# Multi-word expressions in *Hongloumeng* translations: a corpus-assisted structural and functional analysis

*Kanglong Liu*[1] *Joyce Oiwun Cheung*[2]

**Address:** [1,2]AG518, The Hong Kong Polytechnic University, Hunghom, Hong Kong, China
**E-mail:** [1]kl.liu@polyu.edu.hk; [2]joyce.ow.cheung@polyu.edu.hk
**Correspondence:** Kanglong Liu
**Citation:** Liu, Kanglong, and Joyce Oiwun Cheung. 2020. "Multi-word expressions in Hongloumeng translations: a corpus-assisted structural and functional analysis." *Translation Quarterly* 98: 79-101.

## *Abstract*

*Lexical bundles (LBs), which are affirmed by extant linguistic research to be one of the major differences between native and non-native language production, has been gaining momentum in studying translator's style. The current study uses LBs as an indicator to investigate translator's style of two full-length English translations of Hongloumeng in a systematic manner. The two major translation versions were produced by a native English speaker (i.e. David Hawkes) and a non-native English speaker (i.e. Xianyi Yang) respectively. The former has gained wider acceptability among Anglophone readers. We speculate that their translation styles might have been influenced by their respective first languages (i.e. Hawkes being influenced by English, and Yang being influenced by Chinese). Therefore, Hawkes' and Yang's translations of the dialogue part were first analyzed with keyword analysis to find out the LBs which are overused in respective versions (Key-LBs); the Key-LBs were then categorized according to Biber et al.'s (2004) Structural and Functional Classifications. Hawkes is found to have used predominantly verb phrases and stance markers, while EFL features such as the overuse of prepositional phrases and referential markers are spotted in Yang's version. Our research confirms that LBs can serve as a reliable indicator for studying translator's style.*

## 1. Introduction

Acclaimed as one of China's Four Great Classical Novels, the Chinese classic *Hongloumeng* has drawn attention from both literary and translation research over the decades.

The work is widely acknowledged as one of the greatest of Chinese fiction for it paints a vivid picture of the aristocratic families against the social broad background of the late Qing Dynasty (1644-1911). The original chronicles were composed by Xueqin Cao and E. Gao, of which the former wrote the first 80 chapters while the latter finished the remaining 40 (Cao & Gao 1982). It is also a literary work which has been translated by both native and non-native English speakers, hence providing scholars with abundant materials for comparative translation analysis. From 1979 to 2013, over 1300 *Hongloumeng* research articles were published, many of which focus on English translations of this classic (Ran & Yang 2013). There are three full-length versions, including *The Story of the Stone* translated by David Hawkes and his son-in-law John Minford, *A Dream of Red Mansions* by Xianyi Yang and his wife Gladys Yang, and *The Red Chamber Dream* by B. S. Bonsall. This latter version has never been officially published but is archived in The University of Hong Kong Library (Bonsall 2004), whereas the other two published versions have been read by many people across the globe. These two versions are therefore representative of not only English translation of a Chinese classic but also English translation produced by native and non-native translators respectively. Hawkes and Minford are both native English speakers and Sinologists; Hawkes translated the first 80 chapters and Minford finished the last 40. Xianyi Yang is a Chinese speaker and he acted as the main translator of *Hongloumeng* while his wife Gladys Yang helped typewrite his verbal translations line after line (cited in Li, Zhang & Liu 2011). Thus, this translation can be considered as the work of a non-native speaker. Over the years, Hawkes and Minford's version has been highly recommended. For example, from a world literature perspective, Wang (2016) comments that Hawkes and Minford's *Hongloumeng* translation is extremely popular among the broad reading public. It is often argued that the popularity of Hawkes' version can be ascribed to Hawkes' being a native English speaker. In fact, Yang's version is found to be more syntactically similar to the source text than Hawkes' version (Li & Wu 2017). Therefore, this paper aims to investigate whether Hawkes and Yang were influenced by their respective first language (L1) in an attempt to identify how such an L1 transfer (if any) affects the popularity of the novel among native English readers.

