Identifying lexical bundles in Chinese

Methodological issues and an exploratory data analysis

Chan-Chia Hsu and Shu-Kai Hsieh
National Taipei University of Business / National Taiwan University

Recurrent word sequences, referred to as "lexical bundles", may be structurally incomplete, but they serve important communicative functions. Despite the essential roles of lexical bundles in discourse, many methodological issues have been raised in the process of identifying lexical bundles, which is generally frequency-based. The present study identifies three-word and four-word bundles in Chinese conversation and news, and efforts are made to respond to methodological challenges encountered in previous studies. We employ a more sensitive dispersion measure, DP, and an internal association measure, G, which help filter out high-frequency word sequences with no identifiable function and reduce the workload of further manual interventions. An exploratory data analysis is then conducted to compare the distributional patterns of lexical bundles in Chinese conversation and news. In Chinese, both the type number and the density of lexical bundles are higher in conversation than in news. This appears to be a strong cross-linguistic tendency that reflects the real-time pressure speakers face in spontaneous speech. The exploratory data analysis also shows that the elements in Chinese bundles are closely associated with each other. This suggests that lexical bundles are useful phrasal units in Chinese discourse, and thus invites further investigations of how lexical bundles are used in Chinese.

Keywords: lexical bundle, multi-word unit, frequency, dispersion measure DP, word association measure G

1. Introduction

The distributional pattern of co-occurrences is a crucial issue in corpus linguistics. The first comprehensive investigation of recurrent word sequences in spoken English was conducted by Altenberg & Eeg-Olofsson (1990), and a similar study for spoken and written Spanish was carried out by Butler (1997). After frequency data have brought to the fore phraseological patterns that used to go unnoticed by

https://doi.org/10.1075/lali.00019.hsu *Language and Linguistics* 19:4 (2018), **525–548**
ISSN 1606-822X / E-ISSN 2309-5067

linguists, theoretical explanations for the ubiquity of recurrent word sequences are provided (see Conklin & Schmitt 2008): from a sociofunctional perspective, recurrent word sequences serve important discourse functions; from a psycholinguistic perspective, these prefabricated chunks may have been stored in the mental lexicon, readily available to language users to effortlessly and fluently handle online interactions (e.g. Pawley & Syder 1983). The profound significance of recurrent word sequences has led them to assume a central place in linguistic studies.

Biber et al. (1999) conducted an even larger-scale investigation of recurrent word sequences in English.[1] The Longman Spoken and Written English Corpus was used, and two registers were chosen: conversation (British English: c. 4,000,000 words; American English: c. 3,000,000 words) and academic prose (c. 5,300,000 words). A large number of multi-word combinations referred to as lexical bundles (e.g. *I don't know what*, *I was going to*, *do you want to*) were identified, i.e. "recurrent expressions, regardless of their idiomaticity, and regardless of their structural status" (Biber et al. 1999: 990). The method for identifying lexical bundles is mainly frequency-based: 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. A lexical bundle often functions as a "pragmatic head" (Biber & Barbieri 2007: 270), expressing stances and/or textual meanings and providing interpretive frames for propositions that follow. For example, *the fact that* conveys a certain stance in the developing discourse. Moreover, empirical studies have suggested that lexical bundles may be psychologically real units, stored and processed holistically (e.g. Jiang & Nekrasova 2007; Tremblay et al. 2009). The Biberian approach has been adopted to identify lexical bundles in a wide variety of corpora, such as historical corpora (e.g. Culpeper & Kytö 2002), learner corpora (e.g. Cortes 2004), newswire corpora (e.g. Partington & Morley 2004), and textbook corpora (e.g. Chen 2010).

Although the method for identifying lexical bundles seems straightforward, many considerations are involved. Challenging issues include what corpus to use, how to determine the bundle length and establish the quantitative criteria, and whether to make manual interventions. These methodological issues have been raised and discussed critically in previous studies, and some have modified the Biberian method to serve their research purposes or to overcome resource limitations (e.g. the relatively small size of the corpus).

So far lexical bundles in languages other than English have not received adequate attention. Spanish bundles are examined in Tracy-Ventura et al. (2007) and Cortes (2008); Kim (2009) is the first to extend the Biberian approach beyond

1. The data in Altenberg & Eeg-Olofsson (1990) was from the London-Lund Corpus, which consisted of nearly 500,000 words.

Indo-European languages, investigating the use of lexical bundles in Korean. In the field of Chinese linguistics, most phraseological/collocational studies deal with idiomatic expressions or focus on selected phrases or frames, and few, if any, studies have been conducted to identify a comprehensive range of lexical bundles in Chinese, which, like Korean, is typologically distinct from English and can provide valuable cross-linguistic insights. A purely frequency-based approach, which is quite similar to the Biberian approach, is adopted in Tao (2015), where the fifty most frequent three-word chunks in spoken Chinese are listed.[2] However, only a few prominent categories are briefly discussed there.[3]

Thus, the present study aims to fill this gap by identifying lexical bundles in Chinese. We identify three-word and four-word bundles in Chinese conversation and news. Efforts are made to respond to the methodological challenges that have been sketched above. Furthermore, after a list of lexical bundles in Chinese is identified, we conduct an exploratory data analysis to reveal their distributional patterns.

This paper is organized as follows. § 2 gives a comprehensive review of methodological issues and concerns about identifying lexical bundles. § 3 introduces the method for identifying lexical bundles in Chinese, providing reasonable alternatives to some practices in the Biberian approach. § 4 presents the overall distribution of lexical bundles in Chinese. § 5 is the conclusion.

2. Methodological issues in identifying lexical bundles

This section reviews methodological challenges encountered by previous studies in identifying lexical bundles and discusses in considerable detail how these issues have been addressed. Major issues concern the corpus, the length of lexical bundles, the quantitative criteria, and the manual interventions involved.

---

2. 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 huge number of disyllabic (compound) words that can usually be split into two morphemes/words. See Li & Thompson (1981: 10–15) for a more detailed discussion about the relationship among morphemes, syllables, and characters in Chinese.

3. They are metalinguistic 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* 'COPULA a CLASSIFIER'), and epistemic stance markers (e.g. *wo juede wo* 'I feel I').

2.1 Issues relating to the corpus

The corpus is highly influential in any study that aims to identify lexical bundles, for size plays a critical role. 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 that corpus will be identified as a lexical bundle. It is argued that this 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). Another problem with a small corpus can arise from the rounding of the actual converted raw frequency (Chen & Baker 2010).[4] Even with these problems, some studies use a relatively small corpus to identify lexical bundles because of difficulties encountered in the process of data collection (e.g. Biber & Barbieri 2007). In such studies, the results need to be interpreted with some caution. In previous studies on lexical bundles, the corpus sizes range from thousands of words to millions of words.

The content of a corpus also has an impact on the identification of lexical bundles. In Chen & 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 texts 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.

2.2 Issues relating to the length of lexical bundles

Although lexical bundles of any length can be identified, most studies focus on three-word and four-word bundles. Conrad & Biber (2004) suggest that many two-word sequences 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

4. For instance, when the frequency threshold of occurring at least forty times per million words is applied to a small corpus of, say, 40,000 words, 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; that is, the converted raw frequency threshold in this case would be rounded up to 2 times. However, when the rounded frequency threshold here is converted back (i.e. fifty times per million words), the original frequency threshold will be found to have been substantially adjusted.

investigation is also a consideration. For example, to obtain a manageable number of lexical bundles, Biber & Barbieri (2007) identify only four-word bundles.

2.3 Issues relating to the quantitative criteria

After a computer program is used to automatically extract word sequences from the corpus, the Biberian approach requires a frequency threshold to identify lexical bundles.[5] Although any frequency threshold is inevitably criticized as arbitrary, the decision involves many considerations. First, as mentioned earlier, 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. Second, the length of lexical bundles can also be an important factor. In Biber et al. (1999), since five-word and six-word bundles are generally less common, a lower frequency threshold of at least five times per million words (as opposed to ten times per million words for three-word and four-word bundles) is adopted.[6] Third, some studies (e.g. Chen & Baker 2010) consider the limitation of their resources, so a conservative frequency threshold is adopted to obtain a manageable size of lexical bundles for further analysis. In previous studies, the frequency thresholds range from five times per million words to forty times per million words. Hyland (2008) suggests that a frequency threshold of occurring at least twenty times per million words can be regarded as conservative.

The Biberian approach also sets a dispersion threshold to guard against idiosyncrasies used by individual speakers/writers and local repetitions reflecting

---

5. A computer program reads each sentence in the corpus and proceeds one word at a time, automatically extracting three-word and four-word sequences. For example, the sentence *I don't know why he left early today* would have the following four-word sequences extracted by the program:

I don't know why
don't know why he
know why he left
why he left early
he left early today

Then, the program sorts all the sequences extracted from the corpus and creates a frequency table to store the results.

6. Also, in De Cock (1998), different frequency thresholds are set for word sequences of different lengths (i.e. occurring at least ten, five, four, three times in the corpus for two-word, three-word, four-word, and five-word sequences, respectively), so that roughly the same proportion (i.e. 10%–12%) of recurrent sequence types can be identified for each sequence length.

the immediate topic of the discourse.[7] The dispersion threshold is mostly set at occurring in at least five different texts in the corpus, though the whole range is from three texts to twenty texts, with the corpus size being an important factor again. However, because most high-frequency word sequences are found to be widely distributed, the dispersion threshold here may be of little practical effect, as Conrad & Biber (2004) admit. For example, most of the bundles they identify occur in more than thirty texts. Partington & Morley (2004) also see the possibility that a high-frequency word sequence occurs in several corpus texts yet is absent in most of the corpus texts. A more sensitive dispersion threshold is needed for studies on lexical bundles.

In Biber et al. (1999) and most follow-up studies, no internal association measure (e.g. Mutual Information, or MI) is adopted. Biber (2009) expresses reservations about the use of MI: first, MI does not consider the order of the words in a sequence; second, MI is known to privilege low-frequency content words. Nevertheless, the high frequency of a lexical bundle does not ensure its semantic or pragmatic coherence (Wray 2002; Salazar 2014), and growing evidence shows that MI is a reliable indicator of which word sequences have distinctive functions in our language use (e.g. Simpson-Vlach & Ellis 2010). Therefore, Salazar (2014) suggests adopting MI scores to screen out high-frequency word sequences that seem to lack identifiable functions. Since word sequences with relatively less frequent content words have been filtered out by the frequency threshold, few negative effects, if any, are observed in Salazar's (2014) list of lexical bundles.

2.4 Issues relating to manual interventions

Recently, some studies have also made manual interventions in the identification of lexical bundles. Chen & Baker (2010) manually exclude context-dependent sequences and word sequences containing content words already present in the essay questions (e.g. *in the UK and*, *the Second World War*). Salazar (2014), even after adopting an internal association measure (i.e. MI), manually 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 4a*). Though computers identify recurrent patterns based on quantitative criteria, it is the researcher who decides whether the computer-yielded results fulfill the research purpose (Wray 2002; O'Keeffe et al. 2007). Still, such criteria are sometimes criticized as subjective (e.g. Hyland 2012).

---

7. Partington & Morley (2004) suggest that individual idiosyncrasies and local repetitions that are excluded by the dispersion criteria can still sometimes be of interest to discourse analysts.

Another related issue concerns overlaps: for instance, both *it has been suggested* and *has been suggested that* are identified as lexical bundles (Biber et al. 1999). Though aware of this issue, many previous studies do not deal with it, simply listing all the overlapping lexical bundles separately. However, Chen & Baker (2010) decide to combine overlapping lexical bundles into a larger unit. The advantage is that the number of lexical bundles is not inflated, but this approach seems to be taking the theoretical stance that shorter bundles are not listed separately in the mental lexicon, which is still a questionable assumption (Tremblay et al. 2009).

2.5 An interim summary

In summary, the method proposed in Biber et al. (1999) has been widely adopted for identifying lexical bundles, and many studies, given their research purposes and research resources available to them, make appropriate adjustments. The methods of the previous studies on lexical bundles are summarized in Appendix A, although it is not an exhaustive list.

3. Identifying lexical bundles in Chinese

This section describes the procedure for identifying Chinese lexical bundles in the present study and elaborates on how we address the methodological issues mentioned in the previous section.

3.1 Extracting high-frequency word sequences

The present study uses the Academia Sinica Balanced Corpus of Modern Chinese (4th edition), hereafter referred to as the Sinica Corpus.[8] It is sufficiently large for a study on lexical bundles, and all the texts there have been segmented by a system developed by the Chinese Knowledge Information Processing (CKIP) Group.[9] However, the conversation subcorpus is much smaller than the news subcorpus (459,833 words and 6,475,872 words, respectively). Another potential problem is that many conversations are recorded from interviews on the radio or talk shows

8. 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.

9. For more details about the segmentation system, refer to http://ckipsvr.iis.sinica.edu.tw/. Note that not all the segmented texts have been manually checked.

on TV. These are not typical naturally-occurring data, but the speakers appear to behave spontaneously, just as they do in daily conversations. With these inherent limitations, the findings in the conversation subcorpus need to be interpreted with some caution.[10]

For now, we closely follow the Biberian approach to automatically extract three-word and four-word sequences.[11] Only uninterrupted word sequences are regarded as potential bundles; that is, word sequences running across a punctuation mark or a turn boundary are excluded. Although it is possible that some word sequences work over sentence or turn boundaries, Butler (1997) suggests that such sequences are not common. Then, following many previous studies, we include for further analysis word sequences that occur at least twenty times per million words.

3.2 Dispersion thresholds

As in most studies on lexical bundles, the dispersion threshold of occurring in at least five different texts is also adopted in the present study. However, as reviewed in § 2, some word sequences that pass the text count threshold are concentrated in just a small number of texts, and they are usually referential expressions and functionally/pragmatically uninteresting. Similar problems are also observed in our Chinese data. Examples include *women de haizi* 'we POSSESSIVE.MARKER child; our children' and *junshi fayanren shi* 'military spokesman office'. A more sensitive dispersion measure is needed to filter them out.

Gries (2008) presents a comprehensive survey on existing dispersion measures (e.g. Carroll 1970) and then proposes a new one, hereafter referred to as DP. The measure DP can be used even when the corpus is not neatly divided. More importantly, it can distinguish distributional patterns that other dispersion 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 portion of the corpus. Take the three-word sequence *shi yi ge* 'be one CLASSIFIER' in the news subcorpus, for example. Table 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 instance, given that the first corpus

---

10. Large corpora of spoken Chinese are still extremely rare. Two large-scale corpora featuring spontaneous speech in Chinese, i.e. the Spoken Chinese Corpus of Situated Discourse and the Lancaster Los Angeles Spoken Chinese Corpus, are restricted to internal use due to human ethics or for confidentiality reasons (Xu 2015).

11. We do this in the software environment R (see Gries 2009), which is available at http://www.r-project.org/.

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.[12] Second, as shown in the second column, the raw frequency of *shi yi ge* in each portion of the corpus is calculated. Then, the third step is to calculate for each portion of the corpus the absolute difference between the expected percentage and the observed percentage, as shown in the third column. The 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.

Table 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		

With the help of this sensitive dispersion measure, we can identify word sequences that pass the text count threshold but have a rather skewed distribution in the corpus. For instance, although *junshi fayanren shi* 'military spokesman office' occurs in fifty newswire texts, its DP value is quite high (i.e. 0.899). A closer examination shows that word sequences with a relatively high DP value are usually specific referential expressions, such as *junshi fayanren shi*, or routine expressions that occur frequently in certain radio or TV programs but not in daily conversations (e.g. *wo shi xuan* 'I be choose' is used by the participants of a quiz show very often). Thus, the use of DP also helps minimize the weakness of our spoken subcorpus (see § 3.1).

We decide to exclude word sequences with a DP value higher than 0.65 from further analysis. If a lower DP threshold were to be adopted, then we would miss too many useful word sequences in Chinese. The DP threshold here echoes the observation in Gries (2008) that a lexical item with a DP value between 0.4 and 0.8

12. The phrase *corpus part*, which is the exact phrasing in Gries (2008), refers to a particular portion of the corpus.

(e.g. *definition*: 0.795; *properly*: 0.625; *house*: 0.453) is certainly known to all native speakers and advanced learners. Although DP is more reliable than text counts, both measures are used in this study. Still, a few word sequences successfully pass the DP threshold yet fail the text count threshold.

3.3 Association threshold

Since an internal association measure can be a useful indicator of which word sequences have essential communicative functions (see § 2.3), we adopt one in the present study. Unlike Salazar (2014), we do not use MI because this measure does not take the word order of a word sequence into account. We adopt G (Wei & Li 2013) instead, which overcomes the weakness of MI and dispels Biber's (2009) concerns about the use of MI in the identification of lexical bundles.

The following is how the G score of a word sequence is determined. To begin with, a word sequence needs to be transformed into multiple pseudo-bigrams: for example, the three-word sequence *shi yi ge* 'be one CLASSIFIER' 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 be extended and applied to longer multi-word sequences.

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

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

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

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

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

E_1 = $P_{word1} \times P_{word2\ word3}$
= 0.013969 × 0.001507 = 2.11e-05

E_2 = $P_{word1\ word2} \times P_{word3}$
= 0.000560 × 0.004884 = 2.74e-06

$WAP_{shi\ yi\ ge} = \frac{E_1}{E_1 + E_2} \times E_1 + \frac{E_2}{E_1 + E_2} \times E_2$

= 1.89e-05

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

= 3.257

G is comparable to MI, and a G score below zero also means that the elements in a word sequence do not co-occur more frequently than expected by chance alone. Therefore, we set the G threshold at zero. A detailed examination of our data confirms that most word sequences with a G score lower than zero are simply combinations of high-frequency (function) words and do not have identifiable functions (e.g. *women yi ge* 'we one CLASSIFIER').

3.4 Other methodological issues and practical solutions

The four quantitative thresholds outlined above (i.e. occurring at least 20 times per million words, occurring in at least 5 different texts, DP no higher than 0.65, G no lower than 0) do not guarantee that all the word sequences in the current list are readily interpretable in functional/pragmatic terms. Following some previous studies (see § 2.4), we make manual interventions. We decide to manually exclude word sequences that do not have identifiable functions in discourse. Most word sequences manually excluded here are composed purely of high-frequency function words (e.g. *le zhe ge* 'ASPECT.MARKER this CLASSIFIER') or contain specific, arbitrary numbers other than one (e.g. *si zhong qingkuang* 'four type situation').

Another methodological issue discussed in § 2 is related to overlaps. For (almost) complete overlaps (e.g. *yisi shi shuo* 'meaning COPULA say; the meaning is' occurs 25 times in the conversation subcorpus, and its longer counterpart *de yisi shi shuo* 'the meaning of … is' occurs 19 times), the shorter bundle is excluded from the current list because it almost always occurs within the longer one and appears to have the same function.[13] For a pair/set of bundles to be treated as complete overlaps, the cut-off point is set at the frequency threshold; that is, since the frequency threshold for the conversation subcorpus is 10 occurrences (i.e. 20 times per million words), *yisi shi shuo*, which occurs only 6 times not within the longer bundle, is treated as a complete overlap. For other overlaps (e.g. *you yi ge* 'have one CLASSIFIER' occurs 343 times in the conversation subcorpus, and its longer counterpart *hai you yi ge* 'still have one CLASSIFIER' occurs only 38 times), both bundles remain in the list because the shorter one occurs outside the longer one very often and each one has its respective function.[14] Both bundles are considered separately, and their frequencies remain unadjusted.

---

13. Yet it should be noted that our decision does not mean accepting at this point any radical stance on the mental storage of lexical bundles (see § 2.4).

14. While *you yi ge* is usually used to introduce a topic, the longer bundle *hai you yi ge* is used not only to introduce a topic but also to further elaborate by naming another item.

4. Results and discussion

We generally follow the Biberian approach to identify lexical bundles in Chinese, and Table 2 summarizes the results of the whole procedure. In Table 2, icon ☑ stands for passing a threshold. In line with expectations, while there are a massive number of sequence types in the corpus, only a tiny proportion of them are frequently used (see Zipf 1949). Besides, both in conversation and in news, the type number of three-word bundles is much larger than that of four-word bundles. It is strikingly evident that very few four-word sequences in the news subcorpus pass the frequency threshold.

Table 2. Number of word sequences passing each 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**

From a methodological perspective, Table 2 demonstrates to what extent our modification (i.e. the use of DP and G as well as our manual exclusion) of the Biberian approach influences the results. The two additional quantitative measures have very little influence on high-frequency word sequences in news: they screen out only three three-word high-frequency word sequences in news, such as *junshi fayanren shi* 'military spokesman office', which is a specialized referential expression in news articles related to the Ministry of National Defense. By contrast, DP and G have a much more significant influence on high-frequency word sequences in conversation: as can be seen in Table 2, they filter out 155 three-word and 23 four-word high-frequency word sequences in conversation. Even though there are more high-frequency word sequences identified in the conversation subcorpus, the proportion of those excluded by DP and G is relatively higher.

We also carefully examine the 100 most frequent three-word spoken bundles: two were filtered out by the DP threshold and another eight by the G threshold (i.e. ten in total, 10%). One three-word spoken bundle excluded by DP is *wo de haizi* 'I DE child; my child' (DP = 0.67), which has no discourse-level function and is locally repeated simply because of the conversation topic. The eight excluded by G (e.g. *shi zhe ge* 'be this CLASSIFIER', *yi ge shi* 'one CLASSIFIER be'), as mentioned at the end of § 3.3, are semantically vague because they are composed of

high-frequency function words only. Although DP and G do not really exclude a substantial number of high-frequency word sequences and manual interventions are still required, these two measures, as we have argued in § 3, are scientifically more solid than text count and MI, which have been widely adopted in previous studies. Furthermore, since our manual analysis here suggests that DP and G do help filter out high-frequency word sequences that are locally repeated or semantically vague, we can utilize these quantitative measures to reduce the workload of further manual exclusion in an efficient, scientifically grounded manner.

From Table 2, we can also clearly see 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. Table 3 presents the percentages of words in lexical bundles. (The percentages in the parentheses are calculated without removing punctuation marks.)

Table 3. Percentages of words in lexical bundles

	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%)

For each set of lexical bundles, the most frequent one is as follows.

- Three-word bundle in conversation: *shi bu shi* 'A-*not*-A *yes-no* QUESTION' (1,317 times per million words), which is usually employed to elicit a response or yield the conversation floor (see also Tao 2015: 343).
- Three-word bundle in news: *shi yi ge* 'COPULA one CLASSIFIER' (181 times per million words), which often serves as a discourse organizer to summarize the main point after a lengthy discussion (see also Conrad & Biber 2004).
- Four-word bundle in conversation: *mei yi ge ren* 'every one CLASSIFIER person; everyone' (159 times per million words), which is an indefinite referential expression.
- Four-word bundle in news: *you hen da de* 'have very large DE' (24 times per million words), which is used to specify quantity or size.

As can be seen above, some lexical bundles occur with a very high frequency. The frequency distributions of lexical bundles in Chinese are presented in Figure 1.[15]

15. As shown in Table 2, there are only three four-word bundles in news. It would be senseless to draw a boxplot with only three data points, and these bundles will not be discussed below. To make the box shapes clear, lexical bundles occurring more than 200 times per million words are not included. All of them are three-word bundles in conversation.

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.

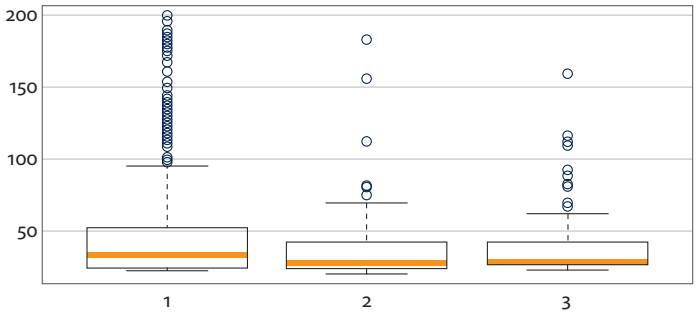


Figure 1. Frequency distributions of lexical bundles

The Shapiro-Wilk normality test shows that the bundle frequencies do not follow normal distributions, so the Mann-Whitney test is performed in Table 4, which presents the frequency means of lexical bundles. It is not surprising that three-word spoken bundles occur more frequently than three-word news bundles ($p = 0.001$). As shown above, the three-word bundle with the highest frequency in conversation occurs approximately seven times more often than that in news. However, the difference between three-word and four-word spoken bundles is not statistically significant ($p = 0.06$).

Table 4. Mean relative frequencies (per million words) of lexical bundles

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

Table 5 shows the text count means of lexical bundles. The text counts are 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. Therefore, the Mann-Whitney test is performed on the text count means.

Table 5. Mean text counts (in percentages) of lexical bundles

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

The large difference between three-word spoken and news bundles achieves statistical significance ($p < 2.2e\text{-}16$), 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.

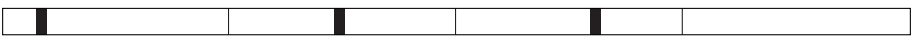
However, DP values show an entirely different tendency. Table 6 presents the DP means of lexical bundles. Only the DP values of four-word spoken bundles follow a normal distribution, so the Mann-Whitney test is still run on the DP means.

Table 6. Mean DP values of lexical bundles

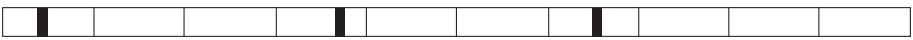
Three-word spoken	Three-word news	Four-word spoken
0.40	0.15	0.42

The DP of three-word news bundles is lower than that of three-word spoken bundles ($p < 2.2e\text{-}16$), but the difference between three-word and four-word spoken bundles is not statistically significant ($p = 0.263$). The DP distributions show that three-word news bundles are more evenly dispersed than three-word spoken bundles.

The reason that text counts and DP values display opposite patterns may be that the former measure is easily susceptible to text lengths. Now consider the following toy example, which is quite similar to the situation in the present study. In Figure 2, the thin bars stand for text boundaries, and the thick bars stand for bundle occurrences.



a. Distribution of lexical bundle *a* in subcorpus A



b. Distribution of lexical bundle *b* in subcorpus B

Figure 2.

The texts in subcorpus A are more than twice as long as those in subcorpus B. There are four texts in subcorpus A, and the bundle *a* occurs in 75% of the texts. There are ten texts in 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*, the same DP values would be obtained. This faithfully reflects that these two bundles are equally well-dispersed. In the present study, the text length difference is even more enormous. On average, each conversation text contains 4,069 tokens, which is almost ten times more than the average number of tokens in the news texts (i.e. 6,475,872/13,800 = 469.2). 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%). The conflicting findings here suggest again that DP is needed to complement text counts in the identification of lexical bundles.

Finally, Table 7 shows the G means of lexical bundles. The G scores of three-word bundles in conversation do not follow a normal distribution, so the Mann-Whitney test is again applied to the G means.

Table 7. Mean G scores of lexical bundles

Three-word spoken	Three-word news	Four-word spoken
3.19	3.76	3.50

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.

5. Conclusion

The present study adopts the Biberian approach to identify lexical bundles in Chinese conversation and news. To effectively deal with methodological issues raised by previous studies adopting the same approach, we employ another more sensitive dispersion measure, DP, and a word association measure, G. We contribute to the methodological discussion pertaining to the identification of lexical bundles by providing a direct contrast between the original approach and the slightly modified approach. It is clearly demonstrated that the two additional measures, especially in the conversation subcorpus (see Table 2), successfully screen out word sequences that occur frequently but lack an identifiable function in discourse. Even with many useful quantitative criteria, manual interventions are still needed to exclude high-frequency word sequences that are semantically/pragmatically vague or completely overlap with a longer lexical bundle. As can be seen in many studies on lexical bundles and other phraseological patterns, "there is no purely automatic way of identifying phrasal units of meaning" (Stubbs 2007: 181). A limitation of this study is that some decisions (e.g. setting quantitative thresholds) involved in the identification of lexical bundles are readily open to criticisms (e.g. arbitrary, subjective), as in other studies on lexical bundles. However, the lexical bundles identified in this study are the results of our strenuous effort to properly tackle methodological issues.

After we identify a list of lexical bundles in Chinese, the results of the exploratory data analysis show intriguing distributional patterns. In Chinese, both the type number and the density of lexical bundles are higher in conversation than in news. This provides cross-linguistic support for a strong tendency that has been observed in English (e.g. Biber et al. 2004) and Spanish (e.g. Butler 1997), i.e. that language users rely on prefabricated chunks more heavily in spoken language than in written language. It is a reasonable strategy for speakers faced with real-time pressure in spontaneous speech (e.g. face-to-face conversations) to retrieve multi-word units that are used repeatedly because of their important functions in discourse and that thus have strong representations in the mental lexicon (Tannen 1982; Johnstone 2002; Conrad & Biber 2004).[16]

Another significant finding is that the internal association score means of lexical bundles identified in this study are higher than three (see Table 7). It has been argued that multi-word combinations achieving that score are useful to speakers (see McEnery et al. 2006). The elements in such word sequences co-occur frequently and are closely associated with each other because of essential communicative functions served by the whole phrasal units. With a few examples (see § 4), we have shown that lexical bundles in Chinese can promote interactions, organize the developing discourse, and have referential uses. In future research, we will delve further into the communicative roles of Chinese bundles in different types of texts.

Acknowledgements

We would like to thank the anonymous reviewers and the editors for their valuable feedback on our manuscript. We also wish to express our deep gratitude to Miao-Hsia Chang, Zhao-Ming Gao, Chia-Lin Lee, and Chia-Rung Lu for their useful comments on our earlier draft. However, any inadequacies that remain in this paper are our own.

16. See Bybee (2007) and Jiang & Nekrasova (2007) for further elaboration on the view that high-frequency multi-word units, such as lexical bundles, are emergent storage/processing units.

Appendix A. 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 & 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 & 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

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)
Nesi & 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 & 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 & 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

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)
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 & 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)

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)
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)

References

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

Biber, Douglas. 2009. A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics* 14(3). 275–311. https://doi.org/10.1075/ijcl.14.3.08bib

Biber, Douglas & Barbieri, Federica. 2007. Lexical bundles in university spoken and written registers. *English for Specific Purposes* 26(3). 263–286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, Douglas & Conrad, Susan & Cortes, Viviana. 2004. If you look at…: Lexical bundles in university teaching and textbooks. *Applied Linguistics* 25(3). 371–405. https://doi.org/10.1093/applin/25.3.371

Biber, Douglas & Johansson, Stig & Leech, Geoffrey & Conrad, Susan & Finegan, Edward. 1999. *Longman grammar of spoken and written English*. London: Longman.

Butler, Christopher S. 1997. Repeated word combinations in spoken and written text: Some implications for functional grammar. In Butler, Christopher S. & Connolly, John H. & Gatward, Richard A. & Vismans, Roel M. (eds.), *A fund of ideas: Recent developments in functional grammar* (Studies in Language and Language Use 31), 60–77. Amsterdam: IFOTT.

Bybee, Joan. 2007. *Frequency of use and the organization of language*. Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195301571.001.0001

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

Chen, Lin. 2010. An investigation of lexical bundles in ESP textbooks and electrical engineering introductory textbooks. In Wood, David (ed.), *Perspectives on formulaic language: Acquisition and communication*, 107–125. London: Continuum.

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

Conklin, Kathy & Schmitt, Norbert. 2008. Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? *Applied Linguistics* 29(1). 72–89. https://doi.org/10.1093/applin/amm022

Conrad, Susan & 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 Reppen, Randi & Fitzmaurice, Susan M. & Biber, Douglas (eds.), *Using corpora to explore linguistic variation* (Studies in Corpus Linguistics 9), 131–145. Amsterdam: John Benjamins. https://doi.org/10.1075/scl.9.09cor

Cortes, Viviana. 2004. Lexical bundles in published and student disciplinary writing: Examples from history and biology. *English for Specific Purposes* 23(4). 397–423. https://doi.org/10.1016/j.esp.2003.12.001

Cortes, Viviana. 2008. A comparative analysis of lexical bundles in academic history writing in English and Spanish. *Corpora* 3(1). 43–57. https://doi.org/10.3366/E1749503208000063

Cortes, Viviana & Csomay, Eniko. 2007. Positioning lexical bundles in university lectures. In Campoy, Mari Carmen & Luzón, María José (eds.), *Spoken corpora in applied linguistics* (Linguistic Insights 51), 57–76. Frankfurt am Main: Peter Lang.

Culpeper, Jonathan & Kytö, Merja. 2002. Lexical bundles in Early Modern English dialogues: A window into the speech-related language of the past. In Fanego, Teresa & Méndez-Naya, Belén & Seoane, Elena (eds.), *Sounds, words, texts, and change*, vol. 2 (Current Issues in Linguistic Theory 224), 45–63. Amsterdam: John Benjamins. https://doi.org/10.1075/cilt.224.06cul

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. https://doi.org/10.1075/ijcl.3.1.04dec

Gries, Stefan Th. 2008. Dispersion and adjusted frequencies in corpora. *International Journal of Corpus Linguistics* 13(4). 403–437. https://doi.org/10.1075/ijcl.13.4.02gri

Gries, Stefan Th. 2009. *Quantitative corpus linguistics with R: A practical introduction*. London: Routledge. https://doi.org/10.4324/9780203880920

Hyland, Ken. 2008. As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes* 27(1). 4–21. https://doi.org/10.1016/j.esp.2007.06.001

Hyland, Ken. 2012. Bundles in academic discourse. *Annual Review of Applied Linguistics* 32. 150–169. https://doi.org/10.1017/S0267190512000037

Institute of Information Science & CKIP Group in Academia Sinica. 2013. *Academia Sinica Balanced Corpus of Modern Chinese*. 4th edn. (http://asbc.iis.sinica.edu.tw/) (Accessed 2016-10-04.)

Jiang, Nan & Nekrasova, Tatiana M. 2007. The processing of formulaic sequences by second language speakers. *The Modern Language Journal* 91(3). 433–445. https://doi.org/10.1111/j.1540-4781.2007.00589.x

Johnstone, Barbara. 2002. *Discourse analysis* (Introducing Linguistics). Malden: Blackwell.

Kim, YouJin. 2009. Korean lexical bundles in conversation and academic texts. *Corpora* 4(2). 135–165. https://doi.org/10.3366/E1749503209000288

Kopaczyk, Joanna. 2012. Applications of the lexical bundles method in historical corpus research. In Pęzik, Piotr (ed.), *Corpus data across languages and disciplines* (Łódź Studies in Language 28), 83–95. Frankfurt am Main: Peter Lang.

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. https://doi.org/10.1075/ijcl.19.2.04len

Li, Charles N. & Thompson, Sandra A. 1981. *Mandarin Chinese: A functional reference grammar*. Berkeley: University of California Press.

McEnery, Tony & Xiao, Richard & Tono, Yukio. 2006. *Corpus-based language studies: An advanced resource book* (Routledge Applied Linguistics). London: Routledge.

Nesi, Hilary & Basturkmen, Helen. 2006. Lexical bundles and discourse signalling in academic lectures. *International Journal of Corpus Linguistics* 11(3). 283–304. https://doi.org/10.1075/ijcl.11.3.04nes

O'Keeffe, Anne & McCarthy, Michael & Carter, Ronald. 2007. *From corpus to classroom: Language use and language teaching*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511497650

Partington, Alan & Morley, John. 2004. At the heart of ideology: Word and cluster/bundle frequency in political debate. In Lewandowska-Tomaszczyk, Barbara (ed.), *Practical applications in language and computers: PALC 2003* (Łódź Studies in Language 9), 179–192. Frankfurt am Main: Peter Lang.

Pawley, Andrew & Syder, Frances Hodgetts. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In Richards, Jack C. & Schmidt, Richard W. (eds.), *Language and communication*, 191–226. London: Longman.

Salazar, Danica. 2014. *Lexical bundles in native and non-native scientific writing: Applying a corpus-based study to language teaching* (Studies in Corpus Linguistics 65). Amsterdam: John Benjamins.

Simpson-Vlach, Rita & Ellis, Nick C. 2010. An academic formulas list: New methods in phraseology research. *Applied Linguistics* 31(4). 487–512. https://doi.org/10.1093/applin/amp058

Stubbs, Michael. 2007. Quantitative data on multi-word sequences in English: The case of the word world. In Hoey, Michael & Mahlberg, Michaela & Stubbs, Michael & Teubert, Wolfgang (eds.), *Text, discourse and corpora: Theory and analysis*, 163–189. London: Continuum.

Tannen, Deborah. 1982. Oral and literate strategies in spoken and written narratives. *Language* 58(1). 1–21. https://doi.org/10.2307/413530

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

Tracy-Ventura, Nicole & Cortes, Viviana & Biber, Douglas. 2007. Lexical bundles in speech and writing. In Parodi, Giovanni (ed.), *Working with Spanish corpora* (Research in Corpus and Discourse), 217–231. London: Continuum.

Tremblay, Antoine & Derwing, Bruce & 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.

Wei, Naixing & 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. https://doi.org/10.1075/ijcl.18.4.03wei

Wood, David. 2010. Lexical clusters in an EAP textbook corpus. In Wood, David (ed.), *Perspectives on formulaic language: Acquisition and communication*, 88–106. London: Continuum.

Wray, Alison. 2002. *Formulaic language and the lexicon*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

Xu, Jiajin. 2015. Corpus-based Chinese studies: A historical review from the 1920s to the present. *Chinese Language and Discourse* 6(2). 218–244. https://doi.org/10.1075/cld.6.2.06xu

Zipf, George Kingsley. 1949. *Human behavior and the principle of least effort: An introduction to human ecology*. Cambridge, MA: Addison-Wesley.

Authors' addresses

Chan-Chia Hsu (corresponding author)
Center for General Education
National Taipei University of Business
321, Sec. 1, Jinan Road
Taipei 10051
Taiwan
chanchiah@gmail.com

Publication history

Date received: 2 November 2016
Date accepted: 11 April 2017