminated from further analyses. A structural framework is proposed, and a functional classification is also proposed for each structural type. For example, it is found that clause constituents are the most common structural type, and they can

[4] Before Altenberg and Eeg-Olofsson (1990) and Altenberg (1998), Manes and Wolfson (1981) focus on 668 compliments recorded from everyday interactions, concluding that compliments in American English are formulaic. For example, only a restricted set of adjectives (e.g., *nice* and *good*) and verbs (e.g., *like* and *love*) is used in compliments, and the most common syntactic pattern is 'noun phrase + *is*/*looks* (*really*) + adjective'. It is argued that the formulaic nature of compliments helps reinforce or create the solidarity among the speakers and makes compliments more readily identifiable wherever they occur in the discourse.

doi:10.6342/NTU201600328

be used as vagueness tags (e.g., *and so on*), qualifying expressions (e.g., *more or less*), connectors (e.g., *on the other hand*), and so forth. Some common collocational frameworks are also identified (e.g., ‘*as* + adverb + *as*’, ‘*a*(*n*) + noun + *of*’). Butler (1997) identifies repeated two-word to five-word combinations in spoken and written Spanish, with the frequency threshold set at 10 times for the smaller corpora and 30 times for the largest corpus (c. 950,000 words).[5] It is found that there are fewer repeated multi-word combinations in written Spanish than in spoken Spanish. It is also found that frequent word sequences in spoken Spanish are largely of an interpersonal nature (e.g., expressing the speaker’s attitude), whereas those in written Spanish are largely of an ideational nature. Interestingly, these findings are later echoed in many studies on lexical bundles in English. Since the late 1990s, an increasing number of studies in the field of second language acquisition have been devoted to second language learners’ use of multi-word combinations. De Cock (1998) is the first quantitative comparison between native and non-native speakers, collecting data from the Louvain International Database of Spoken English Interlanguage (LINDSEI). A recurring pedagogical implication in the literature is that language teachers “have little understanding of the phraseological mechanisms of the language” (Howarth 1998: 186).

Biber et al. (1999) conduct an even larger-scale investigation of recurrent word sequences in English than Altenberg and Eeg-Olofsson (1990) and Altenberg (1998). The Longman Spoken and Written English Corpus is adopted, and two registers are chosen: conversation (British English: c. 4,000,000 words; American English: c. 3,000,000 words) and academic prose (c. 5,300,000 words).[6] Having been widely adopted, the Biberian approach defines lexical bundles as “recurrent expressions,

[5] Butler (1997) consults five corpora.

[6] For a detailed description of the Longman Spoken and Written English Corpus, refer to the first chapter of Biber et al. (1999).

regardless of their idiomaticity, and regardless of their structural status" (Biber et al. 1999:990). For instance, *I don't know what*, *I don't want to*, *I was going to*, *do you want to*, and *are you going to* are the most common four-word bundles in conversation, occurring over 100 times per million words. As can be seen in the definition and the above examples, a lexical bundle is not necessarily a structurally complete unit. Only 15% of the lexical bundles in conversation can be regarded as structurally complete units; in academic prose, the percentage is even lower, i.e., less than 5%. Another important characteristic of lexical bundles is that they are semantically transparent from the constituents and serve important discourse functions (e.g., Biber et al. 2003, Hyland 2008).

The method for identifying lexical bundles is generally frequency-based. A computer program reads each sentence in the corpus and proceeds one word at a time, automatically extracting three-word, four-word, five-word, and six-word sequences. For example, the sentence *But I don't know why we're talking about this* would have the following four-word sequences extracted by the program:

(2.1) But I don't know
I don't know why
don't know why we
know why we're talking
why we're talking about
we're talking about this

Note that two-word contractions (e.g., *we are* → *we're*, *do not* → *don't*) are treated as single words since the method relies on orthographic words, and that lexical sequences running across a turn boundary or a punctuation mark are not included.

doi:10.6342/NTU201600328

Then, the program sorts all the sequences extracted from the corpus and creates a frequency table to store the results. A frequency threshold of at least ten occurrences per million words is set for a multi-word sequence to be recognized as a lexical bundle.[7] Furthermore, a dispersion threshold of occurring in at least five different texts in the corpus is also set, with a view to guarding against individual idiosyncrasies or ongoing topics.

Through a quantitative analysis of lexical bundles in conversation and academic prose, many intriguing distributions emerge. First, it is confirmed that lexical bundles are extremely common in both conversation and academic prose: three-word bundles occur over 80,000 times per million words in conversation, and over 60,000 times per million words in academic prose; four-word bundles occur over 8,500 times per million words in conversation, and over 5,000 times per million words in academic prose. Second, although lexical bundles are extremely common, only a few of them occur with a relatively high frequency (e.g., *I don't know*, *in order to*). Third, both in conversation and in academic prose, the number of three-word bundles is almost ten times larger than that of four-word bundles.

Moreover, lexical bundles are grouped according to their structural correlates. However, it is noted that the structural categories here are not always mutually exclusive. For instance, some bundles in the category 'personal pronoun + verb phrase' (e.g., *I don't know why*) contain a fragment of a *wh*-clause, so these bundles can also be listed in the category 'lexical bundles with *wh*-clauses'. The structural categories for lexical bundles in conversation and in academic prose are summarized in the following table.

[7] Since five-word and six-word bundles are generally much less common, a lower threshold of at least five occurrences per million words is adopted.

doi:10.6342/NTU201600328

Table 2.2. Structural categories for lexical bundles in conversation and in academic prose (Biber et al. 1999).
(Those in bold are shared by the two registers.)

Categories in conversation	Categories in academic prose
1. personal pronoun + lexical verb phrase (+ complement-clause fragment) (e.g., *well I don't know*, *I don't think he*) 2. **pronoun/noun phrase + be +** (e.g., *it's going to be*, *there's a lot of*) 3. verb phrase with active verb (e.g., *let's have a look*, *see if I can*) 4. *yes-no* question fragments (e.g., *do you want to*, *did you have a*) 5. *wh*-question fragments (e.g., *what are you doing*, *what do you mean*) 6. lexical bundles with *wh*-clause fragments (e.g., *don't know what it*) 7. lexical bundles with *to*-clause fragments (e.g., *don't want to go*) 8. **verb + *that*-clause fragment** (e.g., *thought it was a*, *said I don't know*) 9. **adverbial clause fragments** (e.g., *if you want to*, *as long as you*) 10. noun phrase expressions (e.g., *the end of the*, *or something like that*) 11. prepositional phrase expressions (e.g., *at the end of*, *at the same time*) 12. quantifier expressions (e.g., *all the rest of*) 13. other expressions (e.g., *no no no no*, *on and on and on*) 14. meaningless sound bundles (e.g., *da da da da*)	1. noun phrase with *of*-phrase fragment (e.g., *the end of the*, *the size of the*) 2. noun phrase with other post-modifier fragment (e.g., *the way in which*, *the fact that the*) 3. prepositional phrase with embedded *of*-phrase fragment (e.g., *as a result of*, *in the form of*) 4. other prepositional phrase (fragment) (e.g., *at the same time*) 5. anticipatory *it* + verb/adjective phrase (e.g., *it can be seen that*, *it is possible to*) 6. passive verb + prepositional phrase fragment (e.g., *is based on the*) 7. copula *be* + noun/adjective phrase (e.g., *is one of the*, *is due to the*) 8. **(verb phrase +) *that*-clause fragment** (e.g., *should be noted that*) 9. (verb/adjective +) *to*-clause fragment (e.g., *are likely to be*) 10. **adverbial clause fragments** (e.g., *as we have seen*, *if there is a*) 11. **pronoun/noun phrase + *be* +** (e.g., *this is not the*, *there has been a*) 12. other expressions (e.g., *as well as the*, *may or may not*)

doi:10.6342/NTU201600328

The data in Biber et al. (1999) are examined more closely in Biber et al. (2003), and an initial effort is made to propose a functional framework for lexical bundles. As can be seen above, the structural correlates of lexical bundles in conversation are significantly different from those in academic prose, and this corresponds to functional differences between the two registers. In conversation, approximately 90% of lexical bundles are clausal fragments, and these bundles are usually used to express the speaker's stance (e.g., *oh I don't know*) or organize conversational interactions (e.g., *what do you mean*). On the other hand, in academic prose, most bundles are phrasal rather than clausal, and they are usually used to make direct references (e.g., *the end of the hallway*) or identify logical relationships (e.g., *on the other hand*).[8]

Conrad and Biber (2004) still examine the data in Biber et al. (1999), with the aim of updating the functional framework in Biber et al. (2003). As presented in the following table, there are four functional categories in the updated framework. The findings of Biber et al. (2003) are confirmed. With respect to discourse functions, common bundles in conversation and those in academic prose show almost entirely non-overlapping distributions: bundles in conversation usually serve as stance expressions and discourse organizers, while bundles in academic prose are mainly referential expressions. This framework has been adopted in many follow-up studies (e.g., Biber et al. 2004).

[8] Biber et al. (2003) analyze only four-word bundles.

Table 2.3. Functional framework of lexical bundles in conversation and academic prose (Conrad and Biber 2004).

I. **STANCE EXPRESSIONS**: express attitudes or assessments that provide a frame for the interpretation of the following proposition

(1) **epistemic stance**: comment on the knowledge status of the information in the following proposition (e.g., certain, uncertain, or probable/possible)

■ personal (e.g., *I don't know if*, *I think it was*)

■ impersonal (e.g., *the fact that the*)

(2) **attitudinal/modality stance**: express speaker attitudes towards the actions or events described in the following proposition

(i) desire (e.g., *if you want to*, *I don't want to*)

(ii) obligation/directive

■ personal (e.g., *you might want to*, *do you want me*)

■ impersonal (e.g., *it is important to*, *it is necessary to*)

(iii) intention/prediction

■ personal (e.g., *I'm not going to*, *are we going to*)

■ impersonal (e.g., *it's going to be*, *going to have a*)

(iv) ability (e.g., *to be able to*, *can be used to*)

II. **DISCOURSE ORGANIZERS**: reflect relationships between prior and coming texts

(1) **topic introduction/focus**: provide an overt signal that a new topic (or subtopic) is being introduced or is becoming the focus (e.g., *what do you think*, *if you look at*)

(2) **topic elaboration/clarification**: add more information to a topic (e.g., *nothing to do with*), clarify or seek further clarification of what is stated previously (e.g., *what do you mean*), or overtly mark the relationship the speaker/writer sees between discourse units (e.g., *on the other hand*)

doi:10.6342/NTU201600328

Table 2.3. Functional framework of lexical bundles in conversation and academic prose (Conrad and Biber 2004) (continued).

III. **REFERENTIAL EXPRESSIONS**: make direct references to physical or abstract entities or to textual contexts

(1) identification/focus: identify an entity or part of it as noteworthy (e.g., *that's one of the*, *of the things that*, *one of the most*)

(2) imprecision: communicate that the previous text is imprecise (e.g., *or something like that*, *and stuff like that*)

(3) specification of attributes: focus on some particular attribute of the entity

(i) quantity (e.g., *there's a lot of*, *how many of you*, *per cent of the*)

(ii) tangible framing attributes (e.g., *the size of the*, *in the form of*)

(iii) intangible framing attributes (e.g., *the nature of the*, *in the case of*)

(4) time/place/text reference

(i) place reference (e.g., *in the United States*)

(ii) time reference (e.g., *at the same time*, *at the time of*)

(iii) text deixis (e.g., *as shown in figure*)

(iv) multifunctional reference (e.g., *the end of the*, *the beginning of the*)

IV. **SPECIAL CONVERSATIONAL FUNCTIONS**: occur only in the conversation subcorpus

(1) politeness routines (e.g., *thank you very much*)

(2) simple inquiry (e.g., *what are you doing*)

(3) reporting (e.g., *I said to him*)

The method proposed in Biber et al. (1999) has been widely adopted to identify lexical bundles in a corpus. However, different studies may modify the method to serve their research purposes or to overcome the limitations of the corpus. The methods of the follow-up studies are summarized in the following table, which is admittedly not an exhaustive list, and each study will be reviewed immediately afterwards. Methodological concerns for the identification of lexical bundles will be discussed in more detail in the next chapter, which is devoted to the methodology of the present study.

doi:10.6342/NTU201600328

Table 2.4. Methods for identifying lexical bundles (sorted in chronological order).

Study	Language	L1/L2	Length (words)	Frequency threshold (times per million words, if not specifically specified)	Dispersion threshold	Other criteria	Database (e.g., corpus, size, registers)
Biber et al. (1999)	English	L1	3, 4, 5, 6	10 (3-, 4-word bundles); 5 (5-, 6-word bundles)	5 texts	NA	The Longman Spoken and Written English Corpus ■ conversation: c. 4,000,000 words (British English); c. 3,000,000 words (American English) ■ academic prose: c. 5,300,000 words
Cortes (2002)	English	L1	4	20	5 texts	NA	a self-built corpus ■ freshman compositions: 360,704 words
Culpeper and Kytö (2002)	English	L1	3	recur at least 10 times	3 texts	only consider the top 50-ranked bundles in each data set	The Corpus of English Dialogues 1560-1760 ■ late trials: 211,426 words; early trials: 40,727 words ■ late comedy drama: 104,494 words; early comedy drama: 102,817 words
Biber et al. (2004)	English	L1	4	40	5 texts	NA	The TOEFL 2000 Spoken and Written Academic Language Corpus (T2K-SWAL) ■ university classroom teaching: c. 1,248,800 words ■ university textbooks: c. 760,600 words
Cortes (2004)	English	L1, L2	4	20	5 texts	NA	a self-built corpus ■ published academic writing (history): 966,187 words ■ published academic writing (biology): 1,026,344 words ■ student writing (history): 493,109 words ■ student writing (biology): 411,267 words
Partington and Morley (2004)	English	L1	2, 3, 4, 5, 6, 7	occur more than 3 times	NA	NA	The Newspool Corpus ■ editorials: c. 500,000 words ■ press briefings: c. 250,000 words ■ political news interviews: c. 250,000 words

doi:10.6342/NTU201600328

Nesi and Basturkmen (2006)	English	L1	4	10	NA	NA	The British Academic Spoken English Corpus: 882,980 words The Michigan Corpus of Academic Spoken English: 387,818 words
Biber and Barbieri (2007)	English	L1	4	40	3 texts	NA	T2K-SWAL ■ spoken (5 registers): ranging from 39,255 words to 1,248,811 words ■ written (3 registers): ranging from 52,410 words to 760,619 words
Cortes and Csomay (2007)	English	L1	3, 4, 5	20	NA	structural and idiomatic coherence	The Michigan Corpus of Academic Spoken English ■ university speech: c. 1,700,000 words (200 hours) three comparison corpora ■ The Corpus of Spoken Professional American English ■ The Bank of English National Public Radio ■ The Switchboard Corpus
Tracy-Ventura et al. (2007)	Spanish	L1	4	30	20 texts	NA	a self-built corpus ■ sociolinguistic interviews: 2,222,025 words ■ academic texts: 1,002,550 words
Cortes (2008)	English; Argentinian Spanish	L1	4	20	5 texts	NA	a self-built corpus: academic writing (history) ■ English: 1,001,012 words ■ Spanish: 1,003,264 words
Hyland (2008)	English	L1	3, 4, 5	20	10% of texts	NA	a self-built corpus ■ 4 academic disciplines by 3 text types: ranging from 107,700 words to 670,000 words
Kim (2009)	Korean	L1	3	20	5 texts	NA	The Spoken and Written Sejong Corpus ■ conversation: 2,604,054 words ■ academic texts: 3,407,020 words
Chen (2010)	English	L1	4	20	5 texts	NA	The Electrical Engineering Introductory Textbook Corpus: 247,346 words The English for Specific Purposes Textbook Corpus: 99,774 words

Chen and Baker (2010)	English	L1, L2	4	25	3 texts	exclude complete overlaps and complete subsumptions	The Freiburg-Lancaster-Oslo/Bergen Corpus ■ native expert writing: 164,742 words The British Academic Written English Corpus ■ native peer writing: 155,781 words ■ learner writing: 146,872 words
Wood (2010)	English	L1	4	20	NA	NA	a self-built corpus compiled from six textbooks ■ a textual subcorpus: 187,959 words ■ an instructional subcorpus: 391,386 words
Kopaczyk (2012)	Middle Scots	L1	3, 4, 5, 6, 7, 8	occur more than 10 times	10 texts	NA	a compilation of legal and administrative texts: c. 600,000 words ■ The Edinburgh Corpus of Older Scots ■ The Helsinki Corpus of Older Scots ■ an unpublished transcript of a burgh court book from the south-west of Scotland
Leńko-Szymańska (2014)	English	L2	3	7.6 (in COCA)	5 texts or more in any of the learner data sets	NA	■ target bundles: The Corpus of Contemporary American English (c. 425,000,000 words) ■ learner bundles: The International Corpus of Crosslinguistic Interlanguage (6 native languages by 3 proficiency levels; ranging from 4,023 words to 16,089 words)
Salazar (2014)	English	L1, L2	3, 4, 5, 6	10	NA	MI > 0.5; other exclusion criteria (e.g., fragments of other bundles, topic-specific bundles)	■ target bundles: sample texts from the Health Science Corpus (2,082,409 words) ■ non-native bundles: a self-compiled corpus (120,718 words)

As can be seen from the above table, studies on lexical bundles vary greatly in many dimensions of their methods (e.g., the frequency threshold, the corpus size). Also noteworthy is that there are several lines of studies centering around lexical bundles.

First, many studies are concerned with the use of lexical bundles in the academic arena, and most of them consider register differences. Biber et al. (2004) examine corpus texts from university classrooms and textbooks. The use of lexical bundles in conversation, academic prose, university classrooms, and textbooks is compared. The overall patterns are summarized in the following table.

Table 2.5. Functional distribution of common lexical bundles in four registers.

	Stance bundles	Discourse organizers	Referential bundles
Conversation	29	10	3
University classrooms	33	19	32
Textbooks	4	3	20
Academic prose	3	1	15

The extremely high density of lexical bundles in university classrooms is attributed to the fact that classroom teaching relies heavily on both oral bundles and literate bundles. Though being an information-oriented register, classroom teaching is similar to conversation and features many bundles that are fragments of declarative and interrogative clauses. On the other hand, textbooks are like academic prose, featuring many bundles that are noun phrases and prepositional phrases. Furthermore, a strong relationship between the structure and the function of a lexical bundle is revealed again: for example, a noun phrase serves mainly as a referential bundle, and so does a prepositional phrase.

doi:10.6342/NTU201600328

Biber and Barbieri (2007) further extend Biber et al. (2004) and investigate the use of lexical bundles in a wide range of spoken and written university registers: the former includes classroom teaching, classroom management, office hours, study groups, and service encounters, and the latter includes textbooks, course management, and institutional writing. It is found that each register relies on different functional categories. For example, though both being written registers, course management and institutional writing show significantly different functional distributions: over 50% of the lexical bundles in course management are stance bundles (e.g., *students are expected to*, *you are encouraged to*), while nearly 70% of the lexical bundles in institutional writing are referential bundles (e.g., *the first day of*, *the semester in which*). This reflects that the use of lexical bundles is influenced not simply by the physical mode (i.e., spoken versus written) but also by the communicative needs of different registers.

Simpson (2004) also examines lexical bundles in academic speech, identifying lexical bundles typical of academic speech and conducting two within-corpus comparisons (i.e., interactive speech vs. monologic speech, and the speech of professors vs. students). A list of 54 lexical bundles occurring significantly more often in academic speech (e.g., *you can see*, *and so on*, *what I mean*) is provided. Additionally, a functional framework of two main categories, which is summarized in Table 2.6, is proposed: (i) discourse organizing functions, and (ii) interactional functions. These two functions also highlight the dual pragmatic characteristics of academic speech (see also Biber et al. 2004): academic speech is an information-rich genre, so lexical bundles structuring the discourse are common; on the other hand, the interaction between speakers is also important in academic speech, so lexical bundles related to the speakers' interactivity are common as well.

doi:10.6342/NTU201600328

Table 2.6. Selected functions of lexical bundles in academic speech (Simpson 2004).

Function	Subcategory	Example
Discourse organizing functions	focuser, introducing examples	*and in fact*, *what happens is*
	meta-discourse expressions	*when we talk*, *you could say*
	enumerators, temporal sequencers	*the first one*, *the next one*
	contrast and comparison, linking	*exactly the same*, *in the same way*
	cause-effect markers	*so that's why*, *the reason why*
	summarizers	*it turns out* (*that*)
Interactional functions	questions, sentence stems	*what do you mean*, *how do you know*
	explaining, demonstrating	*as you can see*, *I'll show you*
	commands, instructions, advice	*look at this*, *to make sure*
Miscellaneous functions	spatial organizers, locatives	*in this class*, *on the web*
	hedges, mitigators	*in some sense*, *more or less*
	generalizers, vagueness markers	*and so on*, *something like this*

Nesi and Basturkmen (2006) zoom in to investigate lexical bundles used as linking devices in university lectures, a register where the information load is so heavy that connections between prepositions are required to be made clear. It is found that as in university classrooms (see also Biber et al. 2004), both oral bundles and literate bundles occupy a crucial role in university lectures. Like classroom teachers, lecturers face real-time constraints and thus feel the urge to use prefabricated bundles;

doi:10.6342/NTU201600328

on the other hand, a lecture is similar to academic prose in that both are less interactive and more pre-planned. When used as cohesive devices in a lecture, lexical bundles usually signal how an idea or a concept is related to another (e.g., *and this is the*) or signal how a topic or an activity in a lecture is related to another (e.g., *I want to you*). It is also found that lexical bundles often co-occur with conventionally recognized cohesive devices: for example, the bundle *if you look at* is found to co-occur with the linking adverbial *for instance*/*example*.

Cortes and Csomay (2007) then explore the relationship between the position and the function of lexical bundles in university lectures, with the focus on the opening phase. The opening phase of a university lecture "sets students and instructors in the organization of the class about to start, contextualizing the content to be delivered or discussed further" (Cortes and Csomay 2007: 69). It is found that in the opening phase of a university lecture, lexical bundles are often used as discourse organizers to introduce a topic or highlight the forthcoming discourse (e.g., *take a look at*).

Furthermore, disciplinary differences in the use of lexical bundles have also drawn some attention. Cortes (2004) examines lexical bundles in published academic prose from two disciplines, i.e., history and biology. Some bundles are associated more closely with one discipline than with the other. For example, *the power of the*, *the creation of the*, and *in the context of*, all identified in the subcorpus of published history, are strongly related to social events or issues; various stance bundles (e.g., *are likely to be*, *the probability that the*) are used more often in biology to hedge the effect of an affirmation or make it more tentative. Another interesting pattern is that lexical bundles in history are either noun phrases or prepositional phrases, while lexical bundles in biology vary greatly in their structure. Cortes (2004) also compares lexical bundles in publish academic prose and those in student writing. Many lexical bundles frequently occurring in published prose (e.g., *from the perspective of*) are rarely or never used by students. With respect to discourse functions, the way a lexical bundle is used in student writing is sometimes distinctly different from the way it is used in published prose: for example, the bundle *at the same time*, which is usually used for simultaneity in published prose, is often used for addition in student writing.

Hyland (2008) also examines lexical bundles in four disciplines (i.e., electrical engineering, microbiology, business studies, and applied linguistics), and another framework is proposed specifically for lexical bundles in academic writing (see Table 2.7).[9] Many differences in the use of lexical bundles in the four disciplines are observed. For example, nearly half of the lexical bundles in electrical engineering and microbiology (e.g., *was added to the*, *the performance of the*) are research-oriented, describing research methods, specifying some aspects of the research environment, etc., and bundles of this kind usually take the form of ‘noun phrase + *of*’. Based on the findings, Hyland (2008) argues against the assumption that there is a core vocabulary/phraseology in the academic arena (cf. Coxhead 2000, Simpson-Vlach and Ellis 2010), and suggests instead that every discipline has its own phraseology (Swales 1990).

[9] The first two disciplines are applied and pure sciences, while the latter two are social sciences.

doi:10.6342/NTU201600328

Table 2.7. Functional framework of lexical bundles in academic writing (Hyland 2008).

I. RESEARCH-ORIENTED: help writers to structure their activities and experiences of the real world

- **(1) location**: indicate time/place (e.g., *at the same time*, *in the present study*)
- **(2) procedure** (e.g., *the use of the*, *the role of the*, *the purpose of the*)
- **(3) quantification** (e.g., *the magnitude of the*, *a wide range of*, *one of the most*)
- **(4) description** (e.g., *the size of the*, *the surface of the*)
- **(5) topic**: related to the field of research (e.g., *the current board system*)

II. TEXT-ORIENTED: concerned with the organization of the text and its meaning as a message or an argument

- **(1) transition signals**: establish additive or contrastive links between elements (e.g., *on the other hand*, *in addition to the*, *in contrast to the*)
- **(2) resultative signals**: mark inferential or causative relations between elements (e.g., *as a result of*, *it was found that*, *these results suggest that*)
- **(3) structuring signals**: text-reflexive markers which organize discourse units or direct the reader elsewhere in the text (e.g., *in the present study*, *in the next section*, *as shown in figure*)
- **(4) framing signals**: situate arguments by specifying limiting conditions (e.g., *in the case of*, *with respect to the*, *on the basis of*, *in the presence of*)

III. PARTICIPANT-ORIENTED: focus on the writer or the reader of the text

- **(1) stance features**: convey the writer's attitudes and evaluations (e.g., *are likely to be*, *may be due to*, *it is possible that*)
- **(2) engagement features**: address readers directly (e.g., *it should be noted that*, *as can be seen*)

The findings of the studies reviewed above have important implications for language learning and teaching. First, the frequent use of lexical bundles in a wide range of university registers (Biber et al. 2004, Nesi and Basturkmen 2006, Biber and Barbieri 2007) suggests that a second language learner in the university context needs to have an adequate understanding of how lexical bundles function (see also Biber 2006). Second, the finding that a learner sometime misuses lexical bundles (Cortes 2004) suggests that the function of a lexical bundle may not as transparent as expected, and that teachers may not be able to help students to have a good command of lexical bundles simply by exposing them to reading materials. Third, disciplinary differences in the use of lexical bundles (Cortes 2004, Hyland 2008) suggest that a discipline-specific collection of lexical bundles can be of greater relevance to language learners.

The second line of studies on lexical bundles is directly concerned with language learning and teaching and is thus inextricably intertwined with the first line. These studies employ a learner corpus to investigate how (second) language learners use lexical bundles. Cortes (2004), which has been reviewed above, is an example. Earlier, Cortes (2002) creates a corpus of freshman English compositions. It is found that lexical bundles in freshman compositions are structurally similar to those in academic prose (e.g., mostly phrasal rather than clausal), but many bundles in freshman compositions are temporal or locative markers, which also occur frequently in registers other than academic prose. This suggests that the use of formulaic sequences in academic texts can distinguish novice (native) writers from expert writers (see also O’Donnell et al. 2013).

Chen and Baker (2010) examine the use of lexical bundles identified from three corpora of academic writing in English (i.e., published, native students, and non-native students). A quantitative analysis shows that published works exhibit the

doi:10.6342/NTU201600328

widest range of lexical bundles. Besides, native and non-native students are strikingly similar in their use of lexical bundles. For example, their writing features much more discourse organizers than published works, and this is seen as a sign of immature writing. Many other findings here have pedagogical implications, encouraging more emphasis on lexical bundles in curricula and materials.

Leńko-Szymańska (2014) explores the emergence of lexical bundles at the early stages of English learning by examining essays written by non-native students in three grades. Generally, the increase in the use of lexical bundles in the student essays is parallel to the increase in these students' proficiency in English. It is also observed that there seems to be a saturation point beyond which the acquisition of lexical bundles slows down or even does not occur. Additionally, the acquisition of lexical bundles is found to be affected by a learner's learning environment and his/her exposure to a particular register/topic. However, the findings are regarded as tentative and need to be verified with more data.

Salazar (2014) compares lexical bundles identified in the published scientific writing and those identified in essays written by Spanish-speaking learners of English. Not only a frequency threshold but also an association threshold (i.e., the mutual information score) is adopted in the identification of lexical bundles, and a list of exclusion criteria (e.g., fragments of other bundles, such as *on the basis* and *in the case*; topic-specific bundles, such as *amino acid residues* and *the crystal structure*; bundles with random numbers, such as *at least one* and *of the two*) is established to satisfy pedagogical purposes. Besides, lexical bundles sharing components (e.g., *an important role in*, *an essential role in*, *a critical role in*) are treated as variations of a prototypical bundle (i.e., in this case, *a role in*). It is revealed that the way native expert writers use lexical bundles conforms with the conventions of scientific texts. For example, over 30% of the lexical bundles in the expert corpus are noun phrases, and this supports the view that academic writing is noun-centric (e.g., Swales 2008). Many differences exist between expert writers and non-native writers in their use of lexical bundles. The most striking differences are observed in the use of participant-oriented bundles (see Hyland 2008), such as the noticeable underuse of hedging devices (e.g., *it is possible*) in the non-native texts. Several challenges that hinder a successful introduction of lexical bundles in a classroom context are discussed.

Still some studies have directed their attention to lexical bundles in textbooks for English for academic purposes (EAP), with the aim of revealing the gap between the language of EAP materials and the authentic language in the academic world (e.g., Chen 2010, Wood 2010).

To summarize, studies on lexical bundles in learner corpora and textbook corpora are obviously relevant to language learning and teaching (Meunier 2012). Most studies along this line focus on lexical bundles in the academic arena, revealing community-authorized ways of communicating meanings, providing a comprehensive list of target bundles, illustrating essential differences between apprentice writers and expert writers, and arguing for more emphasis on lexical bundles in classrooms and textbooks (Hyland 2012).

The third line of studies on lexical bundles extends to languages other than English. The majority of studies on lexical bundles so far have examined corpus data in English, but studies in other languages, especially those with different typological characteristics, can provide cross-linguistic insights (Aijmer and Altenberg 1996, Johansson 2007). Through cross-linguistic comparisons, different patterns may emerge, and the role of language-specific properties in the distributional patterns of lexical bundles can be better understood.

doi:10.6342/NTU201600328

Tracy-Ventura et al. (2007) examine lexical bundles in Spanish, which are identified in sociolinguistic interviews and academic prose. Two tendencies are shared by both Spanish and English: first, VP-based bundled are typically used as stance expressions and discourse organizers, while NP-based bundles are typically used as referential expressions; second, lexical bundles used in academic prose are mostly referential expressions, while lexical bundles used in conversation are used in various ways. On the other hand, there are also some important differences: first, in Spanish, there are more lexical bundles in academic prose than in conversation; second, more NP/PP-based bundled are identified in Spanish, while more VP-based bundles are identified in English; third, the most common type in spoken Spanish is referential bundles, while stance expressions and discourse organizers are more common in spoken English. Some distributional differences in the use of lexical bundles in these two languages are attributed to their structural differences. First, Spanish relies heavily on *de*-phrase in the modification of a noun, while the two-word construction 'noun + noun' is common in English (e.g., *los procesos de investigation* 'research process'). As a consequence, more NP-based bundles are identified in Spanish than in English, and this contributes to the density of lexical bundles in academic texts of Spanish.[10] Second, since Spanish has gender markings, two NP-based bundles are identified (e.g., *yo creo que el* and *you creo que la*) in cases where only one is identified in English. This explains the dominance of NP-based bundles in Spanish.

Cortes (2008) conducts a comparative analysis of lexical bundles in Spanish and English, and the bundles are identified in a self-compiled corpus of academic writing. It is also found that the number of lexical bundles in the Spanish corpus is more than twice as large as that in the English corpus, and this is attributed to the typological

[10] Tracy-Ventura et al. (2007) identify four-word bundles in the corpus. See Table 2.4 for more methodological details.

doi:10.6342/NTU201600328

differences mentioned above. On the other hand, there are some structural similarities: most bundles in both corpora are prepositional phrases, and the second most frequent type is noun phrases. These support the previous finding that most bundles in academic writing are phrasal.

Kim (2009) identifies lexical bundles in Korean, from both conversation and academic prose. There are fewer lexical bundles in Korean than in English, and this is ascribed to the typological fact that words are highly inflected in Korean. For instance, *gu-run guh gat-ae* ‘(it) seems like it’, one of the most commonly occurring lexical bundle in conversation, has many forms (e.g., *gu-run guh gatah*, *gu-run guh gat-ah-yo*), and the speaker’s choice depends on various factors (e.g., his/her age, the geographical location). A methodological implication is that the frequency threshold for a study on lexical bundles in Korean may need to be lower. Another difference is that in Korean, there are more lexical bundles in academic prose than in conversation.

The fourth line of studies on lexical bundles is concerned with their use in newswire texts. Morley and Partington (2004) examine newspaper editorials, press briefings, and news interviews, and suggest that the use of lexical bundles reflects communicative needs and rhetorical strategies associated with these three registers. For example, the bundle *as we report today* is used to link an editorial to the news report being commented on, and the stance bundle *there can be no* reflects the authoritarian style of editorials. It is also found that the language of press briefings is more formulaic than that of news interviews. A large number of long bundles are identified in press briefings, and many of them are used to avoid difficult or embarrassing questions (e.g., *I’m not aware of any*, *I’m not going to speculate*, *I can’t give you an answer*, *I don’t have any information*, *I’m going to leave it*, *there was no discussion of*). It is demonstrated that superficially similar registers can be differentiated by their use of lexical bundles.

Ansari and Molavi (2013) examine economic and political reports and attempt to reveal their differences in the use of lexical bundles with respect to textual positions.[11] The bundle *according to the* is found to occur in the N2 position, i.e., the second half of any sentence that does not begin a paragraph (see Hoey and O'Donnell 2008), more often in economic reports than in political reports, but the bundle *one of the* shows the opposite trend. However, the conclusion that lexical bundles are primed differently in different disciplines with regard to textual positions is seen as tentative, since only two bundles are considered here.

The fifth line of studies explores the use of lexical bundles in historical corpora. Kopaczyk (2012) summarizes three major challenges to studies along this line, including (i) the relatively small size of a historical corpus, (ii) editorial interventions during the compilation of a historical corpus, and (iii) spelling variants that reflect phonological changes. The first problem has a significant impact on frequency thresholds for identifying lexical bundles (see Chapter 3), and the other two problems can frustrate the automatic identification of every instance of a lexical bundle. Despite these challenges, there have been some intriguing findings. For example, Culpeper and Kytö (2002) observe that the use of lexical bundles in early modern English reflects oral features in written texts, and Kopaczyk (2012) identifies lexical bundles in legal texts from medieval and early modern Scotland. A potential direction along this line is to identify the author or the period of a historical text based on the use of lexical bundles (e.g., Shrefler 2011, Kopaczyk 2013).

Although a large number of studies adopting various corpus-based approaches have demonstrated that lexical bundles can be regarded as building blocks serving important functions in our language use, the assumption that they are stored as

[11] Ansari and Molavi (2013) is not listed in Table 2.4 because only two bundles (i.e., *according to the* and *one of the*) are considered.

doi:10.6342/NTU201600328

unanalyzed chunks in the mental lexicon (e.g., Biber et al. 2004) needs to be verified with psycholinguistic experiments. One of the earliest studies along this line is Schmitt et al. (2004), where a dictation task for native and non-native speakers of English revealed that not all the word sequences frequently occurring in the corpus were reproduced intact. For instance, *in a variety of* was reproduced intact by only 15 native speakers, with 11 participants producing a variation of it. This suggests that corpus-based sequences may not be stored as single units, though the possibility that their holistic storage cannot be revealed in a dictation task cannot be ruled out. The non-native participants' performance was even worse: only four word sequences (i.e., *as a matter of fact*, *in the middle of the, you know, on and off*) were reproduced intact by half or more of the non-native participants. As a result, Schmitt et al. (2004) argue that recurrent sequences and formulaic sequences need to be distinguished: the former are multi-word combinations identified in a corpus but may or may not stored as single units in the mind, while the latter are those stored holistically in the mind but may or may not be identifiable in a corpus analysis.

Some limitations of Schmitt et al. (2004) are pinpointed in Jiang and Nekrasova (2007). First, not all the word sequences in the study (e.g., *to make a long story short*) could qualify as lexical bundles. Second, the word sequences in the study formed a heterogeneous group in many respects. For example, some clusters were from the academic register, while others were from the conversation register. Jiang and Nekrasova (2007) used lexical bundles from corpus-based studies in English (e.g., Biber and Conrad 1999, Cortes 2004) and compared the participants' reaction time in two grammaticality judgment tasks. The results support the holistic representation of lexical bundles in the mind: the lexical bundles were processed faster than the non-formulaic sequences and the ungrammatical sequences. Additionally, as in Schmitt et al. (2004), the native participants still outperformed the non-native

doi:10.6342/NTU201600328

participants. Two possible accounts are provided for the holistic representation of lexical bundles in the mind. First, lexical bundles do not become unanalyzed at a later stage until the language user goes through a frequency-driven chunking process (see also Ellis 2002, 2003). Second, for second language learners, it is also likely that lexical bundles are often introduced as unanalyzed phrases in instruction scenarios.

To further pursue the storage of lexical bundles in the mind, Nekrasova (2009) designed a gap-filling task and a dictation task for native and non-native speakers of English. Generally, some bundles (e.g., *one of the most*, *at the same time*) were found to lean more towards the holistic end than others (e.g., *in the form of*, *in the case of*), and lexical bundles serving as discourse organizers were found to be more familiar to the participants than referential expressions. Still, proficiency effects were observed: in the gap-filling task, the native speakers outperformed the intermediate learners, but did not outperform the advanced learners; in the dictation task, the advanced learners even outperformed the native speakers. There are two implications: first, the representation of a lexical bundle in the mental lexicon can be described in terms of a continuum from a more holistic end to a more compositional end, and multiple criteria should be considered (e.g., how frequently the lexical bundle is appropriately used in a given context, how frequently it occurs in a fixed form); second, different production tasks can reveal different aspects of the mental storage of lexical bundles, and the issue is not necessarily whether a certain task is more accurate than another.

Ellis et al. (2008) not only employed a series of different techniques (i.e., a grammaticality judgment task, a reading aloud task, a priming experiment) but also considered the influence of various variables such as the mutual information (MI) score and the length of a lexical bundle. For native speakers, the MI score is crucial to the processing of a lexical bundle. For non-native speakers, however, the frequency of a lexical bundle is a more dominant variable.

doi:10.6342/NTU201600328

Also with the aim of testing the hypothesis that lexical bundles are stored holistically, Tremblay et al. (2009) conducted a self-paced reading experiment, with the materials presented to the participants in three different manners: i.e., word-by-word, chunk-by-chunk, and sentence-by-sentence. The facilitatory effect of lexical bundles was observed in the second task and the third task, but not in the first task. It is suggested that the holistic nature of a lexical bundle may be disrupted by the word-by-word presentation. Regarding the facilitatory effect of lexical bundles, three accounts are considered. First, a lexical bundle is stored and processed as a single unit. This account gains support from the literature of first language acquisition: children first learn chunks and then analyze them into smaller units at a later stage (e.g., Ellis 1996, 1998). Second, our combinatorial knowledge (i.e., our knowledge of what goes with what) may lie under the facilitatory effect of lexical bundles (see Baayen 2003). In the study, the first-order transitional probability (i.e., $\text{word}_1 \rightarrow \text{word}_2 \rightarrow \text{word}_3$) was controlled, and this variable did not affect the participants' reading time (see also Swinney and Cutler 1979). Nevertheless, it is likely that the second-order transitional probability (i.e., the probability of the occurrence of word_3 after the co-occurrence of word_1 and word_2) or even the third-order transitional probability is playing a role. This appears to be a feasible account since numerous studies have demonstrated language users are good at analyzing linguistic patterns (e.g., Bowers 2005, Libben 2005). The third account is that lexical bundles are stored holistically and our probabilistic knowledge of word strings is also at work.

To sum up, we have seen that a broad variety of corpora (e.g., general corpora, learner corpora, textbook corpora, newswire corpora, historical corpora) have been used to identify lexical bundles. Previous studies characterize the use of lexical bundles in various registers, demonstrating that lexical bundles are not simply an arbitrary output of a frequency-based extraction. The status of lexical bundles in

doi:10.6342/NTU201600328

theoretical linguistics has been recognized, and there are profound implications in many related fields, including language pedagogy (Flowerdew 2012), translation research (Ji 2010), and lexicography (Cowie 1998). Some psycholinguistic studies have also revealed that lexical bundles are stored and processed holistically in the mental lexicon. However, the conclusion is still seen as tentative, and the question of how lexical bundles come to be stored as single units has not been answered adequately. To date, most studies on lexical bundles center around English, especially English for academic purposes (EAP), yet a cross-linguistic perspective (e.g., Tracy-Ventura et al. 2007, Cortes 2008, Kim 2009) can definitely enrich our understanding of the nature of lexical bundles.

2.3 Phraseological Research in Chinese

The current chapter so far has reviewed various approaches to multi-word combinations, with the focus on lexical bundles in English. Now we will proceed to a review on phraseological studies in Chinese.

The first challenge is the issue of how to distinguish between words and phrases in Chinese, which is still under debate (e.g., Lu 1979, Wang 1990, Feng 1997). Some linguists use morphosyntactic criteria, others use semantic criteria, and still others use phonological/prosodic criteria (see Lien 2000). There seems to be no clear boundary between words and phrases in Chinese. In recent years, many studies in the field of natural language processing have adopted complicated computational methods to address this thorny issue (e.g., Han et al. 2000, Cao and Zhou 2004, Wen and Wang 2009, Hu and Tang 2014). These studies, however, aim mainly to develop more precise segmentation tools for Chinese, bearing little relevance to the linguistic perspective on phraseological patterns.

doi:10.6342/NTU201600328

Traditionally, most linguistic research on Chinese phraseology focuses on *shouyu* or *gudingyu*, i.e., idiomatic expressions. Although definitions and coverages vary tremendously from study to study, all research along this line prioritizes fixedness, idiomaticity, syntactic and semantic/pragmatic completeness, and intuitive recognition by native speakers (see Conrad and Biber 2004). Some (e.g., Fu 1985, Huang and Liao 2002) have taken a top-down approach, distinguishing various types of idiomatic expressions (e.g., *chengyu* ‘four-character idioms’, *zhuanming* ‘proper names’, *yanyu* ‘proverbs’, *shuyu* ‘jargon’) and creating a taxonomy that may also include other types of phraseological patterns (e.g., frames). Others have probed into the internal structure of idiomatic expressions (e.g., Lien 1989, Zhou 1994, Zhou 1997, Wang 2009). For instance, Lien (1989) examines four-character compounds in which a pair of antonyms is affixed to a disyllabic word (i.e., antonymous quadrinomials).[12] In fact, idiomatic expressions occur sparsely (Biber et al. 1999, Cortes 2004), so a substantial majority of recurring sequences have been missed in the earlier literature.

A rapidly increasing number of studies have adopted a usage-based approach and conducted corpus-based analyses to examine high-frequency phraseological patterns. For example, Biq (2004) examines the construction ‘V + *ge* + N’ (e.g., *he ge shui* ‘drink some water’), which can be used to express trivialness.[13] Studies of this kind adequately describe selected patterns, reveal many mechanisms in our language use (e.g., metonymy, intersubjectivity, reanalysis, decategorization, grammaticalization), and have pedagogical implications. However, it is not clear exactly how many such

[12] Here is an example:

(a) $[\underline{\textbf{zuo}}_{\text{ANTONYM}}\text{-si}_{\text{VERB}}\text{-}\underline{\textbf{you}}_{\text{ANTONYM}}\text{-xiang}_{\text{VERB}}]_{\text{VERB}}$
left-think-right-think
‘think from different angles’

In (a), the base *sixiang* ‘think’ is a verbal compound formed through the coordination between the bound morpheme *si* ‘think’ and the free morpheme *xiang* ‘think’. The antonymous affixes (i.e., *zuo* ‘left’ and *you* ‘right’) do not change the syntactic category of the resulting quadrinomial. Semantically speaking, *zuo-si-you-xiang* can express “totality” (Lien 1989: 283), i.e., an infinite set of events.

[13] The word *ge* is a general classifier in Chinese.

patterns there are without our conscious notice (Conrad and Biber 2004).

There also exist some teaching materials collecting phraseological patterns useful for second language learners of Chinese (e.g., Wang 1987, Chen and Zhu 2012). These collections are not meant to serve as a comprehensive reference for lexicographical or pedagogical purposes, but aim to highlight patterns second language learners often misunderstand and/or misuse. The decision on which patterns to include is based mainly on the compilers' teaching experience and intuitive judgment. Thus, although the selected patterns are of pedagogical value, the selection process cannot be regarded as scientifically solid.

Recently, efforts have been made to incorporate linguistic and psycholinguistic insights into large-scale extractions of recurrent multi-word combinations from Chinese corpora. Yang (2009) uses quantitative criteria to extract from a large news corpus (c. 80,000,000 words) fixed expressions that are often used as single words. It is suggested that these expressions, referred to as idiom units, are stable in language use and reflect human cognitive rules. An overwhelming majority of idiom units are referential expressions, so it is a pity that many stance expressions and discourse organizers that reflect intriguing dimensions of our language use escape Yang's (2009) attention. Tao (2015) adopts a purely frequency-based method and lists the top 50 three-word chunks in his spoken Chinese data. It is found that most of the component words in these chunks are high-frequency lexemes (e.g., *shi* 'be' and *bu* 'not'). Tao's (2015) approach is quite similar to the Biberian approach, yet only three prominent categories are briefly discussed, i.e., meta-linguistic devices for speaker-addressee interactions (e.g., the *yes-no* interrogative form *shi bu shi*), indefinite expressions involving *yi ge* 'one CLASSIFIER' (e.g., *shi yi ge* 'is a CLASSIFIER'), and epistemic stance markers (e.g., *wo juede wo* 'I feel I').

doi:10.6342/NTU201600328

Li (2014) investigates the use of formulaic expressions in a learner corpus consisting of 300 essays (5,177 words in total) by second language learners of Chinese, and the focus is on three- to six-character discourse markers. A frequency threshold of occurring at least five times in the corpus is adopted; discourse markers that do not exceed the frequency threshold in the learner corpus but achieve a formulaic status in a native corpus are also included in the study. However, the method for identifying formulaic expressions in a native corpus is not adequately described. There are 890 tokens of formulaic discourse markers (342 types) identified in the learner corpus, and there are three main categories: (i) marshalling markers (e.g., *zuotian shangwu* ‘yesterday morning’, *dui henduo ren eryan* ‘to many people’), (ii) connective markers (e.g., *na shihou* ‘at that time’, *tebie shi* ‘in particular’), and (iii) ending markers (e.g., *shenme de* ‘and something like that’, *na jiu hao le* ‘that would be enough’). It is found that the use of formulaic discourse markers reflects a learner’s proficiency level, and this echoes previous findings (see Section 2.2). For example, advanced learners use formulaic discourse markers more frequently than intermediate learners and beginners. Moreover, advanced learners use connective markers more often, whereas intermediate learners and beginners use marshalling markers more often.

To summarize, there have been various approaches to multi-word combinations in Chinese. A general trend is that idiomatic expressions in Chinese appear to have received more attention than multi-word combinations of other kinds. All the findings need to be pieced together so that we can gain a better understanding of Chinese phraseology.

2.4 Summary

Multi-word combinations make up a significant portion in our language use, so they have received a lot of attention. In the current chapter, various approaches and methodologies to explore multi-word combinations have been reviewed (see Section 2.1). Different perspectives are needed since there is a complex taxonomy of multi-word combinations (e.g., two-word collocations, idiomatic expressions, lexical bundles, discontinuous frames).

The major focus of the review is on the Biberian approach, i.e., identifying lexical bundles (i.e., recurrent strings) from a corpus mainly on a frequency-driven basis, and investigating their use in different registers (see Section 2.2). A series of follow-up studies have identified lexical bundles from a broad range of corpora (e.g., general corpora, academic corpora, learner corpora, textbook corpora, newswire corpora, historical corpora). These studies have demonstrated that lexical bundles are not an arbitrary output of a computational extraction; rather, lexical bundles serve vital discourse functions, and their use reflects important characteristics of a register. Some psycholinguistic studies have also suggested that lexical bundles may be stored and processed holistically in the mental lexicon. The findings have far-reaching implications not only for theoretical linguistics but also for related fields such as language education, translation research, and lexicography.

From a cross-linguistic perspective, some attention has been directed to lexical bundles in languages other than English (e.g., Spanish, Korean). However, in the field of Chinese linguistics, most phraseological studies deal with idiomatic expressions or focus on selected phrases or frames, and the use of lexical bundles in Chinese is still an under-researched area.

doi:10.6342/NTU201600328

doi:10.6342/NTU201600328

Chapter 3

Methodology

In the previous chapter, the Biberian approach to multi-word combinations has been reviewed. The identification of lexical bundles is generally frequency-based. This chapter will introduce how their method is modified to identify lexical bundles in Chinese.

3.1 Corpus

A corpus can never be regarded simply as a collection of linguistic data. A corpus is a “social artifact”, as McCarthy (2001:63) has suggested. The corpus is highly influential in any study that aims to identify lexical bundles.

Obviously, the corpus size plays a critical role in the identification of lexical bundles. For instance, if the frequency threshold of occurring at least twenty times per million words is adopted for a corpus of 200,000 words, then a word sequence that occurs only four times in the corpus will be identified as a lexical bundle. It is argued that such a method may be problematic in some registers, and it is suggested that a corpus of at least 1,000,000 words is desirable (Cortes 2002, 2008; Hyland 2012). Even so, due to difficulties encountered in the process of data collection, some studies use a relatively small corpus to identify lexical bundles (e.g., Biber and Barbieri 2007). In such studies, the results need to be interpreted with some caution. As summarized in the literature review (see Table 2.4 in Section 2.2), the corpus sizes in the previous studies on lexical bundles range from thousands of words to millions of words.

Moreover, the contents of a corpus definitely have impacts on the identification of lexical bundles. For example, many history-dependent bundles (e.g., *Mexico and the United States* in the English corpus, *Provincia de Buenos Aires* in the Spanish

doi:10.6342/NTU201600328

corpus) are identified in Cortes (2008), for both corpora consist of academic writing in the discipline of history. In Chen and Baker (2010), it is observed that learner writers use more discourse organizers than expert writers, and this may be attributed to the fact that the articles in the expert corpus are all 2,000-word excerpts, while the texts in the learner corpus are all complete essays that are well-structured.

In view of the issues mentioned above, the present study chooses the Academia Sinica Balanced Corpus of Modern Chinese (the fourth edition), i.e., the Sinica Corpus hereafter.[14] It is sufficiently large for a study on lexical bundles. However, the conversation subcorpus is much smaller than the news subcorpus (i.e., 459,833 words and 6,475,872 words respectively). Another potential issue about the conversation subcorpus is that many conversations are recorded from radio or TV programs.[15] Although these are not typical naturally-occurring data, the speakers can be assumed to behave spontaneously, just as they do in conversations. Still, the findings in the conversation subcorpus need to be interpreted with some caution.

3.2 Automatic Identification of Lexical Bundles in the Corpus

The Biberian approach assumes the concept of word (e.g., the word string *at the same time* is a four-word bundle in English). For languages such as English and Spanish, orthographic words are adopted, i.e., spaces are generally taken as word boundaries.[16] However, in Chinese, the distinction between words and phrases is not always clear (see Section 2.3), and determining word boundaries is never an easy task. In the present study, all the texts in the Sinica Corpus have been automatically segmented by a system developed by the Chinese Knowledge Information Processing

[14] The Academia Sinica Balanced Corpus of Modern Chinese is open to the research community online. It is available at http://asbc.iis.sinica.edu.tw/, where more details about the Sinica Corpus are provided.
[15] Among the 113 conversations, 78 are recorded from radio or TV programs, 35 are interviews, and 2 are talks (but somehow coded as conversations).
[16] For example, words combined with a dash (e.g., *self-control*) and contracted forms (e.g., *don't*) are treated as single words.

doi:10.6342/NTU201600328

(CKIP) Group.[17] Note that not all the segmented texts have been manually checked.

Technically speaking, lexical bundles of any length can be identified; however, most studies focus on three-word and/or four-word bundles. In Conrad and Biber (2004), two-word sequences are not included because many of them are collocations that do not have a distinct discourse-level function. In Cortes (2004), four-word bundles are the focus because they hold many three-word bundles and are much more frequent than five-word bundles. In Hyland (2008), four-word bundles are the focus too, not only because they are far more common than five-word bundles, but also because they offer a clearer range of structures and functions than three-word bundles. Sometimes the scope of the investigation is also a consideration. To obtain a manageable number of lexical bundles, Biber and Barbieri (2007) only identify four-word bundles. Following most studies, the present study focuses on three-word and four-word bundles.[18] As will be shown in the following chapters, three-word and four-word bundles are common in Chinese, and they serve important discourse functions.

The identification of lexical bundles in Chinese is undertaken in R, a free software environment.[19] The program (see Gries 2009) reads through each sentence in the two subcorpora. For each sentence, the program begins with the first word of the sentence and advances one word at a time, identifying and storing all the three-word sequences. Then the same program is used to identify and store all the four-word sequences.

[17] For more details about the segmentation system, refer to http://ckipsvr.iis.sinica.edu.tw/.

[18] In Chinese, almost all the morphemes are monosyllabic, and each syllable is represented by only one character in the writing system. While classical Chinese appears to be a monosyllabic language (i.e., a typical word consists of only one morpheme), modern Chinese has a very large number of disyllabic (compound) words which can usually be broken into two morphemes/words. See Li and Thompson (1981: 10-15) for a more detailed discussion about the relationship among syllables, characters, morphemes, and words in Chinese.

[19] The software R is available at http://www.r-project.org/.

doi:10.6342/NTU201600328

As in previous studies on lexical bundles, only uninterrupted word sequences are regarded as potential bundles. Therefore, word sequences running across a punctuation mark or a turn boundary are excluded in the present study. Although it is possible that some word sequences work over sentence or turn boundaries, Butler (1997) suggests that such sequences are not common.

3.3 Quantitative Thresholds

Studies adopting the Biberian approach to identify lexical bundles are mainly based on frequency data: i.e., a word sequence needs to exceed a certain frequency threshold so that it can be identified as a lexical bundle. In the present study, not only a frequency threshold but also other quantitative thresholds are set to identify lexical bundles in Chinese.

Although any frequency threshold is inevitably criticized as arbitrary, the decision involves many considerations. First, as mentioned in Section 3.1, the size of the corpus is a crucial factor. For a small corpus, a higher frequency threshold may be desirable; otherwise, just a few occurrences of a word sequence would legitimize its status as a lexical bundle. Besides, when a normalized frequency is adopted as a frequency threshold, the rounding of the actual converted raw frequency may substantially revise the original frequency threshold (Chen and Baker 2010).[20] Second, the length of lexical bundles can also be an important factor when we decide on the frequency threshold. For instance, in De Cock (1998), different frequency thresholds are set for word sequences of different lengths (i.e., occurring at least 10, 5, 4, 3 times in the corpus for two-word, three-word, four-word, and five-word

[20] For instance, when the frequency threshold of occurring at least 40 times per million words is applied to a 40,000-word corpus, the converted raw frequency threshold would be 1.6 times in that corpus. For a converted raw frequency to function as an operational threshold, any decimals need to be rounded up or down: i.e., the converted raw frequency threshold in this case would be rounded up to 2 times. However, when the rounded threshold is converted back (i.e., 50 times per million words), the original frequency threshold will be found to have been substantially adjusted.

doi:10.6342/NTU201600328

sequences respectively), so that roughly the same proportion (i.e., 10%-12%) of recurrent sequence types can be identified for each sequence length. Also, in Biber et al. (1999), since five-word and six-word bundles are generally less common, a lower frequency threshold of at least 5 times per million words (as opposed to 10 times per million words for three-word and four-word bundles) is adopted. Third, some studies (e.g., Chen and Baker 2010) consider the limitation of their resources, so a conservative frequency threshold is adopted to obtain a manageable size of lexical bundles for their analysis. As summarized in Table 2.4, frequency thresholds range from 5 times per million words to 40 times per million words. Hyland (2008) suggests that a frequency threshold of occurring at least 20 times per million words can be regarded as conservative. In the present study, the frequency threshold is set at 20 times per million words. The rounding issue mentioned above is not a problem here since both subcorpora are sufficiently large. Additionally, as will be shown in the following chapters, the frequency threshold here is suitable for three-word and four-word bundles and useful in identifying a manageable size of lexical bundles in Chinese.

In addition to a frequency threshold, the Biberian approach to lexical bundles also sets a dispersion threshold to guard against idiosyncrasies used by individual speakers/writers and local repetitions reflecting the immediate topic of the discourse. The dispersion threshold is mostly set at occurring in at least 5 different texts in the corpus (e.g., Biber et al. 1999, Cortes 2002, Biber et al. 2004), though the whole range is from 3 texts to 20 texts (see Table 2.4), with the corpus size being an important factor again. As in most studies, the dispersion threshold of occurring in at least 5 different texts is also adopted in the present study.

However, as Conrad and Biber (2004) admit, the dispersion threshold here may be of little practical effect because most high-frequency word sequences are found to

doi:10.6342/NTU201600328

be widely distributed. For example, most of the bundles identified in Conrad and Biber (2004) occur in more than 30 texts. It is likely that a high-frequency word sequence occurs in several corpus texts yet is absent in most of the corpus texts (Partington and Morley 2004). A more sensitive dispersion threshold is needed for studies on lexical bundles.

There have been many dispersion measures (e.g., Carroll 1970). Gries (2008b) presents a comprehensive survey and then proposes a new dispersion measure, which is referred to as DP hereafter. The measure DP has four appealing characteristics. First, DP is conceptually simple and straightforward. Second, DP can be used even when the corpus is not neatly divided. Third, DP is not restricted to words but is applicable to a wide range of linguistic patterns, such as co-occurrences. Fourth, DP does not blindly output extreme values but can distinguish distributional patterns that other measures fail to. With these strengths, DP is adopted in the present study to complement text counts.

Generally speaking, when we calculate the DP of a word sequence, we consider the difference between the expected proportion and the observed proportion of that word sequence in each corpus part. Here is how the DP of a word sequence is determined. Take the three-word sequence *shi yi ge* 'be one CLASSIFIER' in the news subcorpus, for example. Table 3.1 summarizes the whole procedure. First, the news subcorpus is divided into ten roughly equal parts, as shown in the first column of the table. For example, given that the first corpus part accounts for 9.3% of all the news data, all the occurrences of *shi yi ge* in the first corpus part are supposed to account for 9.3% of its overall occurrences. Second, as shown in the second column, the raw frequency of the word sequence *shi yi ge* in each corpus part is calculated. Then, the third step is to calculate for each corpus part the absolute difference between the expected percentage and the observed percentage, as shown in the third column. The

doi:10.6342/NTU201600328

last step is to sum up all the absolute differences and divide the sum by two, as shown in the last two columns. A DP value always falls between zero and one: the lower the value is, the more evenly dispersed the word sequence is in the corpus.

doi:10.6342/NTU201600328

Table 3.1. Computation of the DP value of the three-word sequence *shi yi ge* ‘be one CLASSIFIER’ in the news subcorpus.

Expected percentage	Observed percentage	Absolute difference	Sum of absolute differences	DP
599,667/6,475,872 = 0.093	108/1,173 = 0.092	|0.093-0.092| = 0.001	0.206	0.206/2 = 0.103
620,416/6,475,872 = 0.096	111/1,173 = 0.095	|0.096-0.095| = 0.001		
637,226/6,475,872 = 0.098	129/1,173 = 0.110	|0.098-0.110| = 0.012		
661,075/6,475,872 = 0.102	161/1,173 = 0.137	|0.102-0.137| = 0.035		
653,741/6,475,872 = 0.101	106/1,173= 0.090	|0.101-0.090| = 0.011		
655,166/6,475,872 = 0.101	78/1,173 = 0.066	|0.101-0.066| = 0.035		
654,488/6,475,872 = 0.101	57/1,173 = 0.049	|0.101-0.049| = 0.052		
670,670/6,475,872 = 0.104	118/1,173 = 0.101	|0.104-0.101| = 0.003		
667,764/6,475,872 = 0.103	174/1,173 = 0.148	|0.103-0.148| = 0.045		
655,659/6,475,872 = 0.101	131/1,173 = 0.112	|0.101-0.112| = 0.011		

The decision on the DP threshold is deferred until Section 4.1, where the distribution of the DP values is presented and discussed in detail. It should be noted that even though the Biberian approach to lexical bundles is interested in uncovering general patterns in our language use, individual idiosyncrasies and local repetitions that are excluded by the dispersion criteria can still be of interest to discourse analysts sometimes (Partington and Morley 2004).

Even when complemented with the dispersion measures, the frequency-based method for identifying lexical bundles still leaves much room for improvement (Salazar 2014).[21] The high frequency of a lexical bundle does not ensure its semantic or pragmatic coherence. The overall high frequency of a lexical bundle may be ascribed simply to the high frequencies of the components, which are often function words. Simpson-Vlach and Ellis (2010) suggest that the mutual information (MI) score is a better indicator of which word sequences have distinctive functions in our language use. Therefore, Salazar (2014) proposes adopting MI scores to screen out high-frequency word sequences that seem to lack identifiable functions.

However, most association measures such as MI scores are confined to measuring the association within two-word sequences. To our knowledge, the latest method of measuring the internal association within multi-word combinations is proposed in Wei and Li (2013), in which the MI measure is refined.[22] The new MI measure is referred to as G hereafter. It is demonstrated that word sequences with the G score higher than three are usually structurally complete and semantically coherent.

The following is how the G score of a word sequence is determined. Let us return to the three-word sequence *shi yi ge* 'be one CLASSIFIER' in the news subcorpus

[21] For more disadvantages of relying on frequency counts to identify other types of multi-word combinations, see Wray (2002).

[22] The association measure used in Salazar (2014) was proposed much earlier, i.e., in 2004. Wei and Li (2013) present an overview of association measures for multi-word combinations and express the need to take the word order into account.

doi:10.6342/NTU201600328

to illustrate the whole procedure, which is presented in (3.1). To begin with, word sequences of various lengths need to be transformed into pseudo-bigrams: for example, a three-word sequence has two dispersion points, i.e., *shi* + *yi ge* and *shi yi* + *ge*. Then, the values needed for the computation of the G score are as follows. All the algorithms here can also be extended to four-word sequences. Still, the decision on the G threshold is deferred until Section 4.2, where the distribution of the G values is presented and discussed in detail.

It should be noted that association measures have limitations. A potential problem is that scores tend to privilege low-frequency words (Biber 2009). However, word sequences with low-frequency components have been filtered out by the frequency threshold. As in Salazar (2014), few negative effects, if any, are observed in the present study as a result of using the G score.

To summarize, four quantitative thresholds are set to identify a candidate list of lexical bundles in Chinese: (i) a frequency threshold (i.e., occurring at least 20 times per million words), (ii) a text count threshold (i.e., occurring in at least 5 different texts), (iii) another dispersion threshold (see Section 4.1), and (iv) a word association threshold (see Section 4.2). In this list still exist some problems that await human judgments. The following section will discuss how these problems can be dealt with.

(3.1) P_{word1} $= P_{shi}$
= the probability of the word *shi* in the corpus
= 90,461/6,475,872 = 0.013969

P_{word3} $= P_{ge}$
= the probability of the word *ge* in the corpus
= 31,628/6,475,872 = 0.004884

$P_{\text{word1 word2}}$ $= P_{shi\ yi}$
= the probability of the sequence '*shi yi*' in the corpus
= 3,627/6,475,872 = 0.000560

$P_{\text{word2 word3}}$ $= P_{yi\ ge}$
= the probability of the sequence '*yi ge*' in the corpus
= 9,759/6,475,872 = 0.001507

$P_{\text{word1 word2 word3}}$ $= P_{shi\ yi\ ge}$
= the probability of the whole sequence in the corpus
= 1,173/6,475,872 = 0.000181

E_1 $= E_{\underline{shi}+\underline{yi\ ge}}$
$= P_{\text{word1}} \times P_{\text{word2 word 3}}$
= 0.013969 × 0.001507 = 2.11e-05

E_2 $= E_{\underline{shi\ yi}+\underline{ge}}$
$= P_{\text{word1 word2}} \times P_{\text{word 3}}$
= 0.000560 × 0.004884 = 2.74e-06

$WAP_{shi\ yi\ ge}$ $= \dfrac{E_1}{E_1+E_2} \times E_1 + \dfrac{E_2}{E_1+E_2} \times E_2$
= 1.89e-05

$G_{shi\ yi\ ge}$ $= MI_{shi\ yi\ ge}$

$$= \log_2\left(\frac{P_{shi\ yi\ ge}}{WAP}\right)$$

= 3.257

3.4 Analysis Procedures

This section will discuss how we deal with some important methodological issues that remain in the current list of lexical bundles. Then, lexical bundles in the finalized list will be classified in two major ways, i.e., according to their structural characteristics and discourse functions.

The first issue is whether the identification of lexical bundles should resist manual interventions based on our intuitive judgment or common sense. Altenberg and Eeg-Olofsson (1990) and Altenberg (1998) adopt structural criteria to eliminate irrelevant word sequences, suggesting that not all recurrent word sequences are phraseologically interesting and that “the sheer bulk of the material makes some sort of selection necessary” (Altenberg 1998: 102). De Cock (1998) even manually excludes literal uses of potential bundles (e.g., *I don't really see them enough you see them an hour*) and refines the frequency data, which is not really feasible in studies based on a large-scale corpus. Simpson’s (2004) list of multi-word combinations includes only word sequences of structural and idiomatic coherence. Chen and Baker (2010) manually exclude word sequences containing content words already present in the essay questions and context-dependent sequences (e.g., *in the UK and*, *the Second World War*). Salazar (2014) excludes ten types of word sequences, such as sequences ending in an article (e.g., *results in a*), bundles with random numbers (e.g., *at least one*), and random section titles (e.g., *figure 4 a*). Such intuition-based criteria are sometimes criticized as subjective (e.g., Hyland 2012). Nevertheless, even though computers identify recurrent patterns based on quantitative criteria, it is the researcher who decides whether the computer-yielded results fulfill the research purpose (e.g., Wray 2002, O’Keeffe et al. 2007).

Since the present study aims to explore the functions of lexical bundles in Chinese, we manually exclude word sequences that do not have identifiable functions.

Most word sequences manually excluded are those composed purely of high-frequency function words (e.g., *le zhe ge* ‘ASPECT.MARKER this CLASSIFIER’). Some word sequences with specific numbers other than one (e.g., *si zhong qingkuang* ‘four type situation’) are also excluded (see also Salazar 2014: 50).

The second issue is that there are two types of overlaps in lexical bundles (Chen and Baker 2010). One is complete overlaps: i.e., shorter lexical bundles are actually derived from longer lexical bundles. For example, the four-word bundles *it has been suggested* and *has been suggested that*, both of which occur 6 times, come from the longer bundle *it has been suggested that*. The other is complete subsumptions, i.e., shorter lexical bundles overlap, and sometimes they are combined together to form a longer lexical bundle. For example, *as a result of*, which occurs 17 times, and *a result of the*, which occurs 5 times, can be combined to form a longer lexical bundle *as a result of the*. In Chen and Baker (2010), overlapping lexical bundles are combined into a longer one to guard against inflated results. However, though aware of the overlapping issue, many previous studies (e.g., Biber et al. 1999, Nesi and Basturkmen 2006) do not deal with it, simply listing all the overlapping lexical bundles separately.

Now we are faced with a dilemma. On the one hand, if we decide to ignore overlaps as most studies do, then the results will be inflated as Chen and Baker (2010) suggest. On the other hand, if we decide to combine two overlapping lexical bundles into a longer one as Chen and Baker (2010) do, then we will be taking the theoretical stance that shorter bundles are not listed separately in the mental lexicon. However, as reviewed in Chapter 2, how lexical bundles are stored and processed in the mental lexicon still remains unanswered (Tremblay et al. 2009). For instance, it is likely that the two four-word bundles *it has been suggested* and *has been suggested that* and the five-word bundle *it has been suggested that* are all separately represented in the mental lexicon. Now we are confronted with a trade-off between producing somewhat inflated results and accepting a still questionable assumption about the mental storage of lexical bundles.

In the present study, a functional perspective is adopted to deal with the above problem. For (almost) complete overlaps (e.g., the three-word bundle *yisi shi shuo* ‘the meaning is’ occurs 25 times in the conversation subcorpus, and the longer bundle *de yishi shi shuo* ‘what someone means is’ occurs 19 times), the shorter bundle is excluded from the list since its function is supposedly the same as that of the longer bundle. Hence, the distributional results will not be inflated in this regard. Yet it should be noted that the decision here does not mean accepting at this point any radical stance on the mental storage of lexical bundles. On the other hand, for other overlaps (e.g., the three-word bundle *you yi ge* ‘there is a’ occurs 343 times in the conversation subcorpus, and the longer bundle *hai you yi ge* ‘there is another one’ occurs only 38 times), both bundles remain in the list since each has its respective function. In the case mentioned here, *you yi ge* is usually used to introduce a topic, while the longer bundle *hai you yi ge* is used not only to introduced a topic but also to further elaborate by naming another item. As in most studies on lexical bundles, the overlapping number is not subtracted, i.e., both bundles are treated as separate ones, and their frequencies remain unadjusted. As a consequence, the distributional results may be somewhat inflated in this regard. Operationally speaking, for a pair of lexical bundles to be treated as complete overlaps, the cut-off point is set at the frequency threshold for each subcorpus. In the above case, *yisi shi shuo* occurs only 6 times not in the longer bundle *de yisi shi shuo*, and this is lower than the frequency threshold for the conversation subcorpus (i.e., 10 occurrences in the subcorpus, or 20 times per million words). Therefore, *yisi shi shuo* and *de yisi shi shuo* are treated as complete overlaps in the present study. With our method, the results may be inflated, but only to a limited extent.

After word sequences that do not have identifiable functions are manually excluded and overlapping bundles are properly dealt with, we get the finalized list of Chinese bundles. These lexical bundles are classified according to their structural characteristics. Although most lexical bundles are not structurally complete, there exist structural associations between them. For instance, both *I don't know how* and *I don't think he* are composed of a personal pronoun, a negated auxiliary, a verb, and a complement clause fragment. Biber et al. (1999) create a structural taxonomy of lexical bundles in English (see Table 2.2 in Section 2.2), which is a useful example for follow-up studies. Since Chinese and English are structurally distinct, the present study does not rely on any a priori structural framework in English.

Additionally, the lexical bundles in the finalized list are classified according to their communicative functions. Conrad and Biber (2004) propose an updated functional framework for lexical bundles in English (see Table 2.3 in Section 2.2), and it has been widely adopted and found to be useful. Therefore, the present study also adopts the same framework to explore the functions of lexical bundles in Chinese. We examine each lexical bundle in concordance lines, interpret its function in the co-text, and classify it under a functional category according to its most common use. However, a few modifications will be made to accommodate our Chinese data (see Section 5.2).

There are some problems with the functional analysis, mostly pertaining to the multifunctionality of lexical bundles (Conrad and Biber 2004, Salazar 2014). It is not always easy to determine exactly which functional category a lexical bundle belongs to. First, a lexical bundle can have different functions when used in different contexts. For example, Aijmer (2008) zooms in to examine the lexical bundle *I don't know*, suggesting that it can be used to avoid a straight answer, preface the speaker's

doi:10.6342/NTU201600328

disagreement, close a topic, and so on. Second, a lexical bundle can have multiple functions even in a single occurrence. For example, the lexical bundle *take a look at* can simultaneously serve as a directive and a topic introducer. Third, assigning a lexical bundle to a single category makes us overlook uses that may be less frequent yet not less pragmatically useful. Thus, instead of classifying lexical bundles according to their most common use, Salazar (2014) sometimes assigns a lexical bundle to multiple categories. However, the present study sticks to assigning a lexical bundle to a single category, mainly for two reasons. First, since the Sinica Corpus is very large, it is not feasible to exhaustively examine every single occurrence of a lexical bundle and identify all of its functions. A more sensible method for the present study is to sample a manageable size of occurrences of a lexical bundle and then determine the lexical sense and the most common function. Second, as have been found in many studies (e.g., Biber et al. 2003, Conrad and Biber 2004), most lexical bundles do have a primary function. Though we do not assign lexical bundles to multiple categories, we are aware of the limitations of our functional analysis.

3.5 Summary

In the present study, a computer code run in R automatically extracts three-word and four-word sequences from the conversation subcorpus and the news subcorpus of the Sinica Corpus. The Biberian approach (Biber et al. 1999), which is generally frequency-based, is adopted to identify lexical bundles, and another more sensitive dispersion measure and a word association measure are also used to complement this approach. Then, a number of high-frequency word sequences are manually excluded due to lacking an identifiable function or completely overlapping with a longer lexical bundle. The lexical bundles in the finalized list are classified according to their structure and function.

doi:10.6342/NTU201600328

We have entered a new era in which “the exploitation of modern computers will be at the center of progress” (Sinclair 2004: 23) in the study of language. Our method is resource-intensive, not only with a computer code running through a large-scale corpus but also with many mathematical algorithms working behind. The whole process involves such a great deal of computer processing that it needs to take place on a cloud server rather than on a personal computer. It is hoped that our method will shed some new light on the identification of lexical bundles.

As can be seen in the current chapter, “there is no purely automatic way of identifying phrasal units of meaning” (Stubbs 2007: 181). There are a number of human decisions involved (e.g., setting thresholds). These decisions are readily open to criticisms (e.g., arbitrary, subjective), as in most studies on lexical bundles. Even so, the lexical bundles identified in the present study are still the results of our effort to properly tackle methodological issues.

doi:10.6342/NTU201600328

Chapter 4

Exploring Quantitative Measures for Lexical Bundles

Following most studies on lexical bundles (see Table 2.4 in Section 2.2), the present study sets the frequency threshold at 20 times per million words. For each of the word sequences that pass the frequency threshold, the text count (i.e., the number of texts in which a given word sequence occurs), the DP value (i.e., another dispersion measure), and the G value (i.e., a word association measure) are calculated (see Section 3.3). In this chapter, the results will be presented, and how the decisions on other quantitative thresholds will be discussed.

4.1 Dispersion Measures for Lexical Bundles

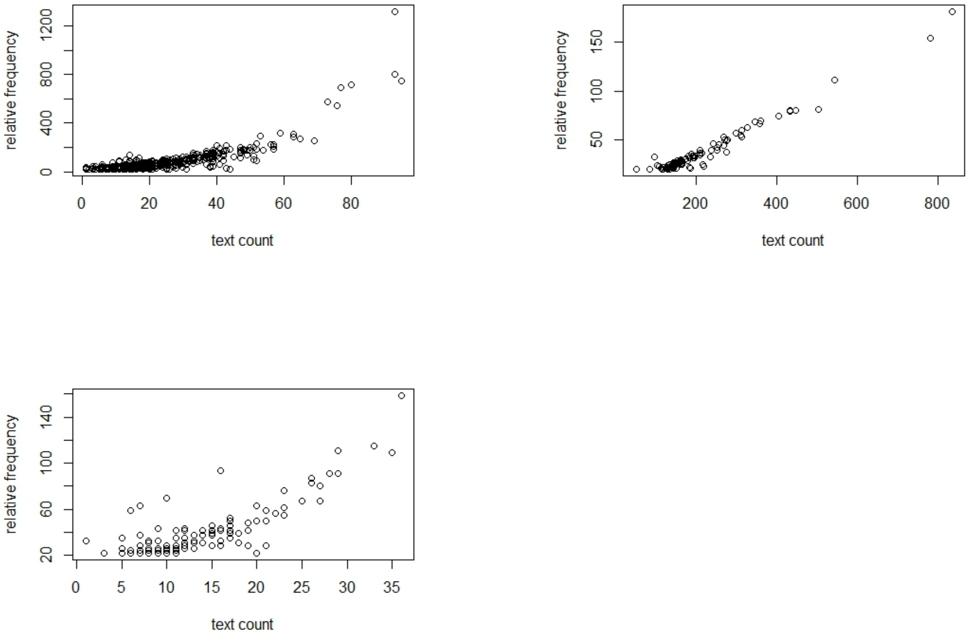
In Biber et al. (1999) and many follow-up studies (e.g., Cortes 2002, Biber et al. 2004, Kim 2009), a dispersion threshold is set at occurring in at least five different corpus texts to guard against individual idiosyncracies and local repetitions. However, the empirical data in the present study expose some problems of adopting text counts as the only dispersion measure in the identification of lexical bundles.

First, text counts and relative frequencies are highly correlated with each other, as illustrated in the following figures.[23]

[23] In the news subcorpus, only three four-word sequences pass the frequency threshold. They are not included in the following discussions.

doi:10.6342/NTU201600328

Figure 4.1. Correlations between text counts and relative frequencies.
(The upper left panel is for three-word spoken sequences, the upper right one is for three-word news sequences, and the lower one is for four-word spoken sequences.)

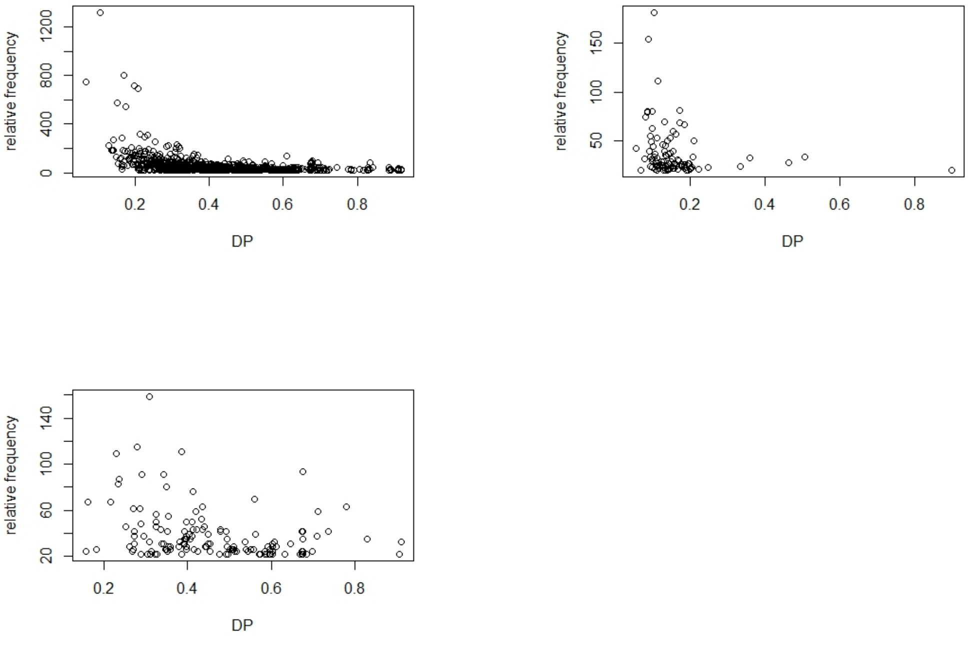


The correlation coefficients for the three sets of word sequences are 0.80, 0.98, and 0.82, respectively. As can be seen from the above figures, almost all the word sequences that pass the frequency threshold also pass the text count threshold. The text count threshold screens out only 26 (out of 1,024) three-word spoken sequences and 2 (out of 143) four-word spoken sequences, and no word sequences in the news subcorpus are excluded here. This suggests that the text count threshold may be of little practical use, and Conrad and Biber (2004) have noticed this.

Second, among sequences that pass the text count threshold, some are local repetitions that simply reflect the immediate topic of the discourse and are functionally/pragmatically uninteresting. Examples include *women de haizi* 'we POSSESSIVE.MARKER child; our children' and *junshi fayanren shi* 'military spokesman office'. Sequences like these do occur in several corpus texts, so they pass the text count threshold; however, they are absent in most of the corpus texts (see also Partington and Morley 2004). A more sensitive dispersion measure is needed to filter them out.

In view of the above problems with text counts, DP appears to be a more reliable dispersion measure. First, as shown in the following figures, the correlation between DP and relative frequencies is not so strong as that between text counts and relative frequencies.

Figure 4.2. Correlations between DP and relative frequencies.
(The upper left panel is for three-word spoken sequences, the upper right one is for three-word news sequences, and the lower one is for four-word spoken sequences.)



Since the correlation coefficients (i.e., -0.36, -0.19, and -0.28, respectively) are much lower, DP can be treated as independent of relative frequencies. Second, DP is more sensitive and can filter out word sequences that pass the text count threshold but have a skewed distribution in the news subcorpus. For instance, although *junshi fayanren shi* 'military spokesman office' occurs in 50 newswire texts, its DP value is rather high (i.e., 0.899). We can set a reasonable DP threshold to exclude such word sequences from further analysis.

To set a reasonable DP threshold, we manually check whether word sequences that pass the text count threshold are of functional/pragmatic value. Take three-word spoken sequences, for example. There are 998 three-word spoken sequences passing the text count threshold. Among these word sequences, five of them have a DP value falling between 0.80 and 0.89, and all the sequences here are either of little functional/pragmatic value (e.g., *ban wo chengzhang* ‘accompany me grow.up’ is simply a TV or radio program title) or have a very low text count (e.g., *zhen de a* ‘really’ occurs in just exactly five corpus texts). Then, 17 word sequences have a DP value falling between 0.70 and 0.79, and only two can be regarded as functionally/pragmatically significant (i.e., *man hao de* ‘very good’ and *zuo bu dao* ‘cannot do it’). Such a procedure is adopted to analyze the remaining word sequences, and the results are presented in the following table.

Table 4.1. Numbers of word sequences at each DP value.
(The first row presents the numbers of word sequences that pass the text count threshold. The numbers in the parentheses are how many word sequences there can be regarded as functionally/pragmatically significant.)

	Three-word spoken		Three-word news		Four-word spoken	
☑ text count threshold	998		101		141	
0.90-0.99	0		0		0	
0.80-0.89	5	(0)	1	(0)	1	(0)
0.70-0.79	17	(2)	0	(0)	4	(2)
0.65-0.69	41	(5)	0	(0)	13	(1)
0.60-0.64	43	(26)	0	(0)	8	(6)

As can be seen from the above table, almost all the word sequences with a DP value higher than 0.65 are functionally/pragmatically uninteresting. As the DP values lower, more word sequences worth our attention emerge. Therefore, it is decided that word sequences with a DP value higher than 0.65 will be excluded from further analysis. Though a few potential bundles are filtered out as a consequence, many word sequences that seem to be just local repetitions can be efficiently eliminated without manual interventions. The DP threshold suggested here, which echoes Gries' (2008b) observation that a lexical item with its DP value falling between 0.4 and 0.8 (e.g., *definition*: 0.795; *formal*: 0.708; *properly*: 0.625; *house*: 0.453) is certainly known to all native speakers and advanced learners, can also be tried in future studies on lexical bundles.

Although previous studies usually adopt text counts as the only dispersion measure, lexical bundles previously identified and related findings are still considered to be solid. As can be seen from the above table, the number of word sequences filtered out through the DP threshold is actually not high: i.e., approximately 6% of the three-word spoken sequences, only one three-word news sequence, and approximately 13% of the four-word spoken sequences. Besides, still a few word sequences successfully pass the DP threshold but fail the text count threshold. Therefore, the text count threshold is not abandoned in the present study, and the DP threshold is treated as complementary to it.

Though the DP threshold in the present study is still arguably arbitrary, our decision is based on a careful analysis of the data. However, even with the help of a more reliable dispersion threshold, some word sequences that do not serve important functions remain in the data. To screen them out, we may resort to quantitative measures that evaluate the internal association between the elements of a word sequence.

doi:10.6342/NTU201600328

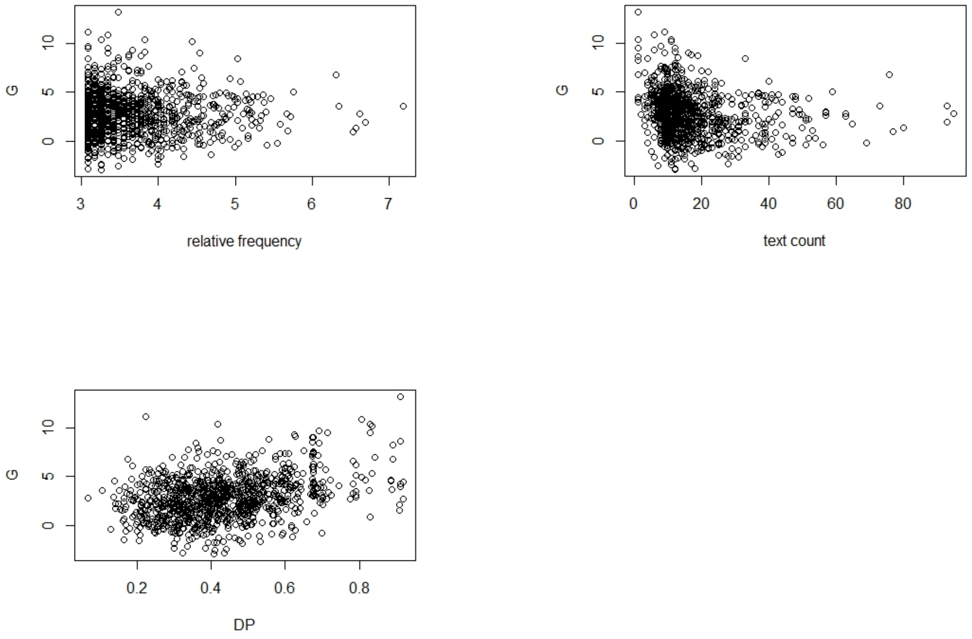
4.2 Association Measure for Lexical Bundles

As have been reviewed in Section 3.3, some studies have suggested that word association measures need to be considered when we identify multi-word combinations of pragmatic value. It has been found that word sequences achieving a stronger internal association are usually structurally complete and have identifiable functions (e.g., Wei and Li 2013, Salazar 2014). In the present study, G is adopted to measure the association between the elements of a multi-word combination (see Section 3.3).

The following figures present the correlations between the G values and the other three quantitative measures of three-word spoken sequences that occur more than 20 times per million words. As the figures show, the correlations between the G values and the other three quantitative measures are quite weak. The correlation coefficients (i.e., -0.02, -0.17, and 0.34, respectively) are low, and a linear regression model shows that it is difficult to predict the G value of a three-word spoken sequence from its relative frequency, text count, and DP altogether (R^2 = 0.18, p < 0.05). For three-word news sequences and four-word spoken sequences, similar patterns are also observed. Therefore, G is regarded as independent of the other quantitative measures.

doi:10.6342/NTU201600328

Figure 4.3. Correlations between G and the other quantitative measures of three-word spoken sequences that occur more than 20 times per million words. (The upper left panel is for the correlation between G and relative frequencies, the upper right one is for the correlation between G and text counts, and the lower one is for the correlation between G and DP.)



Just like the other quantitative thresholds, any word association threshold may be viewed as arbitrary, and different thresholds have been adopted in previous studies. For example, McEnery et al. (2006: 217) argue that word pairs with the MI score higher than three are "more useful for second language learners at beginning and intermediate levels", and Wei and Li (2013) set the G threshold at three when identifying phraseological sequences from academic corpora. For the present study, the G threshold at three is obviously too high: common, useful bundles such as the *yes-no* question structure *shi bu shi* would be excluded. Salazar (2014) adopts MI, and the threshold is set at 0.5. This is much lower, but it indeed appears to be arbitrary.

In fact, a MI score below zero means that these words tend not to co-occur, and this is also true of G. Therefore, the present study adopts a rather conservative threshold, i.e., zero, and excludes word sequences the elements of which do not co-occur more frequently than expected by chance alone.

We manually check whether word sequences that pass all the other three quantitative thresholds but fail the word association threshold are indeed of little functional/pragmatic value.[24] There are 92 three-word spoken sequences with the G value lower than zero, and 40 of them (i.e., 43.5%) do not have identifiable functions (e.g., *women yi ge* 'we one CLASSIFIER'). The association threshold efficiently weeds out a large number of word sequences that do not deserve our attention. Though it appears that many potential bundles are also excluded, there still remains 843 three-word spoken sequences. Given our limited resources, this is a more manageable size for further analysis. Besides, most potential bundles excluded here contain *zhe*/*na ge* 'this/that CLASSIFIER' (e.g., *zhe ge shi* 'this CLASSIFIER is'). They serve mainly as referential expressions and have few discourse-level functions. Through further manual interventions (see Section 3.4), another 200 word sequences (out of 843, i.e., 23.8%) are excluded from the data. It is clear that after the word association threshold is applied, most word sequences remaining in the data (i.e., 76.2%) can be regarded as useful lexical bundles. As for three-word news sequences, only two (e.g., *de yi ge* 'DE one CLASSIFIER') are filtered out by the G threshold; additionally, only four four-word spoken sequences (e.g., *de shi bu shi* 'DE A-*not*-A.QUESTION') are excluded.

A potential problem with word association measures is that word sequences containing low-frequency words can have a rather high value (e.g., Wei and Li 2013). However, such sequences may have been screened out by the frequency threshold at the very beginning of the procedure. As a consequence, just as in Salazar (2014), no negative effects of adopting a word association measure have been observed in the

[24] We do not check whether these sequences are structurally complete, because structural completeness is not a key feature of lexical bundles (Biber et al. 1999).

doi:10.6342/NTU201600328

present study.

4.3 Overall Results of Lexical Bundles in Chinese

The quantitative measures have provided a candidate list of lexical bundles in Chinese. These potential bundles fulfill all the quantitative criteria, including:

(i) reaching the frequency threshold of occurring at least twenty times per million words,

(ii) reaching the text count threshold of occurring in at least five corpus texts,

(iii) getting a DP value no higher than 0.65, and

(iv) reaching the G threshold at zero.

A manual analysis is still needed to exclude word sequences which are not readily interpretable in functional/pragmatic terms.[25] Sequences that remain are identified as lexical bundles. The following table summarizes the whole procedure.

Table 4.2. Numbers of word sequences passing each threshold.
(The icon ☑ stands for passing a threshold.)

	Three-word spoken	Three-word news	Four-word spoken	Four-word news
Types of sequences	165,970	3,044,598	156,078	2,793,826
☑ frequency threshold	1,024	101	143	3
☑ text count threshold	998	101	141	3
☑ DP threshold	935	100	123	3
☑ G threshold	843	98	118	3
☑ manual exclusion	**643**	**87**	**105**	**3**

In line with expectations, while there are a large number of sequence types, only a tiny proportion of them are frequently used. It is strikingly evident that very few four-word sequecnes in news pass the frequency threshold. It is also clear that conversations feature a much wider range of different lexical bundles than newswire

[25] See Section 3.4 for the exclusion criteria.

doi:10.6342/NTU201600328

texts. However, both in conversation and in news, the type number of three-word bundles is much larger than that of four-word bundles.

As for the proportion of corpus data covered by lexical bundles, conversation is also higher than news. The following table presents the percentages of words in lexical bundles.

Table 4.3. Percentages of words in lexical bundles.
(The percentages in the parentheses are calculated without removing punctuation marks.)

	Spoken		News	
Three-word	13.26%	(10.68%)	1.17%	(0.99%)
Four-word	2.01%	(1.62%)	0.03%	(0.03%)
Total	15.27%	(12.30%)	1.20%	(1.02%)

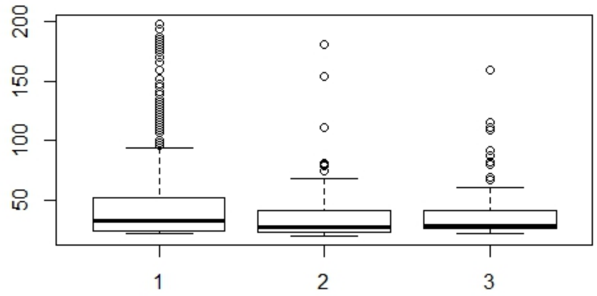
The same tendencies are also observed in English. In spontaneous speech (e.g., face-to-face conversations), speakers face real-time pressure. Therefore, a common strategy is to rely on frequent repetitions of prefabricated chunks such as lexical bundles (Biber et al. 2004, Johnstone 2002, Tannen 1982).

The following figure demonstrates the frequency distributions of lexical bundles.[26]

[26] As shown in Table 4.2, there are only three four-word bundles in news. It is inappropriate to draw a boxplot with only three data points. To make the shapes of the boxes clear, lexical bundles occurring more than 200 times per million words are not included. All of them are three-word bundles in conversation.

doi:10.6342/NTU201600328

Figure 4.4.	Frequency distributions of lexical bundles.
(The boxes from left to right are for three-word bundles in conversation, three-word bundles in news, and four-word bundles in conversation. The numbers on the vertical axis are frequencies per million words.)



Some lexical bundles occur with a very high frequency. As shown in the above figure, most of them are three-word bundles in conversation. The most common bundle in each set is as follows:

(i) three-word bundle in conversation: *shi bu shi* ‘A-*not*-A *yes-no* QUESTION’ (1,317 times per million words),

(ii) three-word bundle in news: *shi yi ge* ‘COPULA one CLASSIFIER’ (181 times),

(iii) four-word bundle in conversation: *mei yi ge ren* ‘every one CLASSIFIER person; everyone’ (159 times),

(iv) four-word bundle in news: *you hen da de* ‘have very large DE’ (24 times).

As some lexical bundles occur much more frequently than others, the Shapiro-Wilk normality test shows that the frequency distributions of the lexical bundles do not follow normal distributions. Thus, the Mann-Whitney test is performed on the means in the following table.

Table 4.4.	Means of relative frequencies (per million words) of lexical bundles.

Three-word spoken	Three-word news	Four-word spoken
55.4	37.9	38.6

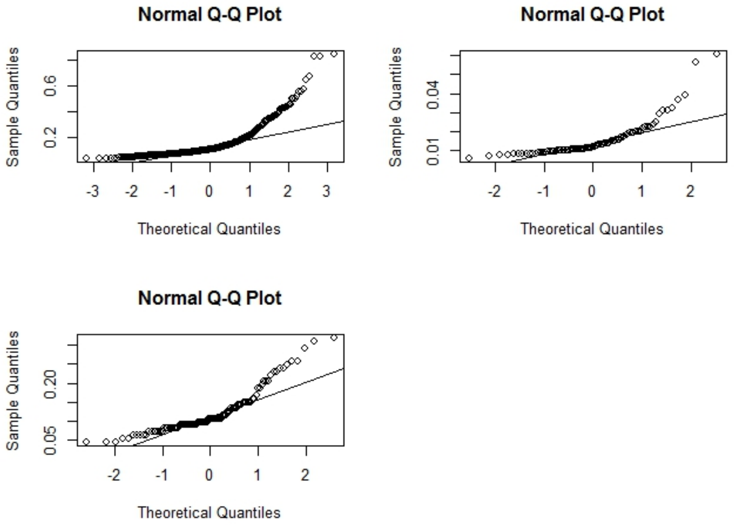
doi:10.6342/NTU201600328

Many three-word bundles in conversation occur with a very high frequency. As shown above, the three-word bundle with the highest frequency in conversation occurs approximately seven times more often than that in news. Therefore, it is not surprising that the relative frequency mean of three-word bundles in conversation is the highest. It is evident that three-word spoken bundles occur more frequently than three-word news bundles (*p* = 0.001), but the difference between three-word and four-word spoken bundles is not statistically significant (*p* = 0.06).

There are two dispersion measures (i.e., text counts and DP) in the present study. Before the dispersion of lexical bundles is discussed, the text counts of lexical bundles need to be normalized against the text numbers of the subcorpora (i.e., 113 conversation texts and 13,800 news texts). For example, *shi bu shi* occurs in 93 conversation texts, so its normalized text count is 0.823 (i.e., 93/113).

Just like frequencies, text counts also have skewed distributions. Some lexical bundles occur in a much larger number of texts than others, as the following figures show.

Figure 4.5. Quantile-quantile plots for text counts of lexical bundles.
(The upper left panel is for three-word spoken bundles, the upper right one is for three-word news bundles, and the lower one is for four-word spoken bundles.)



With skewed distributions, the Mann-Whitney test is performed on the text count means in the following table.

Table 4.5. Means of text counts (in percentages) of lexical bundles.

Three-word spoken	Three-word news	Four-word spoken
15.9%	1.54%	12.2%

The huge difference between three-word spoken and news bundles achieves statistical significance (p < 2.2e-16), and the difference between three-word and four-word spoken bundles is also statistically significant (p = 0.039). It appears that spoken bundles tend to occur in a larger proportion of texts than news bundles do.

However, DP values show an entirely different tendency. The following table presents the DP means of lexical bundles.

Table 4.6. Means of DP values of lexical bundles.

Three-word spoken	Three-word news	Four-word spoken
0.40	0.15	0.42

Only the DP values of four-word spoken bundles follow a normal distribution, so the Mann-Whitney test is still run on the DP means. The DP of three-word news bundles is lower than that of three-word spoken bundles (p < 2.2e-16), but the difference between three-word and four-word spoken bundles is not statistically significant (p = 0.263). Contrary to the finding based on text counts, the DP distributions show that three-word news bundles are more evenly dispersed than three-word spoken bundles.

The reason why text counts and DP values display opposite patterns may be that the former measure is easily susceptible to text lengths.[27] On average, each conversation text contains 4,069 tokens, which is almost ten times more than the

[27] There are 113 texts in the conversation corpus, which contains 459,833 tokens; there are 13,800 texts in the news corpus, which contains 6,475,872 tokens.

doi:10.6342/NTU201600328

average token number of a news text (i.e., 6,475,872/13,800 = 469.2). Now consider the following toy example, which is quite similar to the situation in the present study:

Figure 4.6(a). Distribution of a lexical bundle *a* in the subcorpus A.
(The thin bars stand for text boundaries. The thick bars stand for bundle occurrences.)

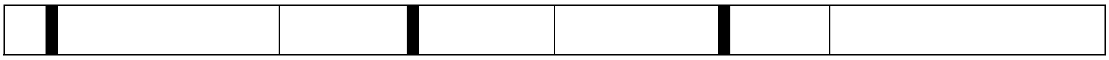
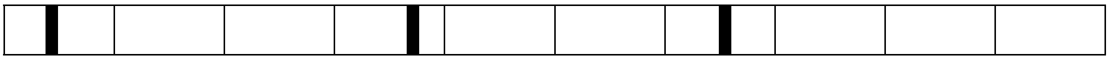


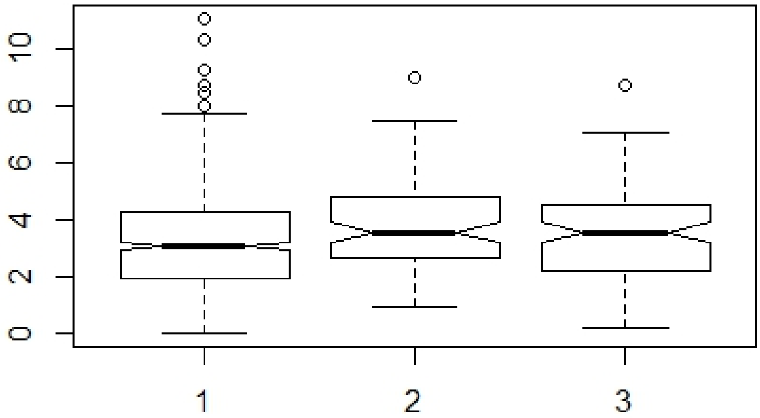
Figure 4.6(b). Distribution of a lexical bundle *b* in the subcorpus B.
(The thin bars stand for text boundaries. The thick bars stand for bundle occurrences.)



The texts in the subcorpus A is more than twice as long as those in the subcorpus B. There are four texts in the subcorpus A, and the bundle *a* occurs in 75% of the texts. There are ten texts in the subcorpus B, and the bundle *b* occurs in merely 30% of the texts. However, if we evenly divide both subcorpora and calculate the DP values for *a* and *b* (see Section 3.3.), then it is evident that the two bundles will be equally well-dispersed. In the present study, the text length difference is even more enormous. As a consequence, it comes as no surprise that the text count difference between three-word conversation and news bundles is dramatic (i.e., 15.9% vs. 1.54%). Based on DP values, three-word news bundles actually seem to be more evenly dispersed than three-word spoken bundles. The conflicting findings here also suggest that DP is needed to complement text counts in the identification of lexical bundles.

The following figure shows the distributions of the word association measures. The G means of three-word spoken bundles, three-word news bundles, and four-word spoken bundles are 3.19, 3.76, 3.50, respectively. The means are all above three, and this reminds us that word sequences with the MI score higher than three are of greater use for second language learners at beginning and intermediate levels (McEnery et al. 2006: 217). The role of Chinese bundles in second language learning needs to be further explored, but this is beyond the scope of the present study.

Figure 4.7.	G distributions of lexical bundles.
(The boxes from left to right are for three-word bundles in conversation, three-word bundles in news, and four-word bundles in conversation.)



The G values of three-word bundles in conversation do not follow a normal distribution, so the Mann-Whitney test is still applied to the G means. The difference between three-word spoken and news bundles achieves statistical significance (*p* = 0.002), and that between three-word and four-word spoken bundles is also statistically significant (*p* = 0.035). That is, the components in news bundles tend to be associated more closely than those in spoken bundles, and the components in longer bundles tend to be associated more closely than those in shorter bundles.

4.4 Summary

In the process of identifying lexical bundles, the present study adds two quantitative thresholds to the Biberian approach. First, DP reflects the dispersion of word sequences more accurately than text counts and weeds out some local repetitions that narrowly pass the text count threshold. Second, the word association measure G filters out many word sequences that contain frequently occurring function words but do not have identifiable functions. These two measures are fairly independent of relative frequencies and text counts and complement the Biberian approach. However, the quantitative measures cannot screen out all semantically/pragmatically vague word sequences, so manual interventions are still needed. In the long run, 838 lexical bundles in total (i.e., 643 three-word spoken bundles, 87 three-word news bundles, 105 four-word spoken bundles, and 3 four-word news bundles) are identified for further analysis.

Echoing previous findings in English (e.g., Biber et al. 1999), the present study shows that lexical bundles in different text types display different distributional patterns. Conversations feature a much wider range of different lexical bundles than newswire texts. As for the proportion of corpus data covered by lexical bundles, conversation is also higher than news. These reflect that in spontaneous speech, speakers are under real-time pressure and thus rely more heavily on prefabricated expressions such as lexical bundles. Regarding the dispersion of lexical bundles, the DP distributions show that news bundles are more evenly dispersed than spoken bundles.

It is also found that news bundles achieve stronger internal associations than spoken bundles. The G means of spoken and news bundles fall around three, which has been argued to be a critical value showing that these multi-word combinations are important in language acquisition (McEnery et al. 2006). The strong association between the elements of lexical bundles confirms that lexical bundles are not merely accidental combinations of high-frequency words (Conrad and Biber 2004). What knits these high-frequency items into lexical bundles is their essential communication functions in language use. The following two chapters will explore the functions of lexical bundles in Chinese and reveal more genre differences.

doi:10.6342/NTU201600328

Chapter 5

Lexical Bundles in Conversation

In this chapter, we will focus on the form and function of lexical bundles in conversation (Sections 5.1 and 5.2). We will also address some intriguing issues, such as the interaction between structural and functional categories, the interaction between quantitative measures and discourse functions, and similarities and differences between Chinese and English in their use of spoken bundles (Section 5.3).

5.1 Structural Classification of Lexical Bundles in Conversation

Just as in English, most lexical bundles in Chinese run across traditional phrase boundaries and appear to be structurally incomplete. Still, lexical bundles in Chinese have strong structural correlates, so they can be grouped into some structural types. In the present study, a functional taxonomy is created for lexical bundles in Chinese. There are three major types, including (i) clausal bundles, (ii) VP-based bundles, and (iii) NP-based bundles (see also Biber et al. 2004).

5.1.1 Clausal Bundles

A prototypical clausal bundle contains two core elements of a clause: i.e., the subject and the verb. Some lexical bundles that miss either one of the two elements but work on the sentential level are also assigned to this category. The following are seven subtypes.

The first subtype is bundles that contain a nominal element (e.g., a pronoun, a noun phrase, or a noun phrase fragment) and a common copula in Chinese, i.e., *shi* ‘be’. Examples include *diyi zhong shi* ‘the first type is’ and *wo de yisi shi* ‘I POSSESSIVE.MARKER meaning be; what I mean is’. Sometimes the copula is preceded

doi:10.6342/NTU201600328

by an adverb (e.g., *women dou shi* ‘we all are’, *ta ye shi* ‘he is also’) or followed by another one or two words (e.g., *zhe shi yi* ‘this is a’, *wo shi yi ge* ‘I am one CLASSIFIER’).

The second subtype is bundles containing a pronoun and a verb that can take a clause as its complement. The most common pronoun here is *wo* ‘I’, but a few noun phrases are also attested in that slot (e.g., *henduo ren shuo* ‘many people say’). As for the verb slot, there are three main types: (i) verbs of thinking, such as *juede* ‘feel’ (e.g., *wo juede zhe* ‘I feel this’), *xiang* ‘think’ (e.g., *wo xiang women* ‘I think we’), *zhidao* ‘know’ (e.g., *women dou zhidao* ‘we all know’); (ii) verbs of perception, such as *kankan* ‘take a look at’ (e.g., *women lai kankan* ‘we come look; let’s take a look at’) and *kandao* ‘see’ (e.g., *women keyi kandao* ‘we can see’); and (iii) verbs of saying, such as *shuo* ‘say’ (e.g., *ta jiu shuo* ‘he then said’), *jiang* ‘say’ (e.g., *ni ganggang jiang* ‘you just said’). As can be seen from the above examples, the subject of the complement clause can be included in a lexical bundle. Some bundles in this subcategory begin with a connector (e.g., *danshi wo juede* ‘but I feel’).

The third subtype is bundles containing a pronoun and a verb other than those mentioned above. Examples include *wo hen xihuan* ‘I very like; I like something very much’ and *wo yao qu* ‘I want go; I want to go’. Sometimes the object of a transitive verb occurs in a lexical bundle (e.g., *wo you yi ge* ‘I have one CLASSIFIER’).

The fourth subtype is bundles that contain a pronoun and a modal expression.[28] There are four modal expressions attested, i.e., *hui* ‘can; will’ (e.g., *wo jiu hui* ‘I then will’), *yao* ‘want to’ (e.g., *women jiu yao* ‘we then want to’), *keyi* ‘can’ (e.g., *ni jiu keyi* ‘you then can’), and *keneng* ‘may’ (e.g., *zhe ge keneng* ‘this CLASSIFIER may’). Some modal concatenations are also attested, such as *ni yiding yao* ‘you must’, *wo*

[28] When the modal expression is followed by a verb (e.g., *wo yao qu* ‘I want go; I want to go’), that bundle is assigned to the previous subcategory.

yiding hui 'I definitely will', and *women bixu yao* 'we must'.

The fifth subtype is bundles beginning with a connector that is followed by a pronoun or a demonstrative (i.e., *zhe* 'this', *na* 'that'). Bundles in this subcategory do not contain a verb, but they are still regarded as clausal bundles. Words in the connector slot usually introduce a clause, and the pronoun or the demonstrative functions as a clause subject. The most common connectors here are *suoyi* 'so' (e.g., *suoyi zhe ge* 'so this CLASSIFIER') and *ranhou* 'then' (e.g., *ranhou wo jiu* 'then I just'). Some adverbial transitional phrases are also attested in the connector slot (e.g., *shishi shang women* 'fact up we; in fact we', *jiben shang wo* 'basically I').

The sixth subtype is *yes*-*no* question bundles in the form of A-*not*-A. There are four different forms attested, i.e., *shi bu shi*, *dui bu dui*, *hao bu hao*, and *you mei you*. There are two reasons why they are regarded as clausal bundles. First, in the formation of an A-*not*-A question in Chinese, two clauses are involved (Li and Thompson 1981: 535).[29] Second, the above A-*not*-A forms can be attached to a clause and function as tag questions very often.[30] Among the four forms of A-*not*-A questions, *shi bu shi* occurs most frequently in four-word bundles (e.g., *shi bu shi keyi*, *women shi bu shi*).

The last subcategory is a miscellaneous one. For example, the three-word bundle *shihou wo jiu* 'when…, then I…' runs across the clause boundary, as shown below.

(5.1) Shengqi de shihou wo jiu hui quan shen de jirou dou hui hen yongli,
'When I feel angry, then the muscles in my whole body become tense,'

[29] For instance, if a Chinese speaker wants to ask whether Zhangsan is a student, then he or she takes the affirmative sentence *Zhangsan shi xuesheng* 'Zhangsan is a student' and its negative counterpart *Zhangsan bu shi xuesheng* 'Zhangsan is not a student', and then concatenates the underlined words, i.e., *Zhangsan shi bu shi xuesheng* 'Is Zhangsan a student'.

[30] Here is an example:
(a) Women de rensheng tai dandiao le, shi bu shi?
'Our life is too monotonous, isn't it?'

There is also an adverbial clause fragment, i.e., *wan le yihou* ‘finish ASPECT.MARKER after; after something finishes’.

(5.2) Mo di wan le yihou jiu gai ni xiwan,

‘After (I) have mopped the floor, then you should wash the dishes,’

Bundles like these do not occur very often, so no specific subcategories are created for them.

5.1.2 VP-based Bundles

A lexical bundle that features a verb but does not include its subject is considered to be a VP-based bundle.[31] There are also seven subcategories in the following.

The first subtype is bundles that contain the common copula *shi* ‘be’. The copula is often preceded by an adverb (e.g., *jiu shi zheyang* ‘just be so’), a connector (e.g., *ye shi hen* ‘also be very’), a modal expression (e.g., *keneng jiu shi* ‘may just be’), or a negator (e.g., *er bu shi* ‘but not be’).

The second subtype is bundles that feature a verb of thinking (e.g., *jiu hui juede* ‘then will feel’) or a verb of saying (e.g., *gen wo jiang* ‘with I tell; tell me’, *gaosu wo shuo* ‘tell I say; tell me’). These verbs usually take a clause as their complement. As the examples show, when a verb of saying occurs, the addressee is often specified in the bundle.

The third subtype is bundles that feature an adjectival verb. A large number of adjectives in Chinese may function as verbs (Li and Thompson 1981), and their verbal

[31] Chinese is a pro-drop language, in which the subject of a sentence may be omitted when it is inferable from the context. Therefore, some predicative bundles, such as *tai bang le* ‘too good SENTENCE-FINAL.MARKER’, may stand alone and thus be treated as a clausal bundle with a zero anaphor. With the same methodological considerations as in the functional analysis of lexical bundles (see Section 3.4), the present study assigns such bundles only to a single category, i.e., VP-based. The clausal category is generally for bundles containing both a subject and (part of) a predicate.

characteristics are manifested in lexical bundles. First, an adjectival verb can be followed by a sentence-final particle and be the nucleus of a VP-based bundle (e.g., *tai bang le* ‘too good SENTENCE-FINAL.MARKER’). Second, just as typical verbs can be, adjectival verbs can be negated by a negative particle (e.g., *bu tai hao* ‘not too good’, *hen bu rongyi* ‘very not easy’). The most common adjectival verb in spoken bundles is *hao* ‘good’.

The fourth subtype is bundles that contain a verb and its object or nominal complement. Most bundles in this category feature *you* ‘have’ (e.g., *you yi ge* ‘have one CLASSIFIER’, *you henduo de* ‘have many DE’, *you zheyang de* ‘have this.kind DE’). Other verbs attested in the data include *biancheng* ‘become’ (e.g., *biancheng yi ge* ‘become one CLASSIFIER’), *zuo* ‘do’ (e.g., *zuo shenme shiqing* ‘do what thing’), and *pengdao* ‘meet’ (e.g., *pengdao zhe zhong* ‘meet this kind’).

The fifth subtype is bundles that feature a modal expression.[32] The modal expression may be preceded by an adverb (e.g., *dou hui you* ‘all will have’), a negative particle (e.g., *bu gan qu* ‘not dare go’), or another modal expression (e.g., *yiding yao you* ‘must need have’).

The sixth subtype is bundles that feature the construction ‘verb + *de* + resultative complement’ (e.g., *shuo de hao* ‘speak DE well’). The negative variation ‘verb + *bu* + resultative complement’ (e.g., *zhao bu dao* ‘cannot find out’) is also attested. Sometimes the verb is not included in the bundle. For example, *de hen hao* ‘DE very good’ can combine with a wide range of verbs, but none of the combinations pass all the quantitative thresholds. Thus, only *de hen hao* is identified as a lexical bundle. Likewise, sometimes the complement is just a fragment (e.g., *jiang de hen* ‘speak DE very’, *bian de hen* ‘become DE very’).

[32] When the modal expression is followed by a noun phrase or a noun phrase fragment (e.g., *hui you yi* ‘will have one’), that bundle is assigned to the previous subcategory.

Lastly, there is also a miscellaneous subcategory for a few VP-based bundles that bear specific constructions in Chinese. For example, the construction V-*yi*-V has a multiplicative reading, and *tan yi tan* 'talk and talk' and *xiang yi xiang* 'think and think' frequently occur in conversation. Some serial verb constructions occur frequently, such as causative constructions (e.g., *rang* *ta* *qu* 'let him go') and verb phrases as the direct object of another (e.g., *bu xihuan chi* 'not like eat').

5.1.3 NP-based Bundles

As the term suggests, a prototypical NP-based bundle is either a complete noun phrase or a noun phrase fragment. Additionally, phrases that modify nouns and prepositional phrases where noun phrases share a heavier semantic load are also assigned to this category. The following are six subtypes.

The first subtype is full noun phrases. Common examples include: (i) 'number + classifier/quantifier + noun' (e.g., *yi ge ren* 'one CLASSIFIER person'); (ii) 'demonstrative + classifier/quantifier + noun' (e.g., *zhe ge wenti* 'this CLASSIFIER question'); (iii) 'pronoun + possessive *de* + noun' (e.g., *women de shehui* 'we DE society'); (iv) juxtaposition of two nouns (e.g., *women liang ge* 'we two CLASSIFIER'); (v) locative phrases (e.g., *wo xin li* 'I heart in; in my mind', *zhe ge shijie shang* 'this CLASSIFIER world on; in this world').[33] As can be seen from the above examples, the neutral classifier *ge* occurs quite frequently in full noun phrases.

The second subtype is noun phrases or noun phrase fragments followed by an adverb. Common adverbs here include *dou* 'all' (e.g., *henduo ren dou* 'many people all'), *jiu* 'then' (e.g., *lingwai yi ge jiu* 'another one CLASSFIER just'), *bu* 'not' (e.g., *wo ye bu* 'I also not'), and *hen* 'very' (e.g., *wo ye hen* 'I also very').

[33] Locative markers function as nouns in classical Chinese and still retain a number of nominal properties in modern Chinese (Huang et al. 2009). Therefore, the present study treats locative phrases as NP-based bundles.

doi:10.6342/NTU201600328

The third subtype is bundles that noun phrases or noun phrase fragments preceded by the modifier marker *de*. Noun phrases can be modified by an adjective (e.g., *kuaile de shiqing* ‘happy DE thing’), a demonstrative (e.g., *zheyang de shiqing* ‘this.kind DE thing’), a noun (e.g., *guozhong de shihou* ‘junior.high.school DE time’), or a verb (e.g., *chi de dongxi* ‘eat DE thing’). Sometimes only the modifier marker *de* is included in lexical bundles (e.g., *de guocheng dangzhong* ‘DE process in; in the process of’, *de yi zhong* ‘DE one kind; one kind of’).

The fourth subtype is bundles that feature modifiers but do not contain noun phrases. Examples include *zui zhongyao de* ‘most important DE’, *zhe fangmian de* ‘this aspect DE’, and *ni shuo de* ‘you say DE; what you say’. Sometimes the modifier marker *de* is not included in lexical bundles (e.g., *yi ge hen zhongyao* ‘one CLASSIFIER very important’).

The fifth subtype is prepositional phrases. The most common preposition in spoken bundles is the locative marker *zai*, which can be used in concrete terms (e.g., *zai jia li* ‘at home inside’) or in abstract terms (e.g., *zai zhe fangmian* ‘in this regard’).

Lastly, there is also a miscellaneous subcategory for a few NP-based bundles. Most bundles in this subcategory are juxtapositions of a pronoun and a nominal time expression (e.g., *na shihou wo* ‘that time I’, *wo mei ci* ‘I every time’).

5.1.4 Others

Still some lexical bundles do not fit comfortably in with the above three main categories. A miscellaneous category is created for them, but their type number is rather low (cf. Section 5.1.5). Some routine formulae in spoken Chinese allow little variation and do not follow traditional phrasal rules. Examples include *tong yi shijian zaijian* ‘the.same one time see.you’, which is used at the end of a TV program, and *dui dui dui* ‘right right right’, which is used as a response in conversation. Some

doi:10.6342/NTU201600328

adverbial phrases (e.g., *dui wo lai jiang* ‘to I come say; to me’, *ye bu tai* ‘also not too’) are also assigned here. Bundles in this category often contain elements that have undergone a grammaticalization process (cf. Traugott and Dasher 2005), so sometimes it is difficult to determine the grammatical status of bundles here.[34]

5.1.5 Structural Distributions of Lexical Bundles

The following tables present the structural distribution of Chinese bundles in conversation. As Salazar (2014) suggests, a structural category may be represented by a wide range of bundle types, and each of them may occur just sporadically; it is also likely that only a few bundle types are assigned to a certain structural category, but each one occurs with a very high frequency. Therefore, both the type distribution and the token distribution are presented.

[34] For example, it appears that *jiang* ‘say’ in *dui wo lai jiang* ‘to me’ does not function as a lexical verb.

Table 5.1(a). Structural distribution of lexical bundle types in conversation.

Structural category	Three-word spoken		Four-word spoken	
1. Clausal bundles	**173**	**(26.9%)**	**38**	**(36.2%)**
(1) noun phrase (fragment) + copula *shi* (+ …)	36	(6%)	14	(13%)
(2) (connector +) pronoun + verb (+ clause fragment)	72	(11%)	9	(9%)
(3) pronoun + verb (+ …)	15	(2%)	3	(3%)
(4) pronoun + modal expression	14	(2%)	0	(0%)
(5) connector + pronoun/demonstrative	28	(4%)	0	(0%)
(6) A-*not*-A question	5	(1%)	11	(10%)
(7) others	3	(0%)	1	(1%)
2. VP-based bundles	**211**	**(32.8%)**	**42**	**(40.0%)**
(1) copula *shi* (+ …)	79	(12%)	23	(22%)
(2) verb (+ clause fragment)	32	(5%)	3	(3%)
(3) adjectival verb	20	(3%)	0	(0%)
(4) verb + noun phrase (fragment)	49	(8%)	15	(14%)
(5) modal expression (+ verb)	10	(2%)	0	(0%)
(6) (verb +) *de*/*bu* + complement (fragment)	10	(2%)	0	(0%)
(7) others	11	(2%)	1	(1%)
3. NP-based bundles	**242**	**(37.6%)**	**23**	**(21.9%)**
(1) noun phrase	107	(17%)	4	(4%)
(2) noun phrase + adverb	25	(4%)	2	(2%)
(3) (modifier +) *de* + noun phrase (fragment)	34	(5%)	9	(9%)
(4) modifier (+ *de*)	45	(7%)	5	(5%)
(5) preposition + noun phrase (fragment)	20	(3%)	2	(2%)
(6) others	11	(2%)	1	(1%)
4. Others	**17**	**(2.6%)**	**2**	**(1.9%)**
TOTAL	**643**	**(100%)**	**105**	**(100%)**

doi:10.6342/NTU201600328

Table 5.1(b). Structural distribution of lexical bundle tokens in conversation.

Structural category	Three-word spoken		Four-word spoken	
1. Clausal bundles	**4752**	**(29.0%)**	**675**	**(36.2%)**
(1) noun phrase (fragment) + copula *shi* (+ …)	830	(5%)	236	(13%)
(2) (connector +) pronoun + verb (+ clause fragment)	1918	(12%)	210	(11%)
(3) pronoun + verb (+ …)	234	(1%)	39	(2%)
(4) pronoun + modal expression	231	(1%)	0	(0%)
(5) connector + pronoun/demonstrative	521	(3%)	0	(0%)
(6) A-*not*-A question	972	(6%)	180	(10%)
(7) others	46	(0%)	10	(1%)
2. VP-based bundles	**5008**	**(30.6%)**	**703**	**(37.7%)**
(1) copula *shi* (+ …)	2173	(13%)	411	(22%)
(2) verb (+ clause fragment)	546	(3%)	46	(2%)
(3) adjectival verb	354	(2%)	0	(0%)
(4) verb + noun phrase (fragment)	1466	(9%)	228	(12%)
(5) modal expression (+ verb)	145	(1%)	0	(0%)
(6) (verb +) *de*/*bu* + complement (fragment)	182	(1%)	0	(0%)
(7) others	142	(1%)	18	(1%)
3. NP-based bundles	**6319**	**(38.6%)**	**461**	**(24.7%)**
(1) noun phrase	2988	(18%)	107	(6%)
(2) noun phrase + adverb	511	(3%)	52	(3%)
(3) (modifier +) *de* + noun phrase (fragment)	762	(5%)	157	(8%)
(4) modifier (+ *de*)	1289	(8%)	111	(6%)
(5) preposition + noun phrase (fragment)	580	(4%)	24	(1%)
(6) others	189	(1%)	10	(1%)
4. Others	**288**	**(1.8%)**	**27**	**(1.4%)**
TOTAL	**16367**	**(100%)**	**1866**	**(100%)**

As can be seen from the above tables, the type distribution and the token distribution display similar tendencies. For three-word bundles in conversation, the most common structural type is NP-based bundles, followed by VP-based bundles and clausal bundles. For four-word bundles in conversation, the most common structural type is VP-based bundles, followed by clausal bundles and NP-bundles. Few bundles fall into the miscellaneous category, and this confirms that lexical bundles have strong structural correlates.

Compared with three-word spoken bundles, a larger proportion of four-word spoken bundles fall into the clausal category. This is not surprising, since four-word bundles have more word slots and are longer. Besides, three-word and four-word spoken bundles prefer different clausal types. The sequence ‘pronoun + clause-taking verb’ is often the core of a clausal bundle. When such a sequence is preceded by a connector or followed by a word in the clause that follows, a three-word bundle forms, i.e., ‘connector + pronoun + verb’ or ‘pronoun + verb + clause fragment’. On the other hand, four-word spoken bundles often feature the copula *shi*, and the clausal type ‘noun phrase (fragment) + copula *shi* (+ …)’ is quite common.

Regarding the NP-based category, the length of bundles appears to influence their structural distribution. Many three-word spoken bundles are noun phrases, mostly in the form of ‘number/demonstrative + classifier + noun’. Still, perhaps with more word slots, four-word spoken bundles often include the modifier marker *de* or even the full modifier (e.g., *jia li de ren* ‘home inside DE person; people at home’). The most common NP-based type of four-word spoken bundles is ‘(modifier +) *de* + noun phrase (fragment)’.

As for the VP-based category, three-word and four-word spoken bundles show similar preferences. The most common types are ‘copula *shi* (+ …)’ and ‘verb + noun phrase (fragment)’. This suggests that the copula *shi* plays an essential role in

doi:10.6342/NTU201600328

VP-based bundles, and that the word order VO is a dominant argument structure in Chinese bundles.

5.2 Functional Classification of Lexical Bundles in Conversation

The functional taxonomy in Conrad and Biber (2004) is adopted to classify lexical bundles in Chinese. Several modifications are made to accommodate the data in Chinese. Now there are three major functional categories: (i) interpersonal bundles, (ii) discourse organizers, and (iii) referential expressions. Each of them will be introduced in detail.

5.2.1 Interpersonal Bundles

It has been widely recognized and often mentioned in semantics textbooks that there are three levels or facets of meaning: (i) referential/propositional/ideational, (ii) textual, and (iii) interpersonal/interactional/expressive. Hyland (2008) proposes a functional framework for lexical bundles in academic writing, and the three main categories can be seen as corresponding to the above three kinds of meaning respectively: (i) research-oriented bundles (referential), (ii) text-oriented bundles (textual), and (iii) participant-oriented bundles (interpersonal). In Conrad and Biber's (2004) functional framework of lexical bundles, however, there are four categories, i.e., (i) stance expressions, (ii) discourse organizers, (iii) referential expressions, and (iv) special conversational functions. Obviously, the first three categories also correspond to the three kinds of meaning. In essence, bundles in the fourth category (e.g., *thank you very much*, *what are you doing*) are also interpersonal, just as stance expressions are. Therefore, the first category and the fourth category are merged in the present study, and the label for this new category is interpersonal bundles. When people interact with others through language, interpersonal bundles are used to

doi:10.6342/NTU201600328

express viewpoints and attitudes as well as maintain interactions and relations with other people.

The first subcategory of interpersonal bundles is epistemic stance bundles that “comment on the knowledge status of the information in the following proposition: certain, uncertain, or probable/possible” (Biber et al. 2004: 389). Epistemic bundles can be either personal or impersonal: personal epistemic bundles explicitly attribute stances to someone (e.g., *wo juede wo* ‘I feel I’), whereas impersonal stance bundles do not (e.g., *shishi shang shi* ‘the fact is that’). The following are two examples:

(5.3) Wo bu zhidao shuo zhe shi yi ge weixie de fangshi,

‘I don’t know that this is a way to threaten,’

(5.4) Dagai jiu shi ni pingchang taishao shuo, suoyi ta bu tai xiguan.

‘Perhaps you just seldom say that, so he isn’t used to it.’

In (5.3), the speaker uses the personal epistemic bundle *wo bu zhidao* ‘I don’t know’ to overtly convey his or her own uncertainty stance. In (5.4), the bundle *dagai jiu shi* ‘perhaps it is just’ also expresses a lack of certainty, but the stance is not directly attributed to the speaker. There are more personal epistemic bundles than impersonal ones. Besides, it is found that uncertainty is expressed much more often than certainty.

The second subcategory of interpersonal bundles is attitudinal/modality bundles that “express speaker attitudes towards the actions or events described in the following proposition” (Biber et al. 2004: 389). Consider the following examples:

(5.5) Wo yao qu xuehao yingwen,

“I want to learn English well,”

doi:10.6342/NTU201600328

(5.6) Wo bu xuyao shuo you tai gao de yi ge qixu, huoxu wo jiu keyi manman jianliqi ziji de zixinxin.

'I don't need to have a too high expectation, and perhaps I can just gradually build up my confidence.'

(5.7) Wo bixu yao huida.

'I must answer.'

In (5.5), the bundle *wo yao qu* 'I want to' expresses the speaker's self-motivated wish and desire. In (5.6), the bundle *wo jiu keyi* 'then I can' expresses the speaker's ability. In (5.7), the bundle *wo bixu yao* 'I must' expresses the speaker's obligation. These are all personal attitudinal/modality bundles. There are also impersonal attitudinal/modality bundles (e.g., *yao you yi* 'there must be one'), but just a few.

The third subcategory of interpersonal bundles is special interactional bundles. The original label of this category in Conrad and Biber (2004) is special conversational functions because bundles here are found to occur only in the conversation subcorpus. However, in Chinese, some bundles assigned to this category are also found in news (e.g., *shi bu shi* 'A-*not*-A QUESTION'). As a result, the original label for this category is not adopted in the present study.

Special interactional bundles are mostly politeness routines (e.g., *xiexie ge wei* 'thank you everyone'), simple inquiries (e.g., *shenme yang de* 'what kind of'), and reporting clauses (e.g., *wo jiu shuo* 'I just said that', *gaosu wo shuo* 'told me that') (Conrad and Biber 2004: 67). In the present study, two brief responses (i.e., *dui dui dui* 'right right right' and *bu shi la* 'not be SENTENCE-FINAL PARTICLE; no') are also assigned to this subcategory. These bundles serve important interactional functions in spoken language. Politeness routines usually occur in specific contexts, and speakers

doi:10.6342/NTU201600328

use them to maintain relations with other people. For example, *tong yi shijian zaijian* ‘see you again at the same time’ is used by TV program hosts at the end of an episode to invite the audience to watch that program again. Simple inquiries elicit responses and keep the conversation going. It is also clear that reporting clauses and brief responses can help to maintain interactions with others.

5.2.2 Discourse Organizers

Discourse organizers “reflect relationships between prior and coming discourse” (Conrad and Biber 2004: 67). There are three subcategories of discourse organizing bundles: (i) topic introduction bundles, (ii) topic elaboration bundles, and (iii) identification bundles.

Topic introduction bundles overtly signal that a new topic or subtopic is being introduced into discourse. In Chinese, a common type of topic introduction bundles is presentational phrases containing *you* ‘have’ and *yi* ‘one’, such as *you yi ge* ‘there is a’, *you yi ci* ‘one time’, and *you yi dian* ‘there is a point’. Here is an example.

(5.8) A: Na yi wei yao fayan shuoshuo ziji de kanfa?
‘Which one wants to talk about personal opinions?’
B: Wo you yi ci zheyang de jingyan, yinwei wo hen nande you jihui pashan,
‘I have such an experience, because I don’t have many opportunities to go mountain climbing,’

In (5.8), the speaker uses *you yi ci* ‘one time’ at the very beginning of his narration, and an personal experience of going mountain climbing with some friends becomes focal and is further elaborated in the following discourse. An indefinite referent in the form of '*yi* + noun' often provides an overt signal that a newly introduced referent is to become prominent in the immediate discourse (see Dooley and Levinsohn 2001). Bundles with the first personal plural pronoun *women* 'we' in the subject position are also common topic introduction bundles. Here is an example:

(5.9) Women lai kankan tongji de jieguo,
'Let's take a look at the statistical results,'

In (5.9), the bundle *women lai kankan* invites the addressees to pay attention to the following new topic.

Topic elaboration bundles serve to provide more information about a topic. A topic elaboration bundle often contains a connector that clearly signals the semantic relation between propositions, such as cause-and-effect connectors *yinwei* 'because' (e.g., *yinwei na shihou* 'because at that time') and *suoyi* 'so' (e.g., *suoyi zhe shi* 'so this is'). Also, a topic elaboration bundle often contains a monosyllabic adverb that can combine clauses (Biq 2015), such as *ye* 'too' (e.g., *ye shi yiyang* 'also be the same') and *jiu* 'then' (e.g., *wo jiu bu* 'I then don't'). Sometimes these adverbs co-occur with a connector (e.g., *suoyi wo jiu* 'so I just'). Some topic elaboration bundles do not contain overt linking markers like those mentioned above, but they feature preposed expressions that serve as "cohesive ties linking the following predication to something in the preceding context" (Dooley and Levinsohn 2001: 38). Examples include *ju ge lizi* 'give an example; for example' and *jiben shang wo* 'basically I'.

doi:10.6342/NTU201600328

Identification bundles “identify an entity or part of it as noteworthy” (Conrad and Biber 2004: 67).[35] They have discourse organizing functions because they draw more attention and make what is being discussed more prominent in the context. The noteworthiness of an entity is often overtly signaled by a high-frequency degree adverb, like *zui* ‘most’, *hen* ‘very’, and *feichang* ‘extremely’ (e.g., *zui*/*hen*/*feichang zhongyao de* ‘most/very important DE’). Adverbs such as *bijiao* ‘more’ (e.g., *bijiao hao de* ‘more good DE; better’), *xiangdang* ‘very’ (e.g., *shi xiangdang de* ‘be very DE’), and *zhenzheng* ‘really’ (e.g., *shi zhenzheng de* ‘be really DE’) are also attested. Another important discourse organizing function of identification bundles is to summarize and emphasize the main point after a lengthy discussion, as shown in the following example.

(5.10) Suoyi wo geren juede, ruguo ni zaoyudao cuozhe de shihou, ni nenggou xunqiu ziji lingwai yi fangmian de chengjiugan, lai mibu ziji zhe zhong shiluogan, yexu shi yi zhong hen hao de, you jianshexing de yi zhong shiying cuozhe de fangshi.

‘So I personally believe that if you encounter frustrations and you can seek a sense of achievement in another area to compensate for your own sense of loss, this may be a very good, constructive way to go through frustrations.’

Bundles serving this function include *shi yi ge*/*jian*/*ge hen* ‘be one CLASSIFIER very’, *zhe shi hen* ‘this is very’, and *zhexie dou shi* ‘these all are’.

[35] This subcategory is originally placed under the main category “referential expressions” (Conrad and Biber 2004); it is moved under the main category “discourse organizers” in Biber and Barbieri (2007).

5.2.3 Referential Expressions

Referential expressions "make direct reference to physical or abstract entities or to the textual context" (Conrad and Biber 2004: 67). This category also includes lexical bundles that identify some particular attribute of an entity or refer to events or processes. There are five subcategories, as enumerated below.

The first subcategory is imprecision bundles. An imprecision bundle is used when the speaker is not to make a specific reference, as shown in the following example.

(5.11) Mo yi ge tongxue hen xinshang ni,

'A certain classmate admires you very much,'

In addition, an imprecision bundle is also used to indicate that more references of the same type can be provided. The bundle *zhe yi lei de* 'this kind of' is an example.

(5.12) Wo qinshi dou he shuiguojiu. Xianzai yijing kanbuqi pijiu zhe yi lei de dongxi le.

'My roommates all drink fruit wine. Now we have come to look down upon things like beer.'

In (5.12), the speaker may think of many cheap wine items, but not all of them are listed.

The second subcategory is phoric bundles, which feature two common demonstratives in Chinese, i.e., *zhe* 'this' and *na* 'that'. This subcategory is not in Conrad and Biber (2004). However, the word sequence '*zhe*/*na* + noun/classifier' occurs frequently in Chinese, so a new subcategory is created for lexical bundles

doi:10.6342/NTU201600328

bearing this combination (e.g., *na zhong ganjue* ‘that kind of feeling’, *zhe ge shehui shang* ‘in this society’). Bundles bearing *zhe* and *na* have various phoric uses (cf. Thompson 1996). Consider the following examples.

(5.13) Women zhe ge jiemu, shi xiangdang kexue de,
‘We this program is really scientific,’

(5.14) Zhe ge shijie shang nameduo zhanluan,
‘In this world there are so many wars,’

(5.15) Yizhi xie bu chulai hao xiang ku. Na ge shihou zhen xiwang shijian zhanting,
‘All the while I couldn’t write down anything, and I wanted to cry. At that time I really wished that time could stop for a short while.’

(5.16) Wo tongshi yao ba zhe ju hua gaosu wo de xiaohai, buguan ni zai deyi de shihou huoshi you cuozhe de shihou, baba, mama yongyuan dou shi ni de peiban zhe gen zhichi zhe.
‘Meanwhile, I want to tell my children these words: whether you are happy or encounter obstacles, Dad and Mom always keep you company and are your supporters.’

In (5.13), the use of *women zhe ge* ‘we this’ is exophoric, i.e., pointing outwards to something in the environment where the conversation takes place. In (5.14), with their knowledge and experience, both the speaker and the listener understand that *zhe ge shijie zhang* ‘in this world’ refers to the world where they live. This kind of reference

doi:10.6342/NTU201600328

is homophoric. In (5.15), *na ge shihou* 'at that time' refers to the situation just mentioned in the context, and this use of reference is anaphoric, i.e., pointing backwards. In contrast, the meaning of the reference item *zhe ju hua* 'these words' in (5.16) is in the following text. This kind of reference is cataphoric, i.e., pointing forwards. Because both anaphoric and cataphoric uses are endophoric (i.e., pointing outwards to the text), the bundles in (5.15) and (5.16) also contribute to the coherence of the discourse.

The third subcategory is entity bundles, which refer to entities of any kind. In Conrad and Biber (2004), there is a subcategory for time/place/text references. That category is made to accommodate more entity types in the present study, and a new label is thus given. Entity bundles can refer to particular places (e.g., *women ban shang* 'in our class', *wo xin li* 'in my mind'), times (e.g., *de shihou ne* 'when…', *chifan de shihou* 'at mealtimes'), people (e.g., *women liang ge* 'we two', *yi ge nuhaizi* 'a girl'), and a wide range of concrete and abstract things (e.g., *yi bi qian* 'a sum of money', *kuaile de shiqing* 'happy things'). As in English (e.g., *at the end of*), some bundles here are multifunctional, with their interpretation depending on the context (e.g., *zai wo de fangjian* 'in my room', *zai wo de yisheng* 'in my whole life', *zai wo de shen shang* 'in my body', *zai wo de jiyi* 'in my memories').

The fourth subcategory is attribute-specifying bundles. Some bundles in this subcategory specifies quantities through definite numeral expressions (e.g., *zhi you yi ge* 'only have one CLASSIFIER'), indefinite numeral expressions (e.g., *you henduo de* 'have many DE'), and distributive numeral expressions (e.g., *mei ge ren* 'every CLASSIFIER person'). Other bundles in this subcategory describe types (e.g., *de yi zhong* 'a kind of'), manners (e.g., *de fangshi lai* 'by means of'), purposes (e.g., *shi wei le* 'for the purpose of'), and so forth. Still other bundles describe quality (e.g., *hen hao de* 'very good DE', *tai bang le* 'too excellent SENTENCE-FINAL.PARTICLE') and similarity (e.g., *wanquan bu yiyang* ‘completely not the.same; completely different’).

The fifth subcategory, which is not in Conrad and Biber (2004), is process bundles. Some lexical bundles in Chinese make references to events or processes, and a new subcategory is created for them. Thompson (1996) presents four main process types, and each of them is attested to occur in lexical bundles: (i) material processes, i.e., physical actions (e.g., *zhao bu dao* ‘cannot find out’); (ii) mental processes, i.e., processes in the mind (e.g., *xiang yi xiang* ‘think and think’, *wo hen xihuan* ‘I like very much’); (iii) relational processes, i.e., processes that indicate attributes or identify entities, usually realized through the copula *shi* (e.g., *shi yi ge* ‘is a’); (iv) verbal processes, i.e., actions that involve message exchanges through language (e.g., *suo jiang*/*shuo de* ‘what is said’).

5.2.4 Functional Distributions of Lexical Bundles

The following tables show the functional distribution of Chinese bundles in conversation. Still, both the type distribution and the token distribution are presented.

Table 5.2(a). Functional distribution of lexical bundle types in conversation.

Functional category	Three-word spoken		Four-word spoken	
1. Interpersonal bundles	**166**	**(25.8%)**	**27**	**(25.7%)**
(1) Epistemic stance bundles	86	(13%)	10	(10%)
(2) Attitudinal/Modality stance bundles	22	(3%)	0	(0%)
(3) Special interactional bundles	58	(9%)	17	(16%)
2. Discourse organizers	**161**	**(25.0%)**	**35**	**(33.3%)**
(1) Topic introduction bundles	18	(3%)	11	(10%)
(2) Topic elaboration bundles	105	(16%)	11	(10%)
(3) Identification bundles	38	(6%)	13	(12%)
3. Referential expressions	**316**	**(49.1%)**	**43**	**(41.0%)**
(1) Imprecision bundles	1	(0%)	1	(1%)
(2) Phoric bundles	92	(14%)	17	(16%)
(3) Entity bundles	96	(15%)	4	(4%)
(4) Attribute-specifying bundles	88	(14%)	17	(16%)
(5) Process bundles	39	(6%)	4	(4%)
TOTAL	**643**	**(100%)**	**105**	**(100%)**

Table 5.2(b). Functional distribution of lexical bundle tokens in conversation.

Functional category	Three-word spoken		Four-word spoken	
1. Interpersonal bundles	**4603**	**(28.1%)**	**496**	**(26.6%)**
(1) Epistemic stance bundles	2222	(14%)	227	(12%)
(2) Attitudinal/Modality stance bundles	376	(2%)	0	(0%)
(3) Special interactional bundles	2005	(12%)	269	(14%)
2. Discourse organizers	**3806**	**(23.3%)**	**655**	**(35.1%)**
(1) Topic introduction bundles	686	(4%)	158	(8%)
(2) Topic elaboration bundles	2061	(13%)	204	(11%)
(3) Identification bundles	1059	(6%)	293	(16%)
3. Referential expressions	**7958**	**(48.6%)**	**715**	**(38.3%)**
(1) Imprecision bundles	23	(0%)	16	(1%)
(2) Phoric bundles	2586	(16%)	245	(13%)
(3) Entity bundles	2165	(13%)	63	(3%)
(4) Attribute-specifying bundles	2164	(13%)	333	(18%)
(5) Process bundles	1020	(6%)	58	(3%)
TOTAL	**16367**	**(100%)**	**1866**	**(100%)**

doi:10.6342/NTU201600328

As can be seen from the above tables, the type distribution and the token distribution display similar tendencies. Referential expressions are the most dominant functional type for both three-word and four-word spoken bundles: phoric bundles are quite common, because the word sequence 'demonstrative *zhe*/*na* + classifier *ge*/*zhong* + noun' occurs frequently; bundles specifying attributes are also common since there are various attributes (e.g., quantity and similarity). It is also found that three-word spoken bundles are often used to refer to entities, usually in the form of 'modifier + *de* + noun' or 'number + classifier/quantifier + noun'. People are referred to most often (approximately 40%), which reflects that conversation is a participant-oriented text type.

Regarding interpersonal bundles, it is found that three-word bundles are used as epistemic stance bundles most often (86 out of 166, 51.8%), with the word sequence 'subject + mental verb' being the core (e.g., *wo juede zhe* 'I feel this', *danshi wo juede* 'but I feel'). On the other hand, four-word spoken bundles are used as special interactional bundles most often (17 out of 27, 63.0%), and they often contain A-not-A question *shi bu shi* (e.g., *shi bu shi ye* 'A-*not*-A.QUESTION also', *women shi bu shi* 'we A-*not*-A.QUESTION'). There are few three-word attitudinal/modality bundles in conversation, and no four-word spoken bundles are found to be used as attitudinal/modality bundles.

As for discourse organizing functions, three-word bundles are used to elaborate most often (105 out of 161, 65.2%). However, four-word bundles show a roughly equal distribution across the three subcategories.

5.3 Discussion

In the preceding sections, the structural and functional classifications of spoken bundles in Chinese are presented. A closer examination suggests that there is a very

doi:10.6342/NTU201600328

strong relationship between the structural types and the communicative functions of spoken bundles in Chinese. The following tables show the interaction between the structural and functional categories. Since the type distributions and the token distributions show very similar tendencies (see Tables 5.1 and 5.2), only the interaction of bundle types is presented here.

Table 5.3(a). Interaction between structural and functional categories of three-word spoken bundles.

	Clausal bundles	VP-based bundles	NP-based bundles	Others	Total
Interpersonal bundles	91 (55%)	58 (35%)	7 (4%)	10 (6%)	166
Discourse organizers	63 (39%)	54 (34%)	37 (23%)	7 (4%)	161
Referential expressions	19 (6%)	99 (31%)	198 (63%)	0 (0%)	316

Table 5.3(b). Interaction between structural and functional categories of four-word spoken bundles.

	Clausal bundles	VP-based bundles	NP-based bundles	Others	Total
Interpersonal bundles	20 (74%)	6 (22%)	0 (0%)	1 (4%)	27
Discourse organizers	13 (37%)	17 (49%)	4 (11%)	1 (3%)	35
Referential expressions	5 (12%)	19 (44%)	19 (44%)	0 (0%)	43

Table 5.3(a) summarizes the interaction between the structural and functional categories of three-word spoken bundles in Chinese. It is clear that three-word interpersonal bundles are usually realized as clausal or VP-based bundles, which are used mainly to express the speaker’s epistemic stance or to report what was said. Similarly, three-word discourse organizers are usually realized as clausal or VP-based

doi:10.6342/NTU201600328

bundles, but the distributional difference across the three main structural categories is not as sharp. When three-word discourse organizers are realized as clausal bundles, the most common form is ‘connector + pronoun/demonstrative’ (26 out of 63, 41.2%). Both elements have discourse organizing functions: the connector overtly signals the semantic relation between propositions, and the pronoun/demonstrative usually points outwards to something in the text. When three-word discourse organizers are realized as VP-based bundles, it is often the case that the copula *shi* co-occurs with a discourse organizing signal (24 out of 54, 44,4%). Finally, it is found that three-word referential expressions are usually realized as NP-based bundles. This is not surprising, since noun phrases are common grammatical devices to serve referential functions. The strong association between interactional bundles and clausal or VP-based categories and that between referential expressions and NP-based categories are also observed in spoken English (Biber et al. 2004).

Table 5.3(b) summarizes the interaction between the structural and functional categories of four-word spoken bundles. First, it is strikingly clear that four-word interpersonal bundles are realized either as clausal bundles or as VP-based bundles. The proportion of four-word interpersonal bundles realized as clausal bundles is even higher, and these bundles are used mainly to express an epistemic stance or make simple inquiries (i.e., A-*not*-A questions). Second, compared with the structural distribution of three-word discourse organizers, that of four-word discourse organizers is more skewed, with VP-based bundles at the top. Third, four-word referential expressions are not as closely associated with NP-based categories as three-word referential bundles are, but they are still rarely realized as clausal bundles.

In addition to the relationship between structural and functional categories, the relationship between the quantitative measures and the communicative functions of spoken bundles in Chinese is also examined. First, the frequency means of the three

doi:10.6342/NTU201600328

functional categories are presented in the following tables.

Table 5.4(a). Means of relative frequencies (per million words) of three-word spoken bundles.

Interpersonal bundles	Discourse organizers	Referential expressions
60.3	51.4	54.8

Table 5.4(b). Means of relative frequencies (per million words) of four-word spoken bundles.

Interpersonal bundles	Discourse organizers	Referential expressions
40.0	40.7	36.2

The Shapiro-Wilk normality test shows that the frequencies in each group are not normally distributed, so the Kruskal-Wallis rank sum test is performed on the means. The results show that the frequency differences in the above two tables are not statistically significant (p = 0.473, p = 0.388, respectively). Second, the DP means of the three functional categories are presented in the following tables.

Table 5.5(a). DP means of three-word spoken bundles.

Interpersonal bundles	Discourse organizers	Referential expressions
0.42	0.39	0.40

Table 5.5(b). DP means of four-word spoken bundles.

Interpersonal bundles	Discourse organizers	Referential expressions
0.44	0.42	0.40

While the DP values of three-word spoken bundles in the three groups do not follow normal distributions, those of four-word spoken bundles in the three groups are normally distributed. The Kruskal-Wallis rank sum test and the one-way ANOVA are thus run respectively. The results show that the DP differences in the above two tables are not statistically significant (p = 0.171, p = 0.276, respectively). Third, the G means of the three functional categories are presented in the following tables.

doi:10.6342/NTU201600328

Table 5.6(a).　G means of three-word spoken bundles.

Interpersonal bundles	Discourse organizers	Referential expressions
3.3	2.9	3.3

Table 5.6(b).　G means of four-word spoken bundles.

Interpersonal bundles	Discourse organizers	Referential expressions
3.2	4.1	3.2

Still, the Kruskal-Wallis rank sum test and the one-way ANOVA are run respectively. The results show that the G differences in the above two tables are statistically significant (p = 0.006, p = 0.036, respectively). In brief, the functions of spoken bundles in Chinese have an impact on their internal association, but do not influence their frequency distribution or their dispersion in the corpus.

The results in the present study reveal many similarities between Chinese and English in the use of spoken bundles (see Biber et al. 1999, Biber et al. 2004). First, some structural types occur frequently in both languages, including 'connector + clause fragment', 'pronoun + verb + complement clause fragment', and 'pronoun + verb phrase fragment'. Second, after minor modifications, the functional framework for English bundles are highly applicable to Chinese bundles, and this suggests that spoken bundles in both languages share similar functions. Listed below are more functional similarities:

(i)　there are more personal stance bundles than impersonal stance bundles;

(ii)　epistemic bundles are usually used to express uncertain stances; and

(iii)　some prepositional phrases serve as multifunctional referential expressions.

Third, both languages show a similar relationship between structural and functional categories for spoken bundles: i.e., stance bundles are usually realized as clausal or VP-based bundles, and referential bundles are usually realized as NP-based bundles.

doi:10.6342/NTU201600328

There are also some differences between Chinese and English in the use of lexical bundles in conversation. First, the following are obvious structural differences:

(i) English has specific question constructions (e.g., '*wh*-word + copula/auxiliary + pronoun') and thus features a large number of *wh*-question bundles, while Chinese is a *wh*-in-situ language and has few *wh*-question bundles;

(ii) passive verb phrases (e.g., *is based on*, *can be used to*) are common in English, while the common passive marker *bei* in Chinese is not attested in spoken bundles identified in the present study;

(iii) English features many NP-based bundles with post-modifier fragments (e.g., *one of the things*, *the way in which*), while Chinese features many NP-based bundles with pre-modifier fragments (i.e., 'modifier + *de* + noun'); and

(iv) since a classifier is usually required to co-occur with a number or a demonstrative before a noun in Chinese, classifiers (e.g., *ge*) occur far more frequently in Chinese bundles than in English bundles.

Second, there are striking differences in the distributional patterns of spoken bundles in the two languages. It is evident that NP-based bundles are much more common in spoken Chinese than in spoken English. Referential expressions, which are closely associated with NP-based categories, are also much more common in spoken Chinese than in spoken English. These distributional differences can be ascribed to structural characteristics of Chinese. In Chinese, the number *yi* 'one' and the demonstratives *zhe* 'this' and *na* 'that' frequently co-occur with a classifier, especially the neutral classifier *ge*. The sequence 'number/demonstrative + classifier' is NP-based and serves referential functions. The variations of the sequence (e.g., *yi ge*, *zhe ge*, *na zhong*) are often combined with another high-frequency word, and a large number of NP-based referential bundles emerge accordingly.

5.4 Summary

This chapter has extensively investigated the use of lexical bundles in spoken Chinese. Just as in English, most spoken bundles in Chinese are not structurally complete and run across traditional grammatical structures. However, these bundles can be systematically grouped according to their structural characteristics (see Table 5.1). More importantly, these bundles serve as building blocks in spoken discourse, facilitating interpersonal communication (e.g., expressing stances, eliciting responses), organizing discourse (e.g., introducing and elaborating topics), and having a variety of referential uses. The fact that lexical bundles in spoken Chinese have well-defined structural and functional correlates confirms that high-frequency word strings have essential linguistic status instead of being repeated again and again randomly (Biber et al. 2004). In addition, the results also lend cross-linguistic support to the strong relationship between the structural and functional categories. Just as in English and Spanish, stance bundles and referential expressions in spoken Chinese are closely associated with clausal/VP-based categories and NP-based categories respectively. Last but not least, a qualitative comparison between Chinese and English suggests that structural characteristics specific to a language have an impact on the distributional patterns of lexical bundles (Tracy-Ventura et al. 2007).

doi:10.6342/NTU201600328

Chapter 6

Lexical Bundles in News

In the previous chapter, lexical bundles in spoken Chinese are examined in detail, and a qualitative cross-linguistic comparison between Chinese and English regarding the use of lexical bundles is made. In this chapter, we will turn to lexical bundles in a common text type of the written mode, i.e., news writing, and the focus will be shifted onto comparisons between spoken and news bundles.

6.1 Structural Classification of Lexical Bundles in News

The structural framework proposed in the previous chapter is not just tailor-made for lexical bundles in spoken Chinese, but is also potentially applicable to Chinese bundles in other text types. The results show that no new category is needed for news bundles in Chinese. However, some structural categories are not attested in the news subcorpus.

6.1.1 Clausal Bundles

There are seven subcategories for clausal bundles, but only three of them are attested in the news subcorpus.

The first one is 'noun phrase + copula *shi* (+ …)'. Among the five news bundles in this subcategory, four bear the demonstrative *zhe* 'this' in the subject position (e.g., <u>*zhe*</u> *ye shi* 'this also is'). The only four-word news bundle in this subcategory is *zui zhongyao de shi* 'most important DE be; the most important thing is that'.

The second subcategory is 'pronoun + verb (+ clause fragment)'. There is only one news bundle in this subcategory, i.e., *you ren shuo* 'have people say; there are some people saying that'. Like *say* in English, the saying verb *shuo* usually takes a clause as its complement.

Lastly, there is an A-*not*-A question bundle in the news subcorpus, i.e., *shi bu shi*. This bundle occurs frequently not only in conversation but also in news.

6.1.2 VP-based Bundles

There are seven subcategories for VP-based bundles. Still, not all of them are attested in the news subcorpus.

The first subcategory is ‘copula *shi* (+ …)’. The copula *shi* here is often preceded by an adverb (e.g., *jiu shi yi* ‘just be one’), a connector (e.g., *ye shi yi* ‘also be one’), a modal expression (e.g., *keyi shuo shi* ‘can say be; it can be said that’), or a negator (e.g., *er bu shi* ‘yet not be’).

The second subcategory is ‘verb + noun phrase (fragment)’. Only the verb *you* ‘have’ is attested. Examples include *you yi ge* ‘have one CLASSIFIER’, *you butong de* ‘have different DE’, and *ye you ren* ‘also have people’. The only four-word news bundle in this subcategory is *you hen da de* ‘have very large DE’.

The third subcategory is ‘verb + *de*/*bu* + complement’. There is only one news bundle falling into this subcategory, i.e., *zhao bu dao* ‘cannot find out’. This bundle occurs frequently not only in conversation but also in news.

6.1.3 NP-based Bundles

There are six subcategories for NP-based bundles. Still, not all of them are found in the news subcorpus.

The first subcategory is noun phrases. Most news bundles falling into this subcategory are time expressions (e.g., *mei ge yue* ‘every CLASSIFIER month’, *qunian tong qi* ‘last.year same period’) The only four-word news bundle in this subcategory is *mei ge ren dou* ‘every CLASSIFIER person all’.

The second subcategory is ‘*de* + noun phrase (fragment)’. Examples include *de qingkuang xia* ‘DE circumstance under; under the circumstances’ and *de yi zhong* ‘DE one kind; one kind of’. In spoken bundles that fall into this subcategory, the word *de* is often preceded by a modifier (see also Section 5.1.3). However, news bundles here all start with *de*.

The third subcategory is ‘modifier (+ *de*)’. Some bundles in this subcategory contain an extremity-signaling adverb (e.g., *zui zhongyao de* ‘most important DE’). Some bundles use nominal modifiers (e.g., *zhe ci de* ‘this time DE’, *ge di de* ‘every place DE’). Some bundles feature relative constructions (e.g., *suo zuo de* ‘PARTICLE do DE’, *suo xu de* ‘PARTICLE need DE’).

The fourth subcategory is ‘preposition + noun phrase (fragment)’. Only the locative preposition *zai* is attested. Examples include *zai zhe ge* ‘at/in/on this CLASSIFIER’ and *zai wanglu shang* ‘on Internet up; on the Internet’.

6.1.4 Structural Distributions of Lexical Bundles

The following tables present the structural distribution of Chinese bundles in news. Both the type distribution and the token distribution are presented. Because there are only three four-word news bundles identified in the present study, only the structural distribution of three-word news bundles is tabulated.

Table 6.1(a). Structural distribution of lexical bundle types in news.

Structural category	Three-word news	
1. Clausal bundles	**6**	**(6.9%)**
(1) noun phrase (fragment) + copula *shi* (+ …)	4	(4.6%)
(2) (connector +) pronoun + verb (+ clause fragment)	1	(1.1%)
(3) pronoun + verb (+ …)	0	(0.0%)
(4) pronoun + modal expression	0	(0.0%)
(5) connector + pronoun/demonstrative	0	(0.0%)
(6) A-*not*-A question	1	(1.1%)
(7) others	0	(0.0%)
2. VP-based bundles	**28**	**(32.2%)**
(1) copula *shi* (+ …)	18	(20.7%)
(2) verb (+ clause fragment)	0	(0.0%)
(3) adjectival verb	0	(0.0%)
(4) verb + noun phrase (fragment)	9	(10.3%)
(5) modal expression (+ verb)	0	(0.0%)
(6) (verb +) *de*/*bu* + complement (fragment)	1	(1.1%)
(7) others	0	(0.0%)
3. NP-based bundles	**53**	**(60.9%)**
(1) noun phrase	20	(23.0%)
(2) noun phrase + adverb	0	(0.0%)
(3) (modifier +) *de* + noun phrase (fragment)	7	(8.0%)
(4) modifier (+ *de*)	22	(25.3%)
(5) preposition + noun phrase (fragment)	4	(4.6%)
(6) others	0	(0.0%)
4. Others	**0**	**(0.0%)**
TOTAL	**87**	**(100.0%)**

doi:10.6342/NTU201600328

Table 6.1(b). Structural distribution of lexical bundle tokens in news.

Structural category	Three-word news	
1. Clausal bundles	**1663**	**(7.8%)**
(1) noun phrase (fragment) + copula *shi* (+ …)	1148	(5.4%)
(2) (connector +) pronoun + verb (+ clause fragment)	145	(0.7%)
(3) pronoun + verb (+ …)	0	(0.0%)
(4) pronoun + modal expression	0	(0.0%)
(5) connector + pronoun/demonstrative	0	(0.0%)
(6) A-*not*-A question	370	(1.7%)
(7) others	0	(0.0%)
2. VP-based bundles	**7349**	**(34.4%)**
(1) copula *shi* (+ …)	5235	(24.5%)
(2) verb (+ clause fragment)	0	(0.0%)
(3) adjectival verb	0	(0.0%)
(4) verb + noun phrase (fragment)	1933	(9.1%)
(5) modal expression (+ verb)	0	(0.0%)
(6) (verb +) *de*/*bu* + complement (fragment)	181	(0.8%)
(7) others	0	(0.0%)
3. NP-based bundles	**12325**	**(57.8%)**
(1) noun phrase	4683	(21.9%)
(2) noun phrase + adverb	0	(0.0%)
(3) (modifier +) *de* + noun phrase (fragment)	1748	(8.2%)
(4) modifier (+ *de*)	5129	(24.0%)
(5) preposition + noun phrase (fragment)	765	(3.6%)
(6) others	0	(0.0%)
4. Others	**0**	**(0.0%)**
TOTAL	**21337**	**(100.0%)**

As can be seen from the above tables, the type distribution and the token distribution display similar tendencies. All the news bundles fit comfortably in with the structural framework for Chinese bundles, and this reconfirms that lexical bundles have strong structural correlates.

For three-word bundles in news, the most common structural type is NP-based bundles, followed by VP-based bundles and clausal bundles. The most common NP-based type is 'modifier (+ *de*)'. Full noun phrases are also quite common, mostly with 'number/demonstrative + classifier' being the core. As for the VP-based categories, 'copula *shi* (+ …)' and 'verb + noun phrase (fragment)' occur most frequently. This trend has been observed in spoken bundles (see Section 5.1.5), suggesting again that the copula *shi* plays an essential role in VP-based bundles, and that the word order VO is a dominant argument structure in Chinese bundles. Only a few three-word news bundles are clausal. It is found that they usually contain the copula *shi* (4 out of 6, 66.7%).

6.2 Functional Classification of Lexical Bundles in News

Like the structural framework presented above, the functional framework in the previous chapter (see Section 5.2) is also potentially applicable to Chinese bundles in various text types. The results show that news bundles in Chinese can be adequately accommodated in this functional framework.

6.2.1 Interpersonal Bundles

The first subcategory is special interactional bundles. There are only three bundles in this subcategory: i.e., *shi bu shi* 'A-*not*-A QUESTION', *shenme yang de* 'what kind DE', and *you ren shuo* 'have people say'. These bundles are also identified in the conversation subcorpus.

The second subcategory is epistemic stance bundles. In news writing, the writer’s epistemic stance can be expressed through modal expressions (e.g., *keyi shuo shi* ‘it can be said that’ conveys certainty) or the negator *bu* ‘not’ (e.g., *bu shi hen* ‘not very’ serves as a mitigator). It is found that stance bundles in news writing are all impersonal.

There are supposed to be three subcategories for interpersonal bundles. However, no attitudinal/modality bundle is identified in the news subcorpus.

6.2.2 Discourse Organizers

The first subcategory is topic introduction bundles. All the news bundles in this subcategory are presentational phrases, i.e., *you yi ge*/*zhong*/*time*/*wei* ‘have one CLASSIFIER; there is a’.

The second subcategory is topic elaboration bundles. The additive adverb *ye* ‘also’ often occurs in elaboration bundles. Examples include *ye you ren* ‘also have people; there are also some people’, which is used to identify another group of people, and *ye jiu shi* ‘also just be; that is to say’, which is used to paraphrase what is just said. The word *ling* ‘another’ also occurs in elaboration bundles very often (e.g., *ling yi ge*/*zhong* ‘another one CLASSIFIER’), and these bundles are used to enumerate more items.

The third subcategory is identification bundles. In news writing, identification bundles are used to highlight something as newsworthy and grab the reader’s attention. Therefore, extremity-signaling adverbs often occur in identification bundles. The adverb *zui* ‘most’ is the most common one (e.g., *zui da de* ‘most large DE’, *zui zhongyao de* ‘most important DE’), and *ji* ‘extremely’ is also attested (e.g., *ji da de* ‘extremely large DE’). Identification bundles are also used after a lengthy elaboration to provide a concise summary or brief comments. Consider *zhe shi yi* ‘this is a’ in the following example.

(6.1) Wo ganjuedao meiyou ren bu zhongshi, bu guanxin ziji guojia de anquan. Suoyou de ren dou shi zai zheyang de jichu shang fabiao ziji de yijian. Suiran you ren leguan, you ren baoliu, dan suoyou ren de chufadian dou shi yiyang. Wo juede zhe shi yi ge **man kexi de xianxiang**.

‘I feel that all people value and show concern for the safety of their own country. Everyone expresses personal opinions on the same basis. Although some are optimistic and some are conservative, all people have the same point of departure. I feel that this is a **very positive phenomenon**.’

In (6.1), the identification bundle *zhe shi yi* is followed by a brief opinion (i.e., *man kexi de* ‘very positive DE’) and a shell noun (i.e., *xianxiang* ‘phenomenon’) that serves as a cohesive device to enclose the preceding discourse.

6.2.3 Referential Expressions

The first subcategory is phoric bundles. No phoric bundles in the news subcorpus feature the demonstrative *na* ‘that’; only *zhe* ‘this; these’ is attested (e.g., *zhe ge wenti* ‘this CLASSIFIER problem’, *zhe ji nian* ‘these several year; in recent years’).

The second subcategory is entity bundles. Most entity bundles in the news subcorpus are time expressions (e.g., *ge yue nei* ‘CLASSIFIER month within; within … months’, *shang ban nian* ‘up half year; the first half year’).[36] A time frame is

[36] In the present study, high-frequency word sequences with specific numbers other than *yi* ‘one’ are manually excluded (see Section 3.4). However, *san ge yue* ‘three CLASSIFIER month’ and *liu ge yue* ‘six CLASSIFIER month’ remain in the data. The former is a quarter, and the latter is half a year. These two are common time frames in news writing, particularly in business news.

necessary for the presentation of a news event and the interpretation of news data (McKane 2006). For example, *seven joyriders were killed* means little unless a time frame such as *in the past year* is added. Entity bundles referring to places (e.g., *shijie ge di* 'world every place; around the world', *zai wanglu shang* 'on Internet up; on the Internet') and people (e.g., *de ren dou* 'DE person all; those who … all') are also attested.

The third subcategory is attribute-specifying bundles. Some bundles in this subcategory describe intangible attributes, including types (e.g., *de yi zhong* 'DE one kind; one kind of'), processes (e.g., *de guocheng zhong* 'DE process middle; in the process of'), and purposes (e.g., *shi wei le* 'be for LE; for the purpose of'). Other bundles describe tangible qualities, such as sizes (e.g., *hen da de* 'very large DE') and quantities (e.g., *zhi you yi* 'only have one').

The fourth subcategory is process bundles. In the news subcorpus, bundles expressing relational processes (see Section 5.2.3) are the most common (5 out of 11, 45.5%) and usually feature the copula *shi* (e.g., *shi yi ge* 'be one CLASSIFIER', *bu shi yi* 'not be one'). In news writing, these copula-bearing bundles are thought to convey the journalist's certain stance, i.e., showing that the information presented is seen as factual (Kaneyasu 2015).

Referential expressions also include imprecision bundles (e.g., *or something like that*). However, imprecision bundles are not attested in the news subcorpus.

6.2.4 Functional Distributions of Lexical Bundles

Table 6.2 presents the functional distribution of Chinese bundles in conversation. Still, both the type distribution and the token distribution are presented. Because there are only three four-word news bundles identified in the present study, only the structural distribution of three-word news bundles is tabulated.[37]

As can be seen from Table 6.2, the type distribution and the token distribution display similar tendencies. For three-word news bundles, the most common functional category is referential expressions, followed by discourse organizers and interpersonal bundles. Entity bundles are the most frequent referential expressions. The high frequency of entity bundles is attributed to its close association with two common patterns ‘modifier + *de* + noun’ and ‘number/demonstrative + classifier/quantifier + noun’. Besides, as mentioned above, time frames are essential elements in news writing, and they are usually classified as entity bundles. As for discourse organizers, identification bundles and topic elaboration bundles are common subcategories. Identification bundles provide a focus for something newsworthy and are often used to attract the reader’s attention. Only a marginal proportion of news bundles are interpersonal bundles, and no attitudinal/modality bundles are identified in the present study.

[37] The three four-word news bundle identified in the present study are *zui zhongyao de shi* ‘most important DE be; the most important thing is that’ (identification bundle), *you hen da de* ‘have very large DE’ (attribute-specifying bundle), and *mei ge ren dou* ‘every CLASSIFIER person all’ (entity bundle).

doi:10.6342/NTU201600328

Table 6.2(a). Functional distribution of lexical bundle types in news.

Functional category	Three-word news	
1. Interpersonal bundles	**6**	**(6.9%)**
(1) Epistemic stance bundles	3	(3.4%)
(2) Attitudinal/Modality stance bundles	0	(0.0%)
(3) Special interactional bundles	3	(3.4%)
2. Discourse organizers	**31**	**(35.6%)**
(1) Topic introduction bundles	4	(4.6%)
(2) Topic elaboration bundles	12	(13.8%)
(3) Identification bundles	15	(17.2%)
3. Referential expressions	**50**	**(57.5%)**
(1) Imprecision bundles	0	(0.0%)
(2) Phoric bundles	5	(5.7%)
(3) Entity bundles	21	(24.1%)
(4) Attribute-specifying bundles	13	(14.9%)
(5) Process bundles	11	(12.6%)
TOTAL	**87**	**(100.0%)**

Table 6.2(b). Functional distribution of lexical bundle tokens in news.

Functional category	Three-word news	
1. Interpersonal bundles	**1334**	**(6.3%)**
(1) Epistemic stance bundles	601	(2.8%)
(2) Attitudinal/Modality stance bundles	0	(0.0%)
(3) Special interactional bundles	733	(3.4%)
2. Discourse organizers	**8344**	**(39.1%)**
(1) Topic introduction bundles	1045	(4.9%)
(2) Topic elaboration bundles	3350	(15.7%)
(3) Identification bundles	3949	(18.5%)
3. Referential expressions	**11659**	**(54.6%)**
(1) Imprecision bundles	0	(0.0%)
(2) Phoric bundles	910	(4.3%)
(3) Entity bundles	4254	(19.9%)
(4) Attribute-specifying bundles	3810	(17.9%)
(5) Process bundles	2685	(12.6%)
TOTAL	**21337**	**(100.0%)**

6.3 Discussion

The structural and functional classifications of both spoken and news bundles in Chinese have been presented. This section will compare lexical bundles in conversation and news.

Structurally speaking, news bundles are more restricted than spoken bundles in two respects. First, some structural categories of lexical bundles are found in conversation but not in news (e.g., bundles with adjectival verbs). Second, structural variation within a subcategory is not as great in news as in conversation. For example, four A-*not*-A question types are identified in conversation, but only one of them (i.e., *shi bu shi*) occurs frequently in news.

Another structural difference is that disyllabic words are preferred in conversation bundles, while monosyllabic words are preferred in news bundles. For example, both *ke shuo shi* 'can say be; it can be said that' and *keyi shuo shi* 'can say be' are identified in news, but the former occurs more frequently. In conversation, only *keyi shuo shi* is identified. The second example is that *ling* 'another' and its disyllabic counterpart *lingwai* 'another' have a complementary distribution in lexical bundles. The former is found only in news bundles (e.g., *ling yi ge* 'another one CLASSIFIER'), whereas the latter is found only in conversation bundles (e.g., *lingwai yi ge* 'another one CLASSIFIER'). Still another finding is that disyllabic connectors (e.g., *ranhou* 'then', *danshi* 'but', *yinwei* 'because', *suoyi* 'so') are found only in conversation bundles, but not in news bundles. It has long been recognized that speaking and writing show enormous differences in vocabulary choices (e.g., Chafe 1985). More specifically, shorter forms are preferred in news writing. It is essential to avoid wordiness so that the paper's space and the reader's time will not be wasted (e.g., Bagnall 1993, McKane 2006).

doi:10.6342/NTU201600328

Regarding the structural distribution of Chinese bundles in conversation and news, an obvious difference is that there are fewer clausal bundles in news than in conversation. Clausal bundles are often used to convey uncertain stances in conversation, but journalists tend to avoid personal uncertain stances in their writing. However, there are also some similarities, as listed below:

(i) the type distribution and the token distribution display similar tendencies;

(ii) NP-based categories are dominant both in conversation and in news; and

(iii) 'copula *shi* (+ …)' and 'verb + noun phrase (fragment)' are two common VP-based categories both in conversation and in news.

Many functional differences are also found in the use of lexical bundles in conversation and news. These differences reflect some general principles of news writing (e.g., Hough 1988, Berner 1992, Bagnall 1993, Wang 1995, Kovach and Rosenstiel 2001, Rich 2005, McKane 2006, Ma 2007, Peng 2008).

The number of interpersonal bundles in news is much lower than that in conversation. First, fewer question bundles are identified in news. Journalists tend to conform to the principle "give us the answers not the questions" (McKane 2006: 110). Second, fewer reporting bundles (i.e., bundles with verbs of saying) are identified in news. In conversation, verbs of saying often take a high-frequency pronoun in the subject position (i.e., <u>*wo*</u> *jiu shuo* 'I then say'), so many reporting bundles gain high frequencies accordingly. However, in news writing, pronouns are dispreferred (Ma 2007: 249). Verbs of saying usually take as subjects specific information sources, which do not occur so frequently as pronouns. Third, epistemic stance bundles occur far more frequently in conversation than in news. Besides, while epistemic stance bundles in conversation are usually personal (i.e., directly attributing stances to the speaker), those in news are all impersonal. Spoken language often use hedges (Dooley and Levinsohn 2001: 17), and common devices include epistemic bundles which

doi:10.6342/NTU201600328

express an uncertain stance. On the other hand, journalists are expected to stick to facts, and it is inappropriate for them to express their own subjective attitudes or emotions. However, note that the rarity of personal epistemic bundles in news can also be recognized as a kind of "faceless" stance (Biber and Finegan 1988: 31) that avoids personal judgments.

Regarding discourse organizers, topic identification bundles occur more frequently than topic elaboration bundles in news, while the opposite trend is observed in conversation. The degree adverb *zui* 'most' is commonly used in news to identify something newsworthy to the reader, as in the sequence '*zui* + adjective + *de*'. For example, the bundle *zui xin de* 'most new DE; the latest', which is identified in news but not in conversation, is used to introduce the latest development of a news event or a popular product.

There are five subcategories for referential expressions in conversation, but one of them is not identified in news, i.e., imprecision bundles. Journalists are required to be specific and avoid ambiguity; otherwise, misunderstandings may arise, and the newspaper's credibility with the reader may be undermined. Likewise, some bundles specifying indefinite quantities, such as *yi da dui* 'one large pile; a lot of' and *you henduo de* 'have many DE; there are many', are identified in conversation but not in news. These bundles may be regarded as space wasters in news writing because no precise information is provided. Another striking difference observed in referential expressions is that entity bundles referring to times are much more common in news than in conversation. As mentioned in Section 6.2.3, timeliness makes an event newsworthy, and time frames are essential for news events.

Even though there are many functional differences in the use of spoken and news bundles, many similarities are still observed. The following are three important similarities:

doi:10.6342/NTU201600328

(i) attitudinal/modality bundles rarely occur, whether in conversation or in news;

(ii) presentational bundles featuring the sequence *you yi* ‘there is one’ are common topic introduction bundles both in conversation and in news; and

(iii) referential expressions are the most dominant category both in conversation and in news.

With regard to the interaction between structural and functional categories, spoken and news bundles share many similarities. The following table illustrates the interaction between structural and functional categories for three-word news bundles.[38]

Table 6.3. Interaction between structural and functional categories of three-word news bundles.

	Clausal bundles	VP-based bundles	NP-based bundles	Total
Interpersonal bundles	2 (33%)	3 (50%)	1 (17%)	6
Discourse organizers	4 (13%)	14 (45%)	13 (42%)	31
Referential expressions	0 (%)	11 (22%)	39 (78%)	50

[38] Since there are only three four-word news bundles, such a table is not created for them. Also, since the type distributions and the token distributions show very similar tendencies (see Tables 6.1 and 6.2), only the interaction of bundle types is presented here.

doi:10.6342/NTU201600328

As can be seen from the above table, there is a relationship between the structural and functional categories of news bundles in Chinese. It is clear that referential expressions are strongly associated with NP-based categories. In contrast, interpersonal bundles appear to be closely associated with VP-based and clausal categories.[39] These tendencies are also observed in spoken bundles and echo findings in English and Spanish. There is also a notable difference, though: while three-word discourse organizers in conversation are distributed more evenly across the three main structural categories (see Table 5.3(a)), those in news are associated more closely with VP-based and NP-based categories. This is attributed to the rarity of clausal bundles in news (see Section 6.1.4).

Some lexical bundles occur frequently both in conversation and in news. It is found that nearly half of the three-word news bundles (43 out of 87, 49%) and all the three four-word news bundles are also identified in the spoken subcorpus. There are two possible explanations for the considerable overlap. First, news writing, albeit being a written mode, aims to convey messages effectively and efficiently, so journalists are told to write the way people talk (Itule and Anderson 1994). Second, there may exist core bundles that serve essential communicative functions in discourse, so they may be identified in all text types. The following table presents the functional distribution of three-word bundles that are identified both in conversation and in news.

[39] Since there are only six interpersonal bundles identified in the news subcorpus, this tendency is considered to be tentative.

doi:10.6342/NTU201600328

Table 6.4. Functional distribution of three-word bundles identified both in conversation and in news.

Functional category		Type number
Interpersonal bundles	Special interactional bundles	3
	Epistemic stance bundles	2
	Total	**5**
Discourse organizers	Topic elaboration bundles	7
	Identification bundles	7
	Topic introduction bundles	2
	Total	**16**
Referential expressions	Attribute-specifying bundles	9
	Entity bundles	6
	Process bundles	4
	Phoric bundles	3
	Total	**22**

The first glance would suggest that compared with the other two functional categories, more referential expressions are identified in both text types. However, given the high frequencies of referential expressions in both text types, the overlap is actually small in proportional terms. Since referential expressions are closely associated with NP-based categories, the relatively small overlap here echoes Dolch’s (1936) argument that nouns are tied to specific subjects, and few of them are of universal use. On the other hand, there are only six interpersonal bundles in news, and five of them are also identified in conversation. That is, most interpersonal bundles in news also occur frequently in conversation, but not vice versa. As mentioned previously, personal epistemic bundles, which are common in conversation, are avoided in news writing.

Finally, regarding the relationship between the quantitative measures and the communicative functions of lexical bundles, conversation and news display very similar tendencies (see Section 5.3). First, the frequency means of the three functional categories, as presented in the following table, are not statistically different (*p* =

0.453). (The Shapiro-Wilk normality test shows that the frequencies in each group are not normally distributed, so the Kruskal-Wallis rank sum test is performed on the means.)

Table 6.5. Means of relative frequencies (per million words) of three-word news bundles.

Interpersonal bundles	Discourse organizers	Referential expressions
34.3	41.6	36.0

Second, the DP means, which are presented in the following table, are not statistically different either (*p* = 0.501). (The Shapiro-Wilk normality test shows that the DP values of discourse organizers are not normally distributed, so the Kruskal-Wallis rank sum test is performed on the means.)

Table 6.6. DP means of three-word news bundles.

Interpersonal bundles	Discourse organizers	Referential expressions
0.16	0.14	0.16

Third, as the following table shows, the G values of discourse organizers are lower than those of the other two categories. This trend is statistically significant in the spoken subcorpus, yet not in the news subcorpus (*p* = 0.165). (The Shapiro-Wilk normality test shows that the G values in each group are normally distributed, so the one-way ANOVA is performed on the means.)

Table 6.7. G means of three-word news bundles.

Interpersonal bundles	Discourse organizers	Referential expressions
4.7	3.4	3.9

In brief, the functions of news bundles in Chinese do not influence their frequency distribution, dispersion degree, or internal association.

doi:10.6342/NTU201600328

6.4 Summary

This chapter has taken an unprecedented step to investigate news bundles in Chinese. It is found that the structural and functional classifications for spoken bundles in Chinese are both applicable to news bundles, and the close relationship between the structural and functional categories of lexical bundles (e.g., referential expressions are strongly associated with NP-based categories) is reconfirmed. Besides, as is the case in conversation, structural resources available in Chinese also profoundly influence the distribution of lexical bundles in news. The frequent occurrences of the sequence ‘number/demonstrative + classifier/quantifier + noun’ in Chinese contribute to the high frequencies of NP-based bundles and referential expressions in both conversation and news.

Interestingly, the conventions of news writing as an information-oriented text type that processes and packages messages in unique ways make the use of news bundles different from as well as similar to that of spoken bundles. On the one hand, the finding that nearly half of the news bundles are also identified in conversation may relate to the general principle that journalists should write in ways people talk (Itule and Anderson 1994). On the other hand, other principles (e.g., sticking to facts, avoid ambiguities, relating news events to readers, using shorter forms) make some bundles occur more/less frequently in news than in conversation. It is evident that the use of lexical bundles in Chinese newswire texts is reflective of complex interactions between language-specific structural properties and genre-specific communicative needs.

doi:10.6342/NTU201600328

Chapter 7

Conclusion

The present study follows the Biberian approach (Biber et al. 1999), which is generally frequency-based, to identify lexical bundles in Chinese, but more quantitative criteria are adopted. A data exploratory analysis of high-frequency word sequences in Chinese has been conducted in Chapter 4, and a more careful analysis of spoken and news bundles in Chinese has been presented in Chapters 5 and 6 respectively. The current chapter will summarize the findings (Section 7.1), highlight theoretical contributions (Section 7.2), suggest some practical applications of the present study (Section 7.3), pinpoint some limitations (Section 7.4), and offer some directions for future research on lexical bundles in Chinese (Section 7.5).

7.1 Main Findings of the Study

The method developed to identify lexical bundles in the present study complements the Biberian approach. First, a more sensitive dispersion measure, i.e., DP (Gries 2008b), is shown to effectively filter out local repetitions. Second, a word association measure, i.e., G (Wei and Li 2013), is adopted to exclude word sequences that simply contain high-frequency function words and lack identifiable functions. However, as Salazar (2014) suggests, even with the help of all the quantitative thresholds, manual interventions are still needed to screen out semantically/pragmatically vague word sequences.

The data exploratory analysis reveals some distributional differences between spoken and news bundles in Chinese. Conversations feature a much wider range of different lexical bundles than newswire texts, and a much larger proportion of spoken data is covered by lexical bundles. These findings are consistent with those in English (e.g., Biber et al. 1999), suggesting that speakers under the real-time pressure of spontaneous speech rely heavily on prefabricated chunks.

A further investigation shows other similarities between Chinese and English in the use of spoken bundles. Just as in English, most spoken bundles in Chinese are structurally incomplete, but they can be closely grouped according to their grammatical characteristics. Additionally, the functional framework for English bundles is highly applicable to Chinese bundles, which are also used mainly to facilitate interpersonal interactions, organize discourses, and identify items. Furthermore, the close relationship between the structural and functional categories of lexical bundles is reconfirmed: stance bundles are strongly associated with clausal and VP-based categories, while referential expressions are strongly associated with NP-based categories. However, a striking difference between Chinese and English is that NP-based bundles are much more dominant in Chinese, and this is attributed to structural characteristics specific to Chinese.

A comparison between spoken and news bundles in Chinese is also made. Many similarities are revealed: news bundles fit comfortably in with the structural and functional frameworks of spoken bundles, and the relationship between the structural and functional categories also exists. Nevertheless, there are also many differences, such as fewer question bundles, reporting bundles, and epistemic stance bundles in the news subcorpus. These differences reflect some general principles of news writing (e.g., being accurate and specific, avoiding personal judgments).

7.2 Theoretical Contributions

This may be the most comprehensive study so far systematically identifying lexical bundles in Chinese (see also Tao 2015). The results agree with those of previous studies adopting the Biberian approach, demonstrating that the components

doi:10.6342/NTU201600328

in lexical bundles co-occur very frequently due to their important functions rather than simply by chance. The major functional categories of lexical bundles identified in English (Conrad and Biber 2004) gain cross-linguistic support since they have been identified not only in Spanish (Tracy-Ventura et al. 2007) and Korean (Kim 2009) but also in Chinese. Moreover, the comparison between news and spoken bundles in Chinese also assists in our understanding of how the use of lexical bundles is significantly influenced by complex interactions between language-specific structural properties and genre-specific communicative needs. We have seen that the dominance of the sequence ‘number/demonstrative + classifier/quantifier + noun’ biases Chinese towards NP-based bundles, and that the conventions of news writing make the distributional patterns of news bundles in Chinese distinctly different from those of spoken bundles.

The present study also lends cross-linguistic support to Bybee’s (2007) usage-based proposal that high-frequency multi-word units such as lexical bundles are emergent storage/processing units in the mental lexicon. A brief overview is provided here. In this model, human beings are said to be innately equipped with complex cognitive mechanisms that empower us to unconsciously categorize linguistic units of varying sizes (e.g., sounds, words, phrases, and even clauses) and perform probabilistic analyses of what units co-occur frequently, and a vast memory that stores our language experience. Given the powerful brain, language users come to know that some words (e.g., *I don’t know*) frequently co-occur because they serve important functions. Through our repeated use, lexical bundles have strong representations in the memory, emerge as storage/processing units, and become more readily accessible to language users. These prefabricated chunks are easy to produce and comprehend because any internal analysis appears to be unnecessary. Speakers rely heavily on these chunks to achieve their fluency. The evidence for the above proposal is from

Scheibman (2000). First, the reduction of the vowel in *don't* to a schwa is found to occur only in lexical bundles such as *I don't know*, *I don't think*, and *why don't you*, where *don't* occurs most frequently.[40] The holistic representation of lexical bundles is thought to be physically realized in articulatory gestures. Second, the three bundles have a meaning that is not the literal combination of their components: for example, the uncertainty stance bundle *I don't know* evolves into a disagreement mitigator (Aijmer 2008). Both the phonetic reduction and the semantic change suggest that the boundaries between the parts of these bundles have become blurred and that these bundles are emergent storage/processing units. However, note that the above proposal does not preclude the possibility that lexical bundles can be compositionally combined.

Some semantic changes observed in the present study support Bybee's (2007) usage-based model, revealing that lexical bundles in Chinese are emergent storage/processing units, achieving strong representations in the mental lexicon. First, like *I don't know*, some Chinese bundles have developed non-compositional meanings. For example, the A-*not*-A question bundle *shi bu shi* is sometimes used to yield the conversation floor rather than elicit a response (Tao 2015: 343). Second, decategorization (Hopper 1991) is seen in some Chinese bundles. For example, the verbal meaning of *shuo* 'say' is bleached in lexical bundles such as *wo juede* <u>*shuo*</u> 'I feel' and *de yisi shi* <u>*shuo*</u> 'what someone/something means is'. This suggests that the word *shuo* has fused with the other elements in these bundles.

Phonetic evidence for lexical bundles being storage/processing units in Chinese is lacking in the present study because only the transcripts of the conversations are available in the Sinica Corpus. However, there exist some clues. For example, in the conversation subcorpus, the word *na* in elaboration bundles such as <u>*na*</u> *wo*/*ni*/*women*

[40] In the immediate context of *don't*, no phonological properties would motivate the vowel reduction.

doi:10.6342/NTU201600328

jiu 'then I/you/we just' and *na wo xianzai* 'then I now' may be the reduction of *name* 'then'. Another clue comes from formulaic routines such as *tong yi shijian zaijian* 'see you again at the same time', which is usually used at the end of a TV program to invite the audience to stay tuned. These formulaic routines may be stored and processed as a whole because they almost always occur in the same intonation unit and allow little variation.

7.3 Practical Applications

The present study has identified a list of spoken and news bundles in Chinese. These bundles can be used to enrich existing language resources in Chinese, and they also serve as important references for language teachers and learners and have pedagogical implications.

There have been some large-scale lexical projects in Chinese, and lexical bundles as emergent units in the mental lexicon may need to be appropriately represented in those projects. The first example is the Chinese Wordnet, where a large number of lexical items in Chinese are associated through various lexical relations.[41] In the Chinese Wordnet, two lexical items can be coded as having a syntagmatic relation when they co-occur in a lexical bundle. In this way, the organization of the Chinese Wordnet will more faithfully reflect the usage-based proposal (Bybee 2007) that due to the pragmatic usefulness of their combination, the elements in a lexical bundle are closely associated in our memory. Note that words bearing a syntagmatic relation need to be distinguished from two-word collocations that usually do not have distinct discourse-level functions. Besides, when the new relation is defined more precisely, further manual interventions might be needed to exclude some lexical bundles that do not fit the definition.

[41] The Chinese Wordnet is available at http://lope.linguistics.ntu.edu.tw/cwn/.

doi:10.6342/NTU201600328

Another lexical resource the present study may contribute to is the DeepLex project (Hsieh 2015), which is a large data matrix collecting a wide array of lexical behaviors (120 variables in total) of approximately 30,000 lexical units (e.g., characters, words, chunks). Both the scope and the size are still evolving, and this project is calling for an open collaboration. The quantitative measures (e.g., the dispersion measure DP and the internal association measure G) in the present study may be potential variables to be represented for all the multi-word units in DeepLex. Besides, DeepLex takes the functional position in determining linguistic units, so lexical bundles that serve important functions and emerge as storage/processing units are candidate entries to be represented there. After incorporated into DeepLex, the data in the present study may be of great value for natural language processing in Chinese.

Recognizing the theoretical significance of lexical bundles can also direct more attention to their pedagogical implications. First, spoken bundles identified in the present study can help second language learners to develop native-like chunking strategies and therefore achieve greater fluency. By recognizing lexical bundles in Chinese, learners will more accurately determine unit boundaries to avoid awkward pauses; besides, learners will expand their repertoire of prefabricated chunks and deal with face-to-face conversations more easily. Second, teachers can draw more attention to communicative functions performed by lexical bundles (e.g., expressing stances, introducing a new topic) to facilitate learners' interaction with others and enhance the discourse organization of their production. Third, teachers can also explicitly address collocational patterns involved in lexical bundles. For example, the bundle *ju ge lizi* 'give an example' reveals that a high-frequency verbal collocate of the noun *lizi* 'example' is *ju* 'give', and the bundle *yi bi qian* 'one CLASSIFIER money' features the most common classifier of *qian* 'money'. Fourth, as lexical bundles are generally semantically transparent, their components may be seen as essential to language learners. After learning the characters and words commonly used in lexical bundles, learners can understand a large number of phrases through the principle of compositionality. Last but not least, the computational method of extracting word strings developed in this study can be applied to the identification of lexical bundles in learner language corpora of Chinese.

7.4 Limitations

The findings in the present study are subject to a number of limitations. First, the size of the conversation subcorpus may not be considered large enough for a study on lexical bundles, and many texts there are not from typical naturally-occurring conversations (see Section 3.1). Therefore, the findings may need to be interpreted with some caution. Second, since the identification of lexical bundles is word-based, the results of any study on lexical bundles in Chinese are potentially susceptible to the design of the word segmentation system adopted (see Section 3.2). Third, although many considerations have been involved in the setting of the quantitative thresholds adopted to filter out word sequences (see Section 3.3), the final decisions may be open to further improvement.

7.5 Suggestions for Future Research

The present study has identified spoken and news bundles in Chinese, and the identification method can be applied to other text types in Chinese, such as academic texts (e.g., journal articles, university textbooks) and online discussions. We will gain more insights into the role of lexical bundles in fulfilling the communicative needs of different text types. Additionally, a variationist perspective can be adopted to compare lexical bundles used in a wide range of Chinese-speaking communities.

Some aspects of lexical bundles have not been adequately addressed, even in English. An intriguing issue is the relationship between the position and the function of lexical bundles (Cortes and Csomay 2007, Ansari and Molavi 2013) (see Section 2.2). The present study has examined the general use of Chinese bundles in news, and a further study can focus on lexical bundles in the leads (i.e., the first one or two sentences) of news reports to investigate how lexical bundles are used to package the most important information of a news report and draw more attention from the reader.

Further psycholinguistic investigations are also needed to explore the storage and processing of lexical bundles in the mental lexicon, so that more converging evidence for the usage-based model (e.g., Bybee 2007) can accumulate. For example, the common lexical bundle *I don't know* is argued to be a storage/processing unit, and corpus data show that *I don't* occurs far more frequently than *don't know* (Bybee 2007). How the probabilistic analysis is run in the brain to closely associate the three words *I*, *don't*, and *know* remains unsolved (Tremblay et al. 2009, Gries 2013). Before we can arrive at a convincing answer, various comprehension and production experiments (e.g., grammaticality judgment tasks, gap-filling tasks, dictation tasks, self-paced reading tasks) would be necessary. Besides, it would be interesting to assess whether the frequency effect can predict the subjects' performance more reliably than other quantitative measures (e.g., dispersion measures, internal association measures).

There are other more practical aspects of lexical bundles. Although lexical bundles are seen as semantically transparent, many studies have revealed that there is a gap between native and non-native speakers' use of lexical bundles (see Section 2.2). The gap may arise from the fact that some lexical bundles have undergone semantic changes and developed non-compositional meanings. Therefore, more empirical

doi:10.6342/NTU201600328

studies are needed to investigate the role of semantic transparency in the acquisition of lexical bundles by Chinese second language learners. The results will have crucial pedagogical implications for which bundles need to be taught explicitly and how to teach lexical bundles. A corpus-based study investigating how lexical bundles in Chinese are translated would also be useful (Ji 2010). A Chinese bundle may be simply translated into its equivalent bundle in another language. Such an equivalent may not always be available, though; we may wonder what strategies expert translators usually adopt. The results of translation studies on lexical bundles will have important implications for the training of novice translators and the development of machine translation. Another intriguing direction is to adopt advanced techniques in the field of data science to examine the idiosyncratic use of lexical bundles on social networking sites. This may help to identify the linguistic patterns of different communities.

There are various kinds of multi-word expressions (see Section 2.1), and lexical bundles are just one of them. A better understanding of the relationship between lexical bundles and other kinds of multi-word units in Chinese needs to be developed. A lexical bundle may be a fragment of a longer lexico-syntactic frame. For instance, the lexical bundle *de guocheng zhong* 'DE process middle; (in) the process of' is part of the frame '*zai* + event + *de guocheng zhong*'. Some lexical bundles may be regarded as the actual realization of a more abstract construction. For instance, many referential expressions are derived from the construction 'number/demonstrative + classifier/quantifier + noun'. A close examination of each lexical bundle and a clear delineation of the relationship between lexical bundles and other types of multi-word units may herald an effective method to systematically identify constructions and lexico-syntactic frames in Chinese.

doi:10.6342/NTU201600328

References

Aijmer, Karin. 2008. “So er I just sort I dunno I think it’s just because…”: A corpus study of *I don’t know* and *dunno* in learners’ spoken English. In: Andreas H. Jucker, Daniel Schreier and Marianne Hundt (eds.), *Corpora: Pragmatics and Discourse*, 151-168. Amsterdam: Rodopi.

Aijmer, Karin and Altenberg, Bengt. 1996. Introduction. In: Karin Aijmer, Bengt Altenberg and Mats Johansson (eds.), *Languages in Contrast: Papers from a Symposium on Text-based Cross-linguistic Studies*, 11-16. Lund: Lund University Press.

Altenberg, Bengt. 1998. On the phraseology of spoken English: The evidence of recurrent word-combinations. In: A. P. Cowie (ed.), *Phraseology: Theory, Analysis, and Applications*, 101-122. Oxford: Oxford University Press.

Altenberg, Bengt and Eeg-Olofsson, Mats. 1990. Phraseology in spoken English: Presentation of a project. In: Jan Aarts and Willem Meijs (eds.), *Theory and Practice in Corpus Linguistics*, 1-26. Amsterdam: Rodopi.

Ansari, Siamak and Molavi, Arezoo. 2013. Textual positions of lexical bundles across newspaper genres. *International Research Journal of Applied and Basic Sciences* 4(9): 2484-2490.

Baayen, R. Harald. 2003. Probabilistic approaches to morphology. In Rens Bod, Jennifer Hay and Stefanie Jannedy (eds.), *Probabilistic Linguistics*, 229-288. Cambridge: MIT Press.

Baroni, Marco and Evert, Stefan. 2008. Statistical methods for corpus exploitation. In: Anke Lüdeling and Merja Kytö (eds.), *Corpus Linguistics: An International Handbook*, 777-803. Berlin: Mouton de Gruyter.

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus*

doi:10.6342/NTU201600328

Linguistics 14(3): 275-311.

Biber, Douglas and Barbieri, Federica. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26(3): 263-286.

Biber, Douglas, Conrad, Susan and Cortes, Viviana. 2003. Lexical bundles in speech and writing: An initial taxonomy. In: Andrew Wilson, Paul Rayson and Tony McEnery (eds.), *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech*, 71-92. Frankfurt am Main: Peter Lang.

Biber, Douglas, Conrad, Susan and Cortes, Viviana. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3): 371-405.

Biber, Douglas and Finegan, Edward. 1989. Styles of stance in English: Lexical and grammatical marking of evidentiality and affect. *Text* 9(1): 93-124.

Biber, Douglas, Johansson, Stig, Leech, Geoffrey, Conrad, Susan and Finegan, Edward. 1999. *Longman Grammar of Spoken and Written English*. London and New York: Longman.

Biq, Yung-O. 2004. Construction, reanalysis, and stance: 'V *yi ge* N' and variations in Mandarin Chinese. *Journal of Pragmatics* 36(9): 1637-1654.

Biq, Yung-O. 2015. Adverbs. In: William S-Y. Wang and Chaofen Sun (eds.), *The Oxford Handbook of Chinese Linguistics*, 414-428. Oxford: Oxford University Press.

Bolander, Maria. 1989. Prefabs, patterns and rules in interaction? Formulaic speech in adult learners' L2 Swedish. In: Kenneth Hyltenstam and Loraine K. Obler (eds.), *Bilingualism across the Lifespan: Aspect of Acquisition, Maturity and Loss*, 73-86. Cambridge: Cambridge University Press.

Bolinger, Dwight. 1961. Syntactic blends and other matters. *Language* 37(3): 366-381.

Bowers, Jeffrey S., Davis, Colin J. and Hanley, Derek A. 2005. Automatic semantic activation of embedded words: Is there a “hat” in “that”? *Journal of Memory and Language* 52(1): 131-143.

Brooks, Patricia and Tomasello, Michael. 1999. How children constrain their argument structure constructions. *Language* 75(4): 720-738.

Butler, Chris. 1997. Repeated word combinations in spoken and written text: Some implications for Functional Grammar. In: Chris Butler, John Connolly, Richard A. Gatward and Roel M. Vismans (eds.), *A Fund of Ideas: Recent Development in Functional Grammar*, 60-77. Amsterdam: Institute for Functional Research into Language and Language Use.

Bybee, Joan L. 1985. *Morphology: A Study of the Relation between Meaning and Form*. Amsterdam and Philadelphia: John Benjamins.

Bybee, Joan. 2007. *Frequency of Use and the Organization of Language*. Oxford: Oxford University Press.

Bybee, Joan L. and Scheibman, Joanne. 1999. The effect of usage on degrees of constituency: The reduction of *don't* in English. *Linguistics* 37(4): 575-596.

Cao, Juan and Zhou, Jing-ye. 2004. A new method for calculating relativity of Chinese strings. *Journal of Chinese Information Processing* 18(4): 55-59. (曹娟、周經野。2004。一種計算漢字串之間相關程度的新方法。《中文信息學報》，第 18 期第 4 卷，頁 55-59。)

Carroll, John B. 1970. An alternative to Juilland’s usage coefficient for lexical frequencies and a proposal for a standard frequency index. *Computer Studies in the Humanities and Verbal Behavior* 3(2): 61-65.

Chafe, Wallace L. 1968. Idiomaticity as an anomaly in the Chomskyan paradigm. *Foundations of Language* 4: 109-125.

Chafe, Wallace L. 1985. Linguistic differences produced by differences between speaking and writing. In: David R. Olson, Nancy Torrance and Angela Hildyard (eds.), *Literacy, Language, and Learning: The Nature and Consequences of Reading and Writing*, 105-123. Cambridge: Cambridge University Press.

Chen, Lin. 2010. An investigation of lexical bundles in ESP textbooks and electrical engineering introductory textbooks. In: David Wood (ed.), *Perspectives on Formulaic Language: Acquisition and Communication*, 107-125. London and New York: Continuum.

Chen, Ru and Zhu, Xiaoya. 2012. *Huayu Changyong Juxing yu Jiegou 330* ‘Common Chinese Patterns 330’. Taipei: Bookman. (陳如、朱曉亞。2012。《華語常用句型與結構 330》。臺北：書林。)

Chen, Yu-Hua and Baker, Paul. 2010. Lexical bundles in L1 and L2 academic writing. *Language Learning & Technology* 14(2): 30-49.

Conklin, Kathy and Schmitt, Norbert. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29(1): 72-89.

Conrad, Susan, and Biber, Douglas. 2004. The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica* 20: 56-71.

Cortes, Viviana. 2002. Lexical bundles in freshman composition. In: Randi Reppen, Susan M. Fitzmaurice and Douglas Biber (eds.), *Using Corpora to Explore Linguistic Variation*, 131-145. Amsterdam and Philadelphia: John Benjamins.

Cortes, Viviana. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23(4): 397-423.

Cortes, Viviana. 2008. A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora* 3(1): 43-57.

Cortes, Viviana and Csomay, Eniko. 2007. Positioning lexical bundles in university

doi:10.6342/NTU201600328

lectures. In: Mari Carmen Campoy and María José Luzón (eds.), *Spoken Corpora in Applied Linguistics*, 57-76. Frankfurt am Main: Peter Lang.

Coulmas, Florian. 1981. *Conversational Routine*. The Hague: Mouton.

Cowie, A. P. 1981. The treatment of collocations and idioms in learners' dictionaries. *Applied Linguistics* 2(3): 223-235.

Cowie, A. P. 1994. Phraseology. In: Ronald E. Asher (ed.), *The Encyclopedia of Language and Linguistics*, 3168-3171. Oxford: Pergamon.

Cowie, A. P. 1998. Phraseological dictionaries: Some east-west comparisons. In: A. P. Cowie (ed.), *Phraseology: Theory, Analysis, and Applications*, 209-228. Oxford: Oxford University Press.

Coxhead, Averil. 2000. A new academic word list. *TESOL Quarterly* 34(2): 213-238.

Crystal, David. 2008. *A Dictionary of Linguistics and Phonetics* (the sixth edition). Oxford: Blackwell.

Cui, Xiliang. 2006. *Hanyu Shouyu yu Zhongguo Renwen Shijie* 'Chinese Idioms and Chinese Humanistic Inner World'. Beijing: Beijing Language and Culture University Press. (崔希亮。2006。《漢語熟語與中國人文世界》。北京：北京語言大學出版社。)

Culpeper, Jonathan and Kytö, Merja. 2002. Lexical bundles in Early Modern English dialogues: A window into the speech-related language of the past. In: Teresa Fanego, Belén Méndez-Naya and Elena Seoane (eds.), *Sounds, Words, Texts, and Change*, 45-63. Amsterdam and Philadelphia: John Benjamins.

De Cock, Sylvie. 1998. A recurrent word combination approach to the study of formulae in the speech of native and non-native speakers of English. *International Journal of Corpus Linguistics* 3(1): 59-80.

Dolch, E. W. 1936. A basic sight vocabulary. *The Elementary School Journal* 36(6): 456-460.

doi:10.6342/NTU201600328

Dooley, Robert A. and Levinsohn, Stephen H. 2001. *Analyzing Discourse: A Manual of Basic Concept*. SIL International.

Ellis, Nick C. 1996. Sequencing in SLA: Phonological memory, chunking, and points of order. *Studies in Second Language Acquisition* 18(1): 91-126.

Ellis, Nick C. 1998. Emergentism, connectionism and language learning. *Language Learning* 48(4): 631-664.

Ellis, Nick C. 2002. Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition* 24(2): 143-188.

Ellis, Nick C. 2003. Construction, chunking, and connectionism: The emergence of second language structure. In: Catherine J. Doughty and Michael H. Long (eds.), *The Handbook of Second Language Acquisition*, 63-103. Oxford: Blackwell.

Ellis, Nick C., Simpson-Vlach, Rita and Maynard, Carson. 2008. Formulaic language in native and second language speakers: Psycholinguistics, corpus linguistics, and TESOL. *TESOL Quarterly* 42(3): 375-396.

Erman, Britt and Warren, Beatrice. 2000. The idiom principle and the open choice principle. *Text & Talk* 20(1): 29-62.

Feng, Shengli. 1997. Prosodically determined word-formation in Mandarin Chinese. *Social Science in China* 18(4): 120-137.

Fillmore, Charles J. 1978. On the organization of semantic information in the lexicon. In: Donka Farkas, Wesley M. Jacobsen and Karol W. Todrys (eds.), *Papers from the Parasession on the Lexicon*, 148-173. Chicago: Chicago Linguistic Society.

Firth, John R. 1957. *Papers in Linguistics 1934-1951*. Oxford: Oxford University Press.

Flowerdew, Lynne. 2012. *Corpora and Language Education*. New York: Palgrave Macmillan.

doi:10.6342/NTU201600328

Foster, Pauline. 2001. Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In: Martin Bygate, Peter Skehan and Merrill Swain (eds.), *Researching Pedagogic Tasks: Second Language Learning, Teaching, and Testing*, 75-93. London and New York: Longman.

Fu, Huaiqing. 1985. *Xiandai Hanyu Cihui* ‘Modern Chinese Lexicon’. Beijing: Peking University Press. (符淮青。1985。《現代漢語詞匯》。北京：北京大學出版社。)

Goldberg, Adele E., Casenhiser, Devin and Sethuraman, Nitya. 2004. Learning argument structure generalizations. *Cognitive Linguistics* 14(3): 289-316.

Granger, Sylviane and Paquot, Magali. 2008. Disentangling the phraseological web. In: Sylviane Granger and Fanny Meunier (eds.), *Phraseology: An Interdisciplinary Perspective*, 27-50. Amsterdam and Philadelphia: John Benjamins.

Gries, Stefan Th. 2008a. Phraseology and linguistic theory: A brief survey. In: Sylviane Granger and Fanny Meunier (eds.), *Phraseology: An Interdisciplinary Perspective*, 3-25. Amsterdam and Philadelphia: John Benjamins.

Gries, Stefan Th. 2008b. Dispersion and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4): 403-437.

Gries, Stefan Th. 2009. *Quantitative Corpus Linguistics with R: A Practical Introduction*. London: Routledge.

Gries, Stefan Th. 2013. 50-something years of work on collocations: What is or should be next… *International Journal of Corpus Linguistics* 18(1): 137-165.

Gries, Stefan Th. 2014. Frequency tables: Tests, effect sizes, and explorations. In: Dylan Glynn and Justyna A. Robinson (eds.), *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*, 365-389. Amsterdam and

doi:10.6342/NTU201600328

Philadelphia: John Benjamins.

Han, Ke-song, Wang, Yong-cheng and Chen, Gui-lin. 2000. Research on fast high-frequency strings extracting and statistics algorithm with no thesaurus. *Journal of Chinese Information Processing* 15(2): 23-30. (韓客松、王永成、陳桂林。2000。無詞典高頻字串快速提取和統計算法研究。《中文信息學報》，第 15 期第 2 卷，頁 23-30。)

Hoey, Michael and O’Donnell, Matthew Brook. 2008. Lexicography, grammar, and textual position. *Journal of Lexicography* 21(3): 293-309.

Hopper, Paul J. 1991. On some principles of grammaticalization. In: Elizabeth C. Traugott and Bernd Heine (eds.), *Approaches to Grammaticalization*, 17-35. Amsterdam: John Benjamins.

Howarth, Peter. 1998. The phraseology of learners’ academic writing. In: A. P. Cowie (ed.), *Phraseology: Theory, Analysis, and Applications*, 161-186. Oxford: Oxford University Press.

Hsieh, Shu-Kai. 2015. DeepLEX: A multi-dimensional Chinese lexical resource. Paper presented at the CKIP 30th Anniversary Workshop on Chinese Lexicon.

Hu, Liang and Tang, Xuri. 2014. Multiword expression extraction based on word relativity. *International Journal of Knowledge and Language Processing* 5(1): 27-40.

Huang, Borong and Liao, Xudong. 2002. *Xiandai Hanyu* ‘Modern Chinese’ (the third edition). Beijing: Higher Education Press. (黃伯榮、廖序東。2002。《現代漢語》增訂三版。北京：高等教育出版社。)

Huang, C.-T. James, Li, Y.-H. Audrey and Li, Yafei. 2009. *The Syntax of Chinese*. Cambridge: Cambridge University Press.

Hunston, Susan. 2002. *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

doi:10.6342/NTU201600328

Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27(1): 4-21.

Hyland, Ken. 2012. Bundles in academic discourse. *Annual Review of Applied Linguistics* 32: 150-169.

Ji, Meng. 2010. *Phraseology in Corpus-based Translation Studies*. Frankfurt am Main: Peter Lang.

Jiang, Nan and Nekrasova, Tatiana M. 2007. The processing of formulaic sequences by second language speakers. *The Modern Language Journal* 91(3): 433-445.

Johansson, Stig. 2007. *Seeing Through Multilingual Corpora: On the Use of Corpora in Contrastive Studies*. Amsterdam and Philadelphia: John Benjamins.

Johnstone, Barbara. 2002. *Discourse Analysis*. Malden: Blackwell.

Kim, YouJin. 2009. Korean lexical bundles in conversation and academic texts. *Corpora* 4(2): 135-165.

Kopaczyk, Joanna. 2012. Applications of the lexical bundles method in historical corpus research. In: Piotr Pęzik (ed.), *Corpus Data Across Languages and Disciplines*, 83-95. Frankfurt am Main: Peter Lang.

Kopaczyk, Joanna. 2013. *The Legal Language of Scottish Burghs: Standardization and Lexical Bundles 1380-1560*. Oxford: Oxford University Press.

Lambrecht, Knud. 1984. Formulaicity, frame semantics, and pragmatics in German binomial expressions. *Language* 60(4): 753-796.

Leńko-Szymańska, Agnieszka. 2014. The acquisition of formulaic language by EFL learners: A cross-sectional and cross-linguistic perspective. *International Journal of Corpus Linguistics* 19(2): 225-251.

Li, Charles N. and Thompson, Sandra A. 1981. *Mandarin Chinese: A Functional Reference Grammar*. Berkeley: University of California Press.

Li, Ming-Yi. 2014. The use of formulaic discourse markers in CSL learners' writing

doi:10.6342/NTU201600328

from the aspect of textual function. *Journal of Chinese Language Teaching* 11(2): 31-57. (李明懿。2014。語篇功能視角之華語非母語學習者寫作使用定式篇章標記語分析。《華語文教學研究》，第 11 期第 2 卷，頁 31-57。)

Libben, Gary. 2005. Everything is psycholinguistic: Material and methodological considerations in the study of compound processing. *Canadian Journal of Linguistics* 50: 267-283.

Lien, Chinfa. 1989. Antonymous quadrinominals in Chinese. *Journal of Chinese Linguistics* 17(2): 263-306.

Lien, Chinfa. 2000. Goucixue wenti tansuo 'Exploring morphological issues'. *Chinese Studies* 18: 61-78. (連金發。2000。構詞學問題探索。《漢學研究》，第 18 期，頁 61-78。)

Lu, Shuxiang. 1979. *Hanyu Yufa Fenxi Wenti* 'Issues about Chinese Grammar'. Beijing: The Commercial Press. (呂叔湘。1979。《漢語語法分析問題》。北京：商務印書館。)

Makkai, Adam. 1972. *Idiom Structure in English*. The Hague: Mouton.

Manes, Joan and Wolfson, Nessa. 1981. The compliment formula. In: Florian Coulmas (ed.), *Conversational Routine*, 115-132. The Hague: Mouton.

McCarthy, Michael. 2001. *Issues in Applied Linguistics*. Cambridge: Cambridge University Press.

McEnery, Tony and Hardie, Andrew. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press.

McEnery, Tony, Xiao, Richard and Tono, Yukio. 2006. *Corpus-based Language Studies*. London: Routledge.

Mel'čuk, Igor. 1998. Collocations and lexical functions. In: A. P. Cowie (ed.), *Phraseology: Theory, Analysis, and Applications*, 23-53. Oxford: Oxford University Press.

doi:10.6342/NTU201600328

Meunier, Fanny. 2012. Formulaic language and language teaching. *Annual Review of Applied Linguistics* 32: 111-129.

Nattinger, James. 1988. Some current trends in vocabulary teaching. In: Ronald Carter and Michael McCarthy (eds.), *Vocabulary and Language Teaching*, 62-82. London and New York: Longman.

Nekrasova, Tatiana M. 2009. English L1 and L2 speakers' knowledge of lexical bundles. *Language Learning* 59(3): 647-686.

Nesi, Hilary and Basturkmen, Helen. 2006. Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics* 11(3): 283-304.

O'Donnell, Matthew Brook, Römer, Ute and Ellis, Nick C. 2013. The development of formulaic sequences in first and second language writing. *International Journal of Corpus Linguistics* 18(1): 83-108.

O'Keeffe, Anne, McCarthy, Michael and Carter, Ronald. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Oppenheim, Nancy. 2000. The importance of recurrent sequences for nonnative speaker fluency and cognition. In: Heidi Riggenbach (ed.), *Perspectives on Fluency*, 220-240. Ann Arbor: University of Michigan Press.

Partington, Alan. 1998. *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam and Philadelphia: John Benjamins.

Partington, Alan and Morley, John. 2004. From frequency to ideology: Investigating word and cluster/bundle frequency in political debate. In: Barbara Lewandowska-Tomaszczyk (ed.), *Practical Applications in Language and Computers*, 179-192. Frankfurt am Main: Peter Lang.

Pawley, Andrew and Syder, Frances Hodgetts. 1983. Two puzzles for linguistic theory:

doi:10.6342/NTU201600328

Nativelike selection and nativelike fluency. In: Jack C. Richards and Richard W. Schmidt (eds.), *Language and Communication*, 191-225. London and New York: Longman.

Saffran, Jennifer R., Aslin, Richard N. and Newport, Elissa L. 1996. Statistical learning by 8-month-old infants. *Science* 274: 1926-1928.

Saffran, Jennifer R. and Wilson, Diana P. 2003. From syllabus to syntax: Multilevel statistical learning by 12-month old infants. *Infancy* 4(2): 273-284.

Salazar, Danica. 2014. *Lexical Bundles in Native and Non-native Scientific Writing: Applying a Corpus-based Study to Language Teaching*. Amsterdam and Philadelphia: John Benjamins.

Scheibman, Joanne. 2000. *I dunno but*: A usage-based account of the phonological reduction of *don't* in American English conversation. *Journal of Pragmatics* 32(1): 105-124.

Schmitt, Norbert, Grandage, Sarah and Adolphs, Svenja. 2004. Are corpus-derived recurrent clusters psycholinguistically valid? In: Norbert Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use*, 127-151. Amsterdam and Philadelphia: John Benjamins.

Shrefler, Nathan. 2011. Lexical bundles and German bibles. *Literary and Linguistic Computing* 26(1): 89-106.

Simpson, Rita C. 2004. Stylistic features of academic speech: The role of formulaic expressions. In: Ulla Connor and Thomas A. Upton (eds.), *Discourse in the Professions: Perspectives from Corpus Linguistics*, 37-64. Amsterdam and Philadelphia: John Benjamins.

Simpson-Vlach, Rita and Ellis, Nick C. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4): 487-512.

Sinclair, John. 2004. *Trust the Text: Language, Corpus and Discourse*. London:

doi:10.6342/NTU201600328

Routledge.

Sorhus, Helen B. 1977. To hear ourselves—Implications for teaching English as a second language. *English Language Teaching Journal* 31(3): 211-221.

Spöttl, Carol and McCarthy, Michael. 2004. Comparing knowledge of formulaic sequences across L1, L2, L3, and L4. In: Norbert Schmitt (ed.), *Formulaic Sequences: Acquisition, Processing and Use*, 191-225. Amsterdam and Philadelphia: John Benjamins.

Stubbs, Michael. 1996. *Text and Corpus Analysis*. Oxford: Blackwell.

Stubbs, Michael. 2007. Quantitative data on multi-word sequences in English: The case of the word world. In: Michael Hoey, Michaela Mahlberg, Michael Stubbs and Wolfgang Teubert (eds.), *Text, Discourse and Corpora: Theory and Analysis*, 163-189. London and New York: Continuum.

Swales, John. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press.

Swales, John. 2008. Foreword. In: Diane Belcher and Alan Hirvela (eds.), *The Oral/Literate Connection: Perspectives on L2 Speaking, Writing, and Other Media Interactions*, v-viii. Ann Arbor: University of Michigan Press.

Swinney, David A. and Cutler, Anne. 1979. The access and processing of idiomatic expressions. *Journal of Verbal Learning and Verbal Behavior* 18(5): 523-534.

Tannen, Deborah. 1982. Oral and literate strategies in spoken and written narratives. *Language* 58: 1-21.

Tannen, Deborah. 1987. Repetition in conversation as spontaneous formulaicity. *Text & Talk* 7(3): 215-243.

Tao, Hongyin. 2015. Profiling the Mandarin spoken vocabulary based on corpora. In: William S-Y. Wang and Chaofen Sun (eds.), *The Oxford Handbook of Chinese Linguistics*, 336-347. Oxford: Oxford University Press.

doi:10.6342/NTU201600328

Thompson, Geoff. 1996. *Introducing Functional Grammar*. London: Arnold.

Tracy-Ventura, Nicole, Cortes, Viviana and Biber, Douglas. 2007. Lexical bundles in speech and writing. In: Giovanni Parodi (ed.), *Working with Spanish Corpora*, 217-231. London and New York: Continuum.

Traugott, Elizabeth C. and Dasher, Richard B. 2005. *Regularity in Semantic Change*. Cambridge: Cambridge University Press.

Tremblay, Antoine, Derwing, Bruce and Libben, Gary. 2009. Are lexical bundles stored and processed as single units? *Working Papers of the Linguistics Circle of the University of Victoria* 19: 258-279.

Van Lancker-Sidtis, Diana and Rallon, Gail. 2004. Tracking the incidence of formulaic expressions in everyday speech: Methods for classification and verification. *Language & Communication* 24(3): 207-240.

Wang, Jihui. 2009. *Gudingyu Yangjiu* 'A study on idiomatic expressions'. Tianjin: Nankai University Press. (王吉輝。2009。《固定語研究》。天津：南開大學出版社。)

Wang, Li. 1990. Ci yu weiyu de jiexian wenti 'Issues about the demarcation between words and phrases'. In: Li Wang (ed.), *Wang Li Wenji* 'Essays by Wang Li' (vol. 16), 236-253. Shandong: Shandong Education Press. (王力。1990。詞與偽語的界限問題。《王力文集》第十六卷，頁 236-253。山東：山東教育出版社。)

Wang, Yu-Chuan. 1987. *Guoyu Sanbei ge Juxing* 'Three Hundred Patterns in Chinese'. Taipei: Mandarin Daily News. (王玉川。1987。《國語三百個句型》。臺北：國語日報社。)

Wei, Naixing and Li, Jingjie. 2013. A new computing method for extracting contiguous phraseological sequences from academic text corpora. *International Journal of Corpus Linguistics* 18(4): 506-535.

Wen, Juan and Wang, Xiao-jie. 2009. Chinese frequent string extraction and application on language model. *Journal of Beijing University of Posts and Telecommunications* 32(5): 10-14. (文娟、王小捷。2009。中文高頻詞串的抽取及其在語言模型中的應用。《北京電郵大學學報》，第 32 期第 5 卷，頁 10-14。)

Wiechmann, Daniel. 2008. On the computation of collostruction strength: Testing measures of association as expressions of lexical bias. *Corpus Linguistics and Linguistic Theory* 4(2): 253-290.

Wood, David. 2010. Lexical clusters in an EAP textbook corpus. In: David Wood (ed.), *Perspectives on Formulaic Language: Acquisition and Communication*, 88-106. London and New York: Continuum.

Wray, Alison. 2002. *Formulaic Language and the Lexicon*. Cambridge: Cambridge University Press.

Wray, Alison. 2013. Formulaic language. *Language Teaching* 46(3): 316-334.

Yang, Jianguo. 2009. *Jiyu Dongtai Liutong Yuliaoku de Hanyu Shouyu Danwei Yanjiu* ‘Studies on Idiom Unit of Chinese based on the Dynamic Circulating Corpus’. Beijing: Beijing Language and Culture University Press. (楊建國。2009。《基於動態流通語料庫的漢語熟語單位研究》。北京：北京語言大學出版社。)

Zhou, Guangqing. 1994. Chengyu neibu xingshi lun ‘A study on the internal structure of idiomatic expressions’. *Journal of Huazhong Normal University: Humanities and Social Sciences* 1994(5): 112-117. (周光慶。1994。成語內部形式論。《華中師範大學學報：哲學社會科學版》，1994 年第 5 期，頁 112-117。)

Zhou, Jian. 1997. Lun siziyu han sanziyu ‘On four-character and three-character expressions’. *Linguistic Researches* 65: 26-31. (周荐。1997。論四字語和三字語。《語文研究》，第 65 期，頁 26-31。)

Appendix: Lexical Bundles Identified in the Present Study

(A) three-word lexical bundles in conversation

Bundle	Relative frequency	Text count	DP	G
是 不 是	1317.87	93	0.106	3.599522
是 一 個	802.4652	93	0.1695	1.899433
有 一 個	745.923	95	0.067	2.856595
一 個 人	571.9468	73	0.152	3.551967
對 不 對	548.025	76	0.176	6.764605
每 一 個	315.3319	59	0.214	5.049246
一 個 很	306.6331	63	0.235	2.560643
是 一 種	289.2354	63	0.165	2.849701
在 這 個	267.4884	65	0.1425	1.734752
什麼 樣 的	234.8679	52	0.312	4.380474
我 想 這	221.8197	57	0.2905	3.025835
那 個 時候	215.2956	40	0.3175	3.997375
我 是 覺得	213.1209	43	0.2855	4.757862
不 是 說	208.7714	57	0.1895	2.575673
我 覺得 這	204.422	51	0.3115	2.185203
我 覺得 我	202.2473	50	0.211	1.319303
所以 我 覺得	197.8979	47	0.322	4.551234
的 一 種	193.5485	41	0.236	1.945624
個 人 都	187.0244	48	0.139	4.518808
有 這 種	184.8497	57	0.1365	2.892089
的 那 種	184.8497	52	0.1425	2.270509
的 這 種	184.8497	44	0.1665	1.640107
很 大 的	182.675	44	0.1685	3.935814
我 不 知道	180.5003	43	0.23	2.909612
有 一 種	178.3256	49	0.1805	3.214166
很 好 的	178.3256	50	0.249	2.75789
這 個 問題	178.3256	48	0.2265	4.309642
你 這 個	176.1509	39	0.2135	0.619761
做 一 個	173.9762	47	0.173	3.312976
並 不 是	169.6268	42	0.2005	4.211729
這 是 一	165.2774	47	0.193	2.774449
每 個 人	158.7533	40	0.189	6.141954
有 一 次	152.2292	41	0.2665	4.32191
位 特別 來賓	152.2292	33	0.36	8.464061

這 個 時候	152.2292	37	0.2265	2.158919
最 重要 的	152.2292	39	0.2955	4.02494
因為 我 覺得	147.8798	34	0.37	4.310944
好 不 好	145.7051	37	0.2605	4.566226
這 個 樣子	145.7051	39	0.2475	3.511911
一 句 話	141.3557	33	0.3575	4.657645
對 對 對	139.181	14	0.61	6.423861
還 有 一	139.181	49	0.201	2.739026
把 這 個	134.8316	35	0.209	3.39749
所以 我 想	134.8316	39	0.3245	4.614536
另外 一 個	132.6569	37	0.3055	4.961253
不 是 很	130.4822	40	0.149	2.281679
也 不 是	128.3075	42	0.191	2.556455
他 這 個	126.1327	30	0.2875	0.061484
我們 這 個	126.1327	39	0.1885	0.893923
那 我 就	126.1327	33	0.3675	2.778119
個 就 是	123.958	36	0.2505	0.630182
就 好 了	121.7833	31	0.27	4.974531
不 好 的	119.6086	31	0.313	1.961095
我 也 是	117.4339	28	0.2305	1.612589
在 家 裡	115.2592	41	0.1845	4.73315
有 這樣 的	115.2592	32	0.1835	3.616335
這 個 社會	115.2592	35	0.2515	4.782383
覺得 這 個	115.2592	38	0.299	1.754066
很 重要 的	113.0845	33	0.2245	3.75272
就 是 要	113.0845	37	0.16	1.575331
我 就 覺得	110.9098	29	0.3185	3.791927
得 很 好	110.9098	38	0.244	4.608196
這 個 人	110.9098	33	0.2715	0.858891
所以 這 個	108.7351	25	0.2675	1.947653
一 個 問題	106.5604	32	0.225	3.686151
是 我 的	106.5604	39	0.251	1.330862
個 人 的	106.5604	31	0.278	1.015104
在 那 個	100.0363	31	0.221	1.584159
可是 我 覺得	97.86161	24	0.282	4.579935
我 也 不	97.86161	36	0.186	2.574829
跟 我 講	97.86161	26	0.3445	5.905146
有 很多 的	95.68691	33	0.2355	2.584906

doi:10.6342/NTU201600328

我	想	我	93.51221	42	0.2575	0.835383
那	我	覺得	93.51221	31	0.3	1.426246
這	一	點	93.51221	28	0.251	4.539219
一	個	月	91.33751	14	0.5495	4.498911
我	就	會	91.33751	25	0.507	2.056979
我	覺得	很	91.33751	28	0.38	2.210868
這	個	東西	91.33751	24	0.3095	3.598983
然後	我	就	91.33751	25	0.3905	4.120525
的	那	一	89.1628	33	0.2985	2.254629
最	好	的	86.9881	26	0.2485	4.095426
那	我	想	84.8134	32	0.342	2.012318
這	件	事情	84.8134	20	0.365	4.592322
一	件	事情	82.6387	29	0.2575	4.292301
這	是	我	82.6387	39	0.256	1.744165
這	種	情況	82.6387	21	0.282	4.674079
也	就	是	80.46399	27	0.266	1.822729
不	一樣	的	80.46399	26	0.294	2.179229
他	那	個	80.46399	25	0.301	0.725788
有	一	天	80.46399	24	0.346	3.925807
有	些	人	80.46399	14	0.494	4.705127
而	不	是	80.46399	25	0.355	3.66365
也	是	很	78.28929	23	0.2735	2.943034
我	覺得	那	78.28929	25	0.2885	1.165979
那	種	感覺	78.28929	24	0.233	5.756191
家	裡	的	78.28929	19	0.3525	2.035028
他	就	說	76.11459	20	0.3485	3.287766
在	一	個	76.11459	28	0.232	0.148812
有	沒	有	76.11459	19	0.2245	5.676584
我	自己	的	76.11459	26	0.3225	1.677345
的	時候	呢	76.11459	18	0.4045	3.120436
是	他	的	76.11459	29	0.2175	0.933152
這	也	是	76.11459	25	0.1555	1.696996
不	是	一	73.93989	30	0.2305	0.058837
太	好	了	73.93989	24	0.271	6.068297
我	覺得	他	73.93989	23	0.5685	1.043057
會	覺得	說	73.93989	24	0.4765	4.603977
一	大	堆	71.76518	22	0.2315	5.365255
因為	這	個	71.76518	29	0.2955	1.47105

doi:10.6342/NTU201600328

你 自己 的	71.76518	22	0.3275	2.653282
其實 我 覺得	71.76518	24	0.308	4.829939
在 這 種	69.59048	23	0.4235	2.351712
就 是 這樣	69.59048	30	0.3485	2.395169
我 覺得 我們	67.41578	21	0.3655	1.692573
所以 我 就	67.41578	23	0.2925	2.316574
是 一樣 的	67.41578	23	0.462	3.219846
是 你 的	67.41578	27	0.1935	1.056986
是 為 了	67.41578	22	0.2585	2.950017
現在 這 個	67.41578	23	0.301	1.96739
他 自己 的	65.24108	19	0.338	2.54074
有 一 位	65.24108	21	0.4525	3.656811
我 要 去	65.24108	14	0.384	2.587881
我 就 很	65.24108	21	0.349	2.311435
我們 來 聽聽	65.24108	21	0.392	6.18833
每 次 都	65.24108	17	0.378	5.03659
最 大 的	65.24108	22	0.2075	4.30326
跟 他 講	65.24108	20	0.4805	5.70017
把 那 個	63.06637	18	0.38	3.601574
很多 人 都	63.06637	24	0.339	5.297021
跟 我 說	63.06637	17	0.5	3.499511
說 得 好	63.06637	18	0.55	5.516048
有 人 說	60.89167	23	0.2785	4.390139
有 那 種	60.89167	25	0.238	1.878388
你 就 是	60.89167	19	0.315	0.293263
我們 來 看看	60.89167	19	0.4875	6.111505
每 天 都	60.89167	19	0.34	5.136998
這 種 事情	60.89167	19	0.2965	5.135673
都 是 一	60.89167	31	0.2205	1.077285
我 就 說	58.71697	15	0.431	1.909729
我 覺得 說	58.71697	19	0.423	1.197575
或者 是 說	58.71697	21	0.434	3.236736
就 是 因為	58.71697	20	0.375	2.340947
就 是 很	58.71697	24	0.27	0.930742
一 個 非常	56.54227	16	0.41	2.796679
不 太 一樣	56.54227	21	0.26	5.84771
他 也 是	56.54227	17	0.3765	1.722333
只 有 一	56.54227	21	0.2895	3.5986

doi:10.6342/NTU201600328

你	那	個	56.54227	19	0.35	0.249175
你	知道	嗎	56.54227	13	0.477	6.04656
我們	兩	個	56.54227	17	0.312	3.238797
我們	都	是	56.54227	23	0.598	2.398849
那	就	是	56.54227	21	0.31	0.145758
是	有	一	56.54227	41	0.217	0.176576
這	個	地方	56.54227	17	0.331	3.954142
這	個	孩子	56.54227	16	0.516	1.395581
都	不	是	56.54227	20	0.338	1.397941
不	是	那麼	54.36756	23	0.3585	3.664857
有	幾	個	54.36756	16	0.4395	3.983387
我	在	想	54.36756	22	0.4345	4.266112
的	時候	就	54.36756	18	0.4775	0.444809
是	那	種	54.36756	21	0.3675	0.759244
這	句	話	54.36756	18	0.4315	3.469518
都	不	知道	54.36756	21	0.3145	3.617135
小	的	時候	52.19286	17	0.41	2.427262
他	就	會	52.19286	20	0.423	2.386453
他	說	他	52.19286	16	0.2995	2.487653
在	我	的	52.19286	18	0.5555	1.955483
我	跟	我	52.19286	15	0.3625	1.686318
我	覺得	你	52.19286	22	0.3105	0.561884
我	覺得	應該	52.19286	14	0.4465	4.134915
我們	一起	來	52.19286	19	0.548	5.364672
沒有	這	個	52.19286	19	0.388	0.6331
那	個	人	52.19286	17	0.3275	1.012339
哪	一	個	52.19286	19	0.3325	3.092476
做	的	事情	52.19286	16	0.4465	4.770146
就	是	說	52.19286	17	0.4445	0.349364
人	都	是	50.01816	24	0.275	1.961142
不	知道	是	50.01816	17	0.3305	0.944741
我	想	我們	50.01816	17	0.4315	1.945145
我	跟	他	50.01816	20	0.393	2.006962
我	說	我	50.01816	12	0.48	1.570102
某	一	個	50.01816	16	0.309	4.856539
這	個	世界	50.01816	18	0.422	5.107114
這	個	事情	50.01816	17	0.2485	2.254491
這	個	節目	50.01816	14	0.3905	4.619278

這樣	一	個	50.01816	18	0.3085	1.085816
變成	一	個	50.01816	20	0.484	4.457711
但是	我	覺得	47.84346	18	0.293	3.734342
是	一	件	47.84346	19	0.3025	2.862796
是	什麼	呢	47.84346	13	0.543	4.226566
這	是	很	47.84346	19	0.212	1.995106
一	個	地方	45.66875	16	0.381	3.650668
也	是	一樣	45.66875	13	0.336	5.097466
在	他	的	45.66875	19	0.3165	1.961745
有	很多	人	45.66875	17	0.5035	3.283663
但是	這	個	45.66875	16	0.3835	1.878212
你	的	孩子	45.66875	12	0.635	3.305776
你	會	覺得	45.66875	18	0.344	3.902263
你們	兩	個	45.66875	14	0.403	5.419597
我	是	說	45.66875	12	0.411	1.249981
那	個	地方	45.66875	16	0.3725	4.426144
那	時候	我	45.66875	16	0.3575	1.684727
來	講	的話	45.66875	16	0.406	5.707867
的	這	一	45.66875	16	0.359	0.606921
看	這	個	45.66875	19	0.329	0.244612
從	這	個	45.66875	17	0.2845	2.186684
都	會	有	45.66875	19	0.3795	2.909876
談	一	談	45.66875	11	0.417	10.32703
一	個	朋友	43.49405	13	0.4015	2.278392
一	個	禮拜	43.49405	12	0.3715	4.361987
不	太	好	43.49405	22	0.4255	3.70295
什麼	都	不	43.49405	23	0.2825	3.90175
今天	非常	謝謝	43.49405	19	0.4245	8.739988
我	每	次	43.49405	14	0.3605	2.088553
我	就	跟	43.49405	14	0.5555	2.73053
我們	再	來	43.49405	17	0.5145	5.596366
找	一	個	43.49405	15	0.2735	3.741132
那	你	就	43.49405	13	0.2535	2.489689
的	時候	啊	43.49405	9	0.6405	1.770806
這	位	同學	43.49405	11	0.5685	4.032036
就	跟	他	43.49405	17	0.4395	2.82424
一	位	要	41.31935	13	0.622	3.844208
一	個	同學	41.31935	15	0.388	2.818326

一 個 家庭	41.31935	11	0.629	4.305171
也 是 這樣	41.31935	16	0.4015	3.235048
不 一樣 了	41.31935	11	0.372	3.278695
今天 這 個	41.31935	13	0.5205	2.213488
他 有 一	41.31935	23	0.2625	1.027088
在 我們 的	41.31935	17	0.4295	2.574893
我 不 曉得	41.31935	15	0.2835	2.895913
我 媽 就	41.31935	11	0.528	3.737177
的 人 都	41.31935	12	0.329	1.668126
是 什麼 樣	41.31935	17	0.5055	2.301253
就 會 有	41.31935	15	0.3775	1.753768
然後 他 就	41.31935	17	0.308	3.913823
想 一 想	41.31935	14	0.4565	7.265442
當 一 個	41.31935	12	0.459	1.208905
跟 他 說	41.31935	15	0.4435	3.249957
謝謝 各 位	41.31935	14	0.52	7.235752
一 種 很	39.14465	16	0.228	2.014836
也 不 敢	39.14465	12	0.474	5.349461
不 是 啦	39.14465	9	0.515	3.036031
他 說 我	39.14465	19	0.4155	0.869999
但是 我 想	39.14465	16	0.275	4.111892
你 就 會	39.14465	16	0.424	2.045357
我 想 他	39.14465	15	0.4595	0.740589
每 一 次	39.14465	13	0.348	5.917298
的 一 點	39.14465	12	0.438	2.070248
非常 重要 的	39.14465	12	0.455	3.804457
前 三 名	39.14465	6	0.624	9.315053
是 這樣子 的	39.14465	17	0.3125	2.190017
是 對 的	39.14465	12	0.409	3.082963
時候 我 就	39.14465	14	0.423	2.746711
這樣 的 事情	39.14465	10	0.3165	4.139447
通常 都 是	39.14465	13	0.4115	4.448902
都 不 一樣	39.14465	15	0.404	3.826445
都 是 在	39.14465	17	0.315	1.541918
滿 好 的	39.14465	16	0.4495	4.483591
覺得 很 奇怪	39.14465	14	0.3185	6.252401
他 講 說	36.96994	13	0.5715	3.312111
他們 兩 個	36.96994	14	0.4045	4.607201

另外 一 種	36.96994	15	0.363	5.488646
可以 說 是	36.96994	14	0.348	3.623824
各 位 同學	36.96994	9	0.5015	7.143494
你 要 去	36.96994	14	0.3595	3.023492
我 那 時候	36.96994	11	0.486	1.291117
我 很 喜歡	36.96994	14	0.446	3.017707
我 是 選擇	36.96994	12	0.6155	3.631713
我 記得 我	36.96994	14	0.328	3.341611
我 就 把	36.96994	11	0.308	3.108537
我 想 可能	36.96994	15	0.451	3.639607
我 覺得 好像	36.96994	14	0.4395	3.50649
我們 今天 的	36.96994	13	0.391	2.025689
找 不 到	36.96994	13	0.389	7.302596
那 一 天	36.96994	13	0.291	3.369593
的 過程 當中	36.96994	11	0.614	3.906845
是 因為 我	36.96994	20	0.331	0.734919
是 哪 一	36.96994	16	0.5035	2.069618
是 站 在	36.96994	12	0.29	2.971854
真的 是 很	36.96994	12	0.4485	3.903055
基本 上 我	36.96994	13	0.518	1.92029
第一 種 是	36.96994	14	0.6155	4.217311
這 是 我們	36.96994	13	0.39	2.366079
這 都 是	36.96994	15	0.3135	0.478067
這 種 人	36.96994	11	0.5065	1.677726
這 種 東西	36.96994	15	0.3225	4.734823
這 種 情形	36.96994	13	0.3245	5.55868
都 沒有 了	36.96994	11	0.432	3.535329
就 不 一樣	36.96994	13	0.369	2.331565
就 覺得 說	36.96994	12	0.4755	2.805107
還 有 什麼	36.96994	12	0.4745	3.131112
覺得 應該 是	36.96994	13	0.439	4.034063
也 是 滿	34.79524	11	0.3875	5.155078
它 就 是	34.79524	13	0.377	2.293501
同 一 個	34.79524	14	0.369	1.294648
你 一定 要	34.79524	12	0.4255	2.992808
完 了 以後	34.79524	8	0.649	7.700967
我 個人 的	34.79524	13	0.3225	2.801201
我 都 不	34.79524	14	0.446	1.383885

我	就	不	34.79524	24	0.2695	0.500037
我	說	你	34.79524	13	0.549	1.860012
我	覺得	最	34.79524	15	0.4525	2.801616
我們	那	個	34.79524	11	0.4795	0.313817
的	一	面	34.79524	11	0.456	3.174846
的	方式	來	34.79524	15	0.442	3.90117
是	自己	的	34.79524	17	0.3355	0.405232
是	非常	的	34.79524	15	0.432	1.87879
是	這樣	的	34.79524	17	0.3415	0.887408
是	錯	的	34.79524	12	0.395	4.098028
看	不	到	34.79524	9	0.526	5.199621
剛剛	講	的	34.79524	15	0.393	3.259109
這	個	機會	34.79524	13	0.3675	4.008189
都	是	要	34.79524	13	0.4455	1.401709
就	把	它	34.79524	9	0.5465	2.912982
就	會	覺得	34.79524	14	0.389	3.137055
就	覺得	很	34.79524	13	0.339	2.991945
會	比較	好	34.79524	10	0.627	4.159282
一	段	時間	32.62054	13	0.2955	4.810023
一定	要	有	32.62054	12	0.2125	2.436561
又	不	是	32.62054	10	0.347	3.001781
也	有	很多	32.62054	12	0.286	3.817257
不過	我	想	32.62054	10	0.469	5.205439
比較	好	的	32.62054	13	0.339	2.096913
它	是	一	32.62054	11	0.392	3.171898
因為	我	想	32.62054	12	0.484	2.785593
在	這	一	32.62054	19	0.3255	1.604442
有	一	點	32.62054	13	0.3895	2.887412
你	就	可以	32.62054	14	0.3805	3.260195
我	也	覺得	32.62054	12	0.4795	3.575775
我	就	想	32.62054	22	0.2715	2.631975
我	想	應該	32.62054	14	0.427	4.198978
我	會	覺得	32.62054	14	0.316	2.132931
我	還	有	32.62054	13	0.3695	0.021307
那	事實	上	32.62054	10	0.54	2.582724
事實	上	是	32.62054	10	0.474	1.21124
所以	我	是	32.62054	8	0.6175	0.013944
所以	那	個	32.62054	15	0.209	1.459348

要 做 什麼	32.62054	11	0.415	4.377564
做 什麼 事情	32.62054	12	0.482	6.505993
問題 的 時候	32.62054	12	0.4895	3.43668
這 方面 的	32.62054	13	0.4455	3.05776
這 個 可能	32.62054	13	0.392	0.919149
這 個 階段	32.62054	12	0.5395	4.265501
這 種 感覺	32.62054	8	0.465	4.059282
就 是 這樣子	32.62054	10	0.6275	2.723251
會 有 一	32.62054	28	0.472	1.462759
整 個 的	32.62054	11	0.4925	1.143354
講 的 話	32.62054	13	0.3705	4.550411
讓 我 覺得	32.62054	9	0.5415	3.408924
也 是 有	30.44584	15	0.3895	0.399858
大 的 一	30.44584	11	0.5125	1.683778
什麼 事情 都	30.44584	12	0.276	5.331016
他 就 不	30.44584	17	0.4445	1.249345
他 就 跟	30.44584	12	0.308	3.297961
他 說 那	30.44584	11	0.5555	1.707084
可能 就 是	30.44584	13	0.2715	2.034446
可能 就 會	30.44584	11	0.426	4.3884
用 這 個	30.44584	18	0.454	1.0945
有 一定 的	30.44584	6	0.4465	4.400579
完全 不 一樣	30.44584	12	0.364	7.99522
快樂 的 事情	30.44584	8	0.6095	7.362048
我 也 會	30.44584	12	0.361	1.930992
我 不 喜歡	30.44584	7	0.424	2.183424
沒有 一 個	30.44584	14	0.5645	0.169202
那 時候 就	30.44584	11	0.506	2.182486
那 種 很	30.44584	12	0.459	2.047813
所 講 的	30.44584	8	0.639	3.944996
所以 我 不	30.44584	12	0.2565	0.758721
所以 這 是	30.44584	13	0.379	2.170231
的 很 好	30.44584	16	0.3165	0.035536
的 是 什麼	30.44584	12	0.3495	0.803861
很多 人 說	30.44584	10	0.568	3.290269
怎麼 說 呢	30.44584	13	0.4225	6.511208
為 了 要	30.44584	13	0.3375	3.567574
問 他 說	30.44584	12	0.531	4.578169

doi:10.6342/NTU201600328

這 個 能力	30.44584	12	0.467	4.26384
這些 都 是	30.44584	12	0.406	4.20554
都 不 太	30.44584	13	0.329	3.796714
都 是 很	30.44584	16	0.2985	1.499309
就 可以 了	30.44584	11	0.3315	3.565717
就 有 一	30.44584	16	0.239	0.432795
就 會 很	30.44584	13	0.4135	2.373899
然後 每 次	30.44584	10	0.507	4.972457
然後 那 個	30.44584	11	0.559	1.582497
跟 我 媽	30.44584	12	0.5475	4.840658
整 個 社會	30.44584	11	0.366	7.565716
還 有 他	30.44584	11	0.4995	0.991077
一 個 東西	28.27113	9	0.406	2.245448
一 個 結論	28.27113	10	0.527	5.152529
一 個 新	28.27113	14	0.382	3.123295
一般 來 講	28.27113	13	0.405	7.696141
也 有 一	28.27113	15	0.475	1.438142
也 會 有	28.27113	11	0.428	2.684927
大家 一起 來	28.27113	12	0.409	6.262874
大家 都 知道	28.27113	12	0.519	5.616741
太 棒 了	28.27113	10	0.468	6.591266
心 裡 的	28.27113	10	0.563	1.740792
他 講 的	28.27113	10	0.308	2.121662
可是 這 個	28.27113	10	0.39	0.982526
因為 我 自己	28.27113	12	0.334	3.894018
在 家 裡面	28.27113	14	0.341	5.214606
有 不同 的	28.27113	11	0.437	3.066797
你 就 要	28.27113	10	0.417	2.493003
我 的 感覺	28.27113	8	0.618	2.044969
我 是 想	28.27113	13	0.312	2.900382
我 就 可以	28.27113	10	0.432	1.843456
我們 都 知道	28.27113	11	0.572	5.216646
沒有 這 種	28.27113	15	0.316	2.302138
那 我 是	28.27113	12	0.467	0.106245
那 我們 就	28.27113	9	0.443	2.78202
那 還 有	28.27113	11	0.482	0.985286
事實 上 我	28.27113	16	0.488	1.120302
所 說 的	28.27113	11	0.397	3.970232

是 第一 個	28.27113	12	0.345	1.688436
個 都 是	28.27113	9	0.446	0.012322
第一 個 是	28.27113	12	0.379	1.521184
這 一 段	28.27113	12	0.383	4.047827
這 個 原因	28.27113	10	0.39	3.655755
這 個 單元	28.27113	9	0.598	5.562971
就 夠 了	28.27113	11	0.409	5.661507
會 跟 他	28.27113	12	0.555	3.216472
說 的 話	28.27113	11	0.336	3.87672
一 筆 錢	26.09643	5	0.6325	5.216358
了 很多 的	26.09643	10	0.3935	2.445534
也 是 蠻	26.09643	10	0.6135	5.847222
大概 都 是	26.09643	8	0.5175	3.888563
不 知道 怎麼	26.09643	15	0.5515	4.083856
不 知道 為什麼	26.09643	11	0.258	5.351705
不 是 每	26.09643	11	0.4245	2.684853
不 喜歡 吃	26.09643	5	0.6425	4.490401
反正 就 是	26.09643	12	0.3705	4.480947
他 那 時候	26.09643	8	0.569	2.011037
他 真的 是	26.09643	10	0.3965	2.663112
他 跟 我	26.09643	11	0.519	1.839368
他 說 你	26.09643	11	0.291	1.562573
成為 一 個	26.09643	8	0.4235	5.159892
有 沒有 什麼	26.09643	8	0.5365	2.863394
你 剛剛 講	26.09643	10	0.5915	4.552182
我 一定 會	26.09643	11	0.535	2.434177
我 不 太	26.09643	11	0.391	1.107721
我 在 家	26.09643	10	0.5245	1.976028
我 爸 就	26.09643	6	0.606	3.113382
我 的 小孩	26.09643	14	0.54	3.455584
我 看 過	26.09643	10	0.5305	2.663957
我 第一 次	26.09643	9	0.606	2.451488
我 想 大家	26.09643	10	0.4735	3.498776
我 想 很多	26.09643	10	0.4755	3.242116
我 說 他	26.09643	8	0.385	1.63354
我 還 要	26.09643	9	0.5425	2.576885
我 覺得 他們	26.09643	11	0.389	2.315528
我們 大家 都	26.09643	12	0.4055	3.247533

我們	可以	看到	26.09643	11	0.2785	5.57458
我們	的	社會	26.09643	10	0.4915	5.291037
我們	班	上	26.09643	8	0.6165	5.542366
我們	現在	就	26.09643	10	0.5845	2.721536
我們	這	一	26.09643	10	0.3265	2.143089
我們	節目	當中	26.09643	8	0.5915	5.681985
那	個	老師	26.09643	7	0.5355	3.396456
所以	在	這	26.09643	13	0.5045	3.83234
的	特別	來賓	26.09643	10	0.634	1.400225
非常	謝謝	大家	26.09643	9	0.5865	7.391294
很	不	好	26.09643	8	0.515	2.268231
後來	我	就	26.09643	6	0.5405	4.615867
是	你	自己	26.09643	11	0.499	1.670251
是	我	自己	26.09643	14	0.479	0.86292
問	我	說	26.09643	9	0.4165	4.542047
這	個	過程	26.09643	11	0.3585	4.342555
這	個	題目	26.09643	8	0.591	5.012493
這	種	事	26.09643	25	0.2965	2.731781
都	有	他	26.09643	11	0.4795	2.036336
都	是	這樣	26.09643	20	0.253	2.491478
都	是	這樣子	26.09643	12	0.367	3.855335
就	不	好	26.09643	14	0.3465	1.393853
就	會	說	26.09643	10	0.5665	1.665791
然後	你	就	26.09643	11	0.2675	3.936209
然後	那	時候	26.09643	7	0.6005	4.288265
會	有	一些	26.09643	13	0.432	3.630977
像	這樣	的	26.09643	11	0.3365	3.81248
講	得	很	26.09643	9	0.422	4.890893
覺得	我	很	26.09643	10	0.441	3.572135
變	得	很	26.09643	10	0.4995	5.509629
一	個	女孩子	23.92173	10	0.359	4.322031
一	個	機會	23.92173	8	0.6095	3.712857
了	半	天	23.92173	10	0.432	5.793885
下	一	代	23.92173	9	0.417	7.475036
下	一	次	23.92173	8	0.433	4.370472
也	有	人	23.92173	10	0.584	2.85538
也	是	不	23.92173	10	0.509	0.060108
大家	都	是	23.92173	8	0.42	1.030889

doi:10.6342/NTU201600328

不 是 什麼	23.92173	11	0.341	0.719631
不 敢 去	23.92173	6	0.611	4.851957
心 裡 就	23.92173	10	0.398	2.9038
心情 不 好	23.92173	6	0.514	6.736051
比較 好 一點	23.92173	11	0.328	6.939948
吃 的 東西	23.92173	10	0.238	5.402133
回來 的 時候	23.92173	10	0.327	4.591504
在 同 一	23.92173	10	0.42	4.652058
年輕 的 時候	23.92173	9	0.429	5.559606
有 些 同學	23.92173	9	0.6105	5.717956
而且 我 覺得	23.92173	10	0.408	3.423027
你 有 什麼	23.92173	9	0.34	2.316345
你 的 生活	23.92173	9	0.596	4.151573
你 想想 看	23.92173	8	0.536	5.327598
你 覺得 是	23.92173	9	0.579	1.617089
我 都 沒有	23.92173	8	0.403	1.192997
我 覺得 如果	23.92173	12	0.417	2.253275
我們 必須 要	23.92173	8	0.486	4.273569
那 也 是	23.92173	10	0.279	0.459489
那 我 現在	23.92173	9	0.492	3.020982
那 個 時代	23.92173	9	0.427	5.315384
事實 上 我們	23.92173	7	0.49	2.831616
所以 他 就	23.92173	7	0.633	2.928754
所以 我 會	23.92173	11	0.476	1.740391
所以 說 我	23.92173	12	0.5855	2.405297
的 人 就	23.92173	8	0.513	0.844602
的 身 上	23.92173	8	0.504	2.413576
非常 的 好	23.92173	13	0.465	3.184228
很 小 的	23.92173	10	0.437	3.88691
很多 都 是	23.92173	8	0.434	3.313971
怎麼 回 事	23.92173	12	0.338	7.740348
是 不 一樣	23.92173	13	0.338	0.63999
是 不 好	23.92173	13	0.293	0.097596
看 得 到	23.92173	11	0.27	5.470038
要 有 一	23.92173	16	0.415	0.787862
國中 的 時候	23.92173	9	0.5945	6.318648
這 一 方面	23.92173	9	0.421	4.76818
這 個 小孩子	23.92173	9	0.497	2.544432

這	種	問題	23.92173	8	0.5115	3.9967
就	行	了	23.92173	9	0.443	5.50515
就	是	所謂	23.92173	10	0.334	4.192808
就	像	一	23.92173	8	0.435	2.395808
給	他	一	23.92173	16	0.426	1.603059
會	更	好	23.92173	8	0.476	5.140698
碰到	這	種	23.92173	10	0.478	6.218576
種	就	是	23.92173	11	0.419	0.58022
讓	他	去	23.92173	10	0.437	3.797666
一	個	小孩	21.74703	15	0.3215	2.006814
一	個	小孩子	21.74703	10	0.4005	2.753312
一	個	工作	21.74703	6	0.5925	3.150957
一	個	例子	21.74703	10	0.2865	4.473472
一	個	階段	21.74703	15	0.5745	3.512086
一	個	滿	21.74703	9	0.4765	2.232274
一般	都	是	21.74703	6	0.5225	4.727556
上	來	講	21.74703	9	0.4695	3.5084
也	不	太	21.74703	9	0.2515	3.38425
也	不	好	21.74703	9	0.2105	2.090418
也	沒有	什麼	21.74703	9	0.6265	4.118669
大概	就	是	21.74703	10	0.4105	3.499223
不	在	家	21.74703	5	0.5755	2.328704
不	知道	你	21.74703	11	0.4175	1.1515
不	知道	這	21.74703	10	0.3865	0.679259
不	是	為	21.74703	8	0.5105	0.066246
不	喜歡	的	21.74703	7	0.4675	1.405734
不過	我	覺得	21.74703	10	0.5215	3.942361
不管	你	是	21.74703	7	0.4175	3.496624
太	大	的	21.74703	8	0.5205	3.202114
他	會	覺得	21.74703	8	0.5895	2.763401
全部	都	是	21.74703	8	0.4325	2.911421
吃飯	的	時候	21.74703	7	0.5995	5.855966
因為	那	時候	21.74703	8	0.6125	3.988788
在	這	方面	21.74703	6	0.4905	4.547852
好	的	方式	21.74703	8	0.3855	3.031774
好	的	朋友	21.74703	7	0.5745	3.464729
有	一	段	21.74703	10	0.3985	3.82063
有	各	種	21.74703	9	0.3305	3.387628

有　自己　的	21.74703	14	0.5805	0.686162
有　關係　的	21.74703	7	0.5925	2.512781
考　不　上	21.74703	5	0.5065	6.53667
你　也　是	21.74703	9	0.3885	0.393001
你　的　意思	21.74703	8	0.5375	3.49622
你　是　一	21.74703	10	0.5155	0.383612
你　為什麼　不	21.74703	10	0.4285	4.080745
你　給　他	21.74703	7	0.5875	2.304175
你　說　的	21.74703	10	0.4405	0.959312
你　覺得　他	21.74703	7	0.5855	2.759568
告訴　他　說	21.74703	7	0.5905	4.221533
告訴　我　說	21.74703	9	0.3915	4.354559
我　也　很	21.74703	10	0.3975	1.679479
我　心　裡	21.74703	9	0.4005	1.731367
我　必須　要	21.74703	8	0.4945	2.075614
我　自己　也	21.74703	9	0.3835	2.940253
我　自己　是	21.74703	10	0.3845	0.619733
我　每　天	21.74703	9	0.3235	1.302219
我　並　不	21.74703	9	0.5005	1.276667
我　是　很	21.74703	10	0.3905	0.093979
我　是　認為	21.74703	6	0.4105	4.195169
我　個人　覺得	21.74703	5	0.5905	4.961871
我　現在　的	21.74703	11	0.3975	1.230673
我　覺得　現在	21.74703	9	0.3105	2.028694
我們　就　要	21.74703	9	0.4015	3.131655
更　好　的	21.74703	9	0.3975	2.745773
沒有　辦法　去	21.74703	8	0.5015	3.951027
那　本　書	21.74703	5	0.6015	4.749701
那　我　說	21.74703	10	0.3165	1.20194
那　個　樣子	21.74703	8	0.4855	2.088281
所以　這　種	21.74703	9	0.5245	2.186196
的　立場　來	21.74703	8	0.4895	4.420733
的　時候　也	21.74703	10	0.4765	0.464118
的　觀眾　朋友	21.74703	9	0.4745	3.143366
非常　的　重要	21.74703	7	0.5745	7.035649
很　不　容易	21.74703	8	0.6065	5.019343
很　高　的	21.74703	8	0.3935	2.191609
很多　很多　的	21.74703	8	0.4835	4.129551

是	另外	一	21.74703	16	0.3335	1.430833
是	因為	你	21.74703	12	0.4075	1.734311
是	所謂	的	21.74703	10	0.3885	1.774303
是	很	大	21.74703	9	0.3915	1.178391
是	相當	的	21.74703	7	0.5055	2.64512
是	真正	的	21.74703	8	0.3915	2.703489
是	最	好	21.74703	7	0.4875	2.396331
要	去	看	21.74703	10	0.4125	3.735166
首先	我們	來	21.74703	10	0.3825	5.561776
個	所謂	的	21.74703	9	0.4805	2.607029
得	非常	好	21.74703	9	0.5235	5.170188
這	一	次	21.74703	9	0.5185	1.35916
這	件	事	21.74703	26	0.3925	4.703307
這	個	年齡	21.74703	11	0.5755	4.770126
這	個	情況	21.74703	8	0.4645	3.090704
這	個	觀念	21.74703	9	0.5975	3.596781
這	幾	個	21.74703	7	0.4175	2.581752
都	不	敢	21.74703	8	0.5905	4.420704
都	還	沒有	21.74703	8	0.2195	4.424718
就	不	太	21.74703	8	0.4905	1.880195
就	告訴	他	21.74703	10	0.4865	3.715084
就	是	像	21.74703	11	0.2275	0.783139
就	問	他	21.74703	9	0.5085	4.266456
就	會	想說	21.74703	9	0.3085	5.238912
就	說	你	21.74703	11	0.4985	1.963344
就	講	說	21.74703	6	0.5905	2.316382
最	不	喜歡	21.74703	6	0.4745	5.643289
最	主要	是	21.74703	8	0.3155	3.431322
最後	一	個	21.74703	9	0.3105	3.421515
會	有	什麼	21.74703	8	0.5735	2.995843
會	有	很多	21.74703	8	0.4805	2.867629
會	覺得	很	21.74703	16	0.5675	3.153725
實在	是	很	21.74703	8	0.4985	4.281073
說	他	很	21.74703	12	0.4325	3.383825
說	我	要	21.74703	8	0.5175	2.334142
靠	這	個	21.74703	7	0.5405	4.135211
舉	個	例子	21.74703	9	0.2245	11.10124
謝謝	您	的	21.74703	10	0.5735	4.512701

doi:10.6342/NTU201600328

還 不 知道	21.74703	9	0.4405	2.794982

(B) four-word lexical bundles in conversation

Bundle	Relative frequency	Text count	DP	G
每 一 個 人	158.7533	36	0.309	6.17541
是 一 個 很	115.2592	33	0.28	3.575457
我 想 這 個	110.9098	29	0.385	3.568749
這 是 一 個	108.7351	35	0.2285	3.176061
你 是 不 是	91.33751	29	0.2905	2.485948
這樣 的 一 個	91.33751	28	0.342	4.890907
每 個 人 都	86.9881	26	0.2355	6.606436
還 有 一 個	82.6387	26	0.234	4.097853
我 覺得 這 個	80.46399	27	0.35	2.35762
我 跟 你 講	69.59048	10	0.5595	4.553194
也 是 一 個	67.41578	25	0.2155	3.384482
就 是 一 個	67.41578	27	0.161	1.753402
一 個 人 的	60.89167	23	0.271	1.407955
有 一 個 很	60.89167	23	0.287	3.29617
個 很 好 的	56.54227	22	0.3245	4.113104
也 是 一 種	54.36756	23	0.3535	5.081206
是 最 重要 的	52.19286	17	0.4325	3.466246
是 不 是 有	50.01816	17	0.398	1.136512
是 這 個 樣子	50.01816	21	0.3245	3.317365
一 個 很 大	45.66875	17	0.44	4.679144
我 是 一 個	45.66875	15	0.3255	0.980019
重要 的 一 個	45.66875	17	0.2525	4.584568
一 個 很 重要	43.49405	16	0.4215	5.092421
好 的 一 個	43.49405	16	0.4135	2.837521
我 也 不 知道	43.49405	12	0.3355	4.457853
跟 我 講 說	43.49405	9	0.4775	5.510729
不 是 這 個	41.31935	17	0.272	0.813718
我 有 一 個	41.31935	15	0.352	0.853248
我 是 覺得 說	41.31935	16	0.4785	4.293009
的 意思 是 說	41.31935	12	0.492	4.334026
是 不 是 也	41.31935	14	0.3935	2.457551
我 覺得 這 是	39.14465	18	0.449	2.766958
來 看 一 看	39.14465	15	0.563	6.136832

doi:10.6342/NTU201600328

不 是 很 好	36.96994	15	0.41	4.257322
並 不 是 說	36.96994	14	0.397	4.341777
這 是 我 的	36.96994	13	0.2735	3.68832
一 點 就 是	34.79524	11	0.394	4.801977
有 一 個 人	34.79524	17	0.3915	1.99959
我們 是 不 是	34.79524	12	0.3975	1.77593
是 什麼 樣 的	34.79524	12	0.495	2.210366
這 一 類 的	34.79524	12	0.406	4.272919
有 很 大 的	32.62054	16	0.3085	3.497256
我 的 意思 是	32.62054	8	0.5375	4.120012
的 一 件 事情	32.62054	13	0.3805	3.201994
不 好 的 時候	30.44584	8	0.602	4.43159
同 一 時間 再見	30.44584	14	0.647	8.723036
在 這 個 時候	30.44584	12	0.343	4.391818
有 一 句 話	30.44584	12	0.4545	3.712888
我 想 這 是	30.44584	13	0.3935	2.177913
我 覺得 應該 是	30.44584	12	0.39	4.324033
是 每 一 個	30.44584	18	0.339	1.528125
就 是 一 種	30.44584	12	0.2715	2.71685
不 是 一 個	28.27113	11	0.261	0.568979
有 什麼 樣 的	28.27113	12	0.443	2.926893
而 不 是 說	28.27113	11	0.495	5.196943
我 覺得 那 個	28.27113	11	0.511	2.097062
是 不 是 要	28.27113	12	0.444	1.214996
是 有 一 個	28.27113	21	0.379	0.194046
是 很 好 的	28.27113	15	0.355	2.384858
這 是 一 種	28.27113	10	0.397	2.5909
就 跟 我 講	28.27113	7	0.591	4.251167
最 重要 的 是	28.27113	10	0.445	2.362644
會 有 這 種	28.27113	19	0.358	4.526165
跟 他 講 說	28.27113	10	0.613	5.505442
對 我 來 講	28.27113	10	0.604	7.109762
另外 一 個 就	26.09643	8	0.558	3.567475
只 有 一 個	26.09643	12	0.27	4.327108
有 一 個 問題	26.09643	10	0.509	4.090043
有 這麼 一 個	26.09643	5	0.55	5.017219
有 這樣 的 一	26.09643	13	0.5095	3.382299
的 那 種 感覺	26.09643	11	0.416	2.937435

是 一 個 非常	26.09643	10	0.347	3.90672
是 不 是 真的	26.09643	10	0.506	4.238044
是 不 是 就	26.09643	9	0.358	0.279004
是 不 是 會	26.09643	9	0.506	1.3719
是 很 重要 的	26.09643	9	0.5	2.950939
這 個 世界 上	26.09643	11	0.5955	5.774761
這 個 社會 上	26.09643	11	0.396	5.27142
都 是 一 個	26.09643	11	0.3485	1.883948
了 是 不 是	23.92173	10	0.267	0.832359
不 是 那 個	23.92173	10	0.516	1.375333
他 是 一 個	23.92173	11	0.156	1.207889
我 跟 你 說	23.92173	8	0.454	4.139318
的 那 一 種	23.92173	7	0.5445	3.348541
的 這 個 問題	23.92173	10	0.512	1.726346
是 這 個 意思	23.92173	9	0.512	4.150199
個 人 都 是	23.92173	10	0.313	2.139433
家 裡 的 人	23.92173	6	0.6035	5.440101
對 我 來 說	23.92173	10	0.425	5.502468
一 個 人 在	21.74703	11	0.2885	2.00235
也 不 是 說	21.74703	9	0.3275	4.001522
在 這 種 情況	21.74703	6	0.5725	4.719271
我 有 一 次	21.74703	9	0.3855	2.109081
我們 來 看 一	21.74703	9	0.5965	5.167391
那 個 時候 我	21.74703	8	0.4925	1.677132
來 做 一 個	21.74703	9	0.5715	4.209439
所以 我 是 覺得	21.74703	5	0.6025	5.085575
是 一 件 很	21.74703	9	0.3035	4.775012
是 一 種 很	21.74703	8	0.3265	3.301915
是 不 是 可以	21.74703	7	0.4765	2.706641
這 個 是 一	21.74703	10	0.5725	1.341507
這 個 時候 呢	21.74703	5	0.6335	4.931374
都 有 一 個	21.74703	8	0.4975	2.412913
嘛 對 不 對	21.74703	7	0.5855	5.156127
講 一 句 話	21.74703	7	0.5985	5.696346

doi:10.6342/NTU201600328

(C) three-word lexical bundles in news

Bundle	Relative frequency	Text count	DP	G
是 一 個	181.1325	836	0.103	3.257198
最 大 的	154.1093	781	0.088	3.552598
是 一 種	111.0266	545	0.112	4.570535
有 一 個	81.37836	503	0.171	2.283519
的 情況 下	79.9886	447	0.099	4.004295
最 重要 的	79.9886	434	0.0855	4.186816
另 一 個	78.75325	432	0.0845	5.76645
並 不 是	74.12071	405	0.0795	4.226244
一 個 人	68.56166	346	0.1705	2.193458
一 個 月	62.69377	326	0.098	3.762664
很 大 的	59.45098	312	0.153	3.543935
是 不 是	57.13471	298	0.1605	2.986331
這 也 是	54.97286	311	0.0915	4.510331
重要 的 是	53.27426	313	0.1095	4.246669
每 個 人	52.96542	268	0.146	4.872013
而 不 是	50.34031	277	0.138	4.352881
也 就 是	50.03148	271	0.2095	4.60612
可 說 是	48.95055	276	0.0955	6.006263
每 一 個	46.47986	242	0.126	4.794782
三 個 月	44.93568	257	0.134	5.473421
這 是 一	44.16359	267	0.1015	3.222383
個 月 的	42.61941	252	0.0535	1.476467
在 這 個	39.83988	237	0.1545	2.615273
的 一 種	39.06779	209	0.132	1.096104
還 有 一	37.52361	276	0.1465	3.082098
另 一 種	36.28826	209	0.106	6.103413
的 最 大	36.13385	213	0.142	1.45037
的 過程 中	35.36175	188	0.1345	2.90493
在 一 個	34.58966	208	0.135	0.94042
也 不 是	33.97199	191	0.104	3.474982
的 各 種	33.97199	198	0.0935	1.395589
什麼 樣 的	33.66316	180	0.2075	4.326312
是 為 了	33.66316	192	0.506	3.438634
也 是 一	32.42781	236	0.1095	2.139243
去年 同 期	32.42781	96	0.358	9.003726

一 段 時間	32.11897	193	0.0785	5.622348
最 高 的	31.5013	183	0.0975	2.265489
最 好 的	30.11154	172	0.1695	3.977488
這 個 問題	29.95712	162	0.099	5.8263
有 一 天	29.03061	161	0.148	4.166056
有 一 種	28.56736	178	0.127	2.791654
很 好 的	28.56736	163	0.109	3.24253
世界 各 地	28.25852	152	0.1225	7.108474
找 不 到	27.94968	160	0.462	7.507037
有 一 次	27.17759	160	0.141	3.228033
這 次 的	26.71434	152	0.1935	1.135487
所 做 的	26.55992	151	0.1555	4.349914
極 大 的	26.55992	154	0.197	3.63948
個 月 內	26.4055	142	0.158	5.084847
更 大 的	26.09667	152	0.1215	3.793682
有 不同 的	25.94225	154	0.183	3.448043
另 一 位	25.17016	148	0.138	5.450476
的 人 都	25.01574	150	0.1635	2.556788
各 地 的	24.86132	142	0.126	2.015888
這 就 是	24.86132	144	0.1905	3.554236
多 年 來	24.7069	216	0.134	5.035884
年 前 的	24.7069	143	0.1635	1.37072
有 一 位	24.24365	146	0.1335	2.988562
上 半 年	23.93481	104	0.3345	5.463909
可以 說 是	23.78039	140	0.1835	5.575707
最 新 的	23.78039	129	0.1765	2.719014
在 這 種	23.62598	146	0.092	2.166709
所 需 的	23.62598	129	0.15	2.96487
下 半 年	23.16272	108	0.2465	6.062471
也 有 人	23.0083	145	0.1275	2.895511
多 年 的	22.85389	219	0.097	2.453689
有 興趣 的	22.54505	142	0.2055	3.100426
六 個 月	22.39063	126	0.1535	5.463869
有 人 說	22.39063	118	0.114	4.966566
就 是 一	22.39063	144	0.203	1.840357
每 個 月	22.23621	116	0.185	5.684594
是 一 位	21.92738	134	0.1465	2.661361
只 有 一	21.61854	183	0.156	3.358885

不 是 一	21.46412	186	0.1385	1.298418
最 主要 的	21.46412	125	0.1985	3.941075
三 年 前	21.3097	131	0.1675	4.994269
只 是 一	21.15529	150	0.138	3.549949
所 造成 的	21.15529	126	0.1415	4.010377
是 一 項	21.15529	126	0.1055	3.710385
一 個 很	21.00087	112	0.2225	3.259371
是 另 一	20.69203	143	0.1435	2.406893
的 一 大	20.38319	116	0.128	2.209444
這 幾 年	20.38319	115	0.11	5.605083
不 是 很	20.07436	117	0.19	4.151507
在 網路 上	20.07436	83	0.1945	5.265197
更 好 的	20.07436	114	0.0675	3.507968
是 一 大	20.07436	129	0.133	2.160203

(D) four-word lexical bundles in news

Bundle	Relative frequency	Text count	DP	G
有 很 大 的	24.7069	149	0.1545	4.278463
每 個 人 都	24.7069	140	0.175	7.45799
最 重要 的 是	23.93481	145	0.1695	4.63654

doi:10.6342/NTU201600328