Native and non-native speakers' linguistic choices may differ in many aspects, in which lexical bundles (LBs) is one of the significant features distinguishing EFL English from native English (e.g. Chen & Baker 2010, Wei 2007). Lexical bundles, also known as multi-word expressions (MWEs), n-grams, and formulaic sequences, mean *recurring lexical sequences in a register* (Biber, Conrad & Cortes 2004). Wei (2007) and Chen and Baker (2010) find that Chinese speakers use more prepositions to construct recurrent sequences than their native English counterparts; Chen and Baker (2010) further point out that second language (L2) students (i.e. Chinese speakers in their study) overuse certain LBs which native English speakers seldom use. On the other hand, Biber (2009) finds that 50% of the LBs used in native spoken English are structured as "personal pronoun + verb components", which means the predominant LB

structure in spoken English is verb-phrase based. As Chinese is a topic-prominent language (Yip 1995), it is not surprising that Chinese speakers adhere to the topic-prominence convention that they tend to use prepositions combined with a bare noun phrase in the topic position to ensure grammaticality in English. On the other hand, English is a subject-prominent language which often structures sentences in a subject-predicate relation (Ibid), thus half of the LBs in spoken English are made up of "pronoun + verb" (Biber 2009). As a result, comparative study on LB structures can inform us about different linguistic choices by L1 and L2 speakers. Moving beyond study of the functional aspects, Biber and Barbieri (2007) find that LBs in spoken English are mainly used in asserting stances, while Wei and Lei (2011) and Pan and Liu (2019) find that Chinese speakers under-use participant-oriented LBs. Since previous literature has demonstrated that LBs can be used effectively as a parameter to systematically compare L1 and L2 production, this paper has followed this approach to find out the typological differences between native and non-native translators (i.e. David Hawkes and Xianyi Yang in the current case study). In fact, LBs has been demonstrated as an effective indicator for investigating translator's style (Mastropierro 2018). As an indicator, LBs has also been used in *Hongloumeng* translation research. Based on the first 15 chapters of *hongloumeng* translations, Liu and Afzaal (forthcoming) has demonstrated that Hawkes' translation is embedded with a greater number and variety of lexical bundles than the one by the Yang couple. Although their study has shown major differences in the use of LBs between the translations of Hawkes and Yang, it is believed that a study taking all 80 chapters into consideration should yield more rigorous results.

Egbert and Mahlberg (2020) find that dialogue usually contains more personal pronouns and present tense, and it can be assumed that fictional dialogues are closer to our daily conversation than the narration component. Therefore, the current study is based on the dialogue part of both translations (all 80 chapters) for comparative analysis on translator's style. The representation of LBs in the dialogue part of respective translation can serve as a departure point for the identification of the "*translator's idiosyncrasies and conscious interpretive or unconscious idiolectal choices*" (Munday 2012, 144)

## 2. Framework

This study adopted the Structural and Functional Classifications framework proposed by Biber et al. (2004) to look into how Hawkes and Yang used lexical bundles (also known as multi-word expressions) differently. Structural Classification is a system which broadly categorizes expressions into several groups based on their parts-of-speech (POS) distributions. For expressions which contain at least one verb component, they are classified as verb-phrase based (VP-based). For the others which do not have any verb components, they are classified as noun-phrase based (NP-based) if a noun component comes before prepositions or other POS

components. If a preposition comes before nouns, the expression is classified as prepositional-phrase based (PP-based). As for those without any verbs, nouns, or prepositions, they are classified as others. While structural classification is useful in differentiating the POS structures favored by different speakers (i.e., translators in the current study), functional classification enables us to compare translators' styles in terms of their communicative goals. Expressions can be broadly categorized into stance, discourse markers, referential, and special conversational functions, depending on the context. Sometimes an expression can perform more than one function, for example, *I want to* can be a discourse marker which introduces a topic; alternatively, *I want to* can also express desire. To decide on the major function of an expression, we insist on context-based annotation. In the current study, each sentence in which an expression occurs has been studied before we ultimately annotated the expression with its most frequent function.

## 3. Data and procedure

This paper inherited the English Chinese Parallel Translation Corpus (ECPTC) which was built by Li et al. (2011); the corpus was compiled by either scanning hard copies or downloading soft copies from the Internet. It consists of three parts running in parallel, namely the original Chinese texts, the translation by Hawkes and Minford, and the translation by Yang. Since Hawkes only translated the first 80 chapters and Minford translated the remaining 40, only the first 80 chapters were used in the current study to facilitate comparison of how a native English translator's (i.e. Hawkes in this case) use of lexical bundles differs from that of a non-native English translator (i.e. Yang in this case).

A self-written Python programme was utilized to automatically extract the dialogues using punctuation (in this case, quotation marks) to separate fictional speech and narration. The data were then manually proofread to ensure accuracy. Upon completion, we have compiled two corpora, namely, the *Yang Dialogue corpus* (YD) and the *Hawkes Dialogue corpus* (HD). YD consists of 219,478 tokens (i.e. the total number of orthographic words separated by spaces and punctuations) and 9,801 types (i.e. the number of distinct words in the corpus), while HD has 280,682 tokens and 10,734 types (see Table 1). Although Hawkes used more words to translate the first 80 chapters, by dividing the number of types by tokens (i.e. type/token ratio or TTR) we can actually see a higher TTR in YD, showing that Yang has used a wider range of distinct words. As YD and HD differ in size, Standardized TTR (sTTR) of the two corpora were also calculated by working out the average of all the TTRs per 1,000 words. YD has a higher sTTR than HD, confirming that Yang indeed used more distinct words than Hawkes did. In order to find out the different representation of Hawkes' and Yang's lexical bundles, we used WordSmith 8.0 (Scott 2020) to firstly turn both corpora into index files which were then used to generate lists of three-word and four-word clusters and their corresponding frequen-

cies. We call these multi-word clusters *lexical bundles* (LBs). YD only has 32,692 tokens and 6,235 types of 3-word LBs and 5,972 tokens and 1,413 types of 4-word LBs, which are smaller than those of Hawkes (60,538 tokens and 10,498 types of 3-word LBs and 12,867 tokens and 2,931 types of 4-word LBs) due to YD's smaller corpus size. If we compare their TTRs of LBs, we can see that YD has a higher TTR in both 3-word and 4-word LBs than HD and it shows that Yang used more distinct LBs.

Table 1: Statistics of the dialogues translated by Yang and Hawkes

| Measures | YD | HD |
|---|---|---|
| Tokens | 219,768 | 280,716 |
| Types | 9,801 | 10,730 |
| TTR[1] | 4.47 | 3.82 |
| STTR[2] | 42.14 | 39.28 |
| Tokens of 3-word LBs | 32,692 | 60,538 |
| Types of 3-word LBs | 6,235 | 10,498 |
| TTR of 3-word LBs | 19.07 | 17.34 |
| Tokens of 4-word LBs | 5,972 | 12,867 |
| Types of 4-word LBs | 1,413 | 2,931 |
| TTR of 4-word LBs | 23.66 | 22.78 |

[1] TTR=Type/token ratio

[2] sTTR=Standardized type/token ratio

Intrigued by the LBs which were distinctively used by each translator, we conducted two rounds of key-LB analysis (i.e. the first time comparing the YD LBs against the HD LBs as reference corpus, and the second the HD LBs against the YD LBs as reference corpus) by using the built-in keyword analysis of WordSmith 8.0 (Scott 2020). For both rounds, only the LBs whose frequencies are not fewer than three or larger than 10% of the tokens were shortlisted for further analyses. LBs having passed the keyness tests in the analyses (i.e. gaining a BIC>2.5 for p-value<0.1, and Log-likelihood>6.63 for p-value<0.01) would be the Key-LBs, meaning that these LBs have an unusually high frequency in their respective corpus. Among these LBs one can easily find content expressions such as *Our Old Lady* which are irrelevant for function and structural analysis of the LBs; therefore, LBs which contain character names and places were redacted, leaving us with 57 and 139 LB types in YD and HD respectively. We applied Biber et al.'s (2004) *Structural Classification* (i.e. NP-based, VP-based, PP-based and others) and *Functional Classification* (i.e. stance, discourse organizers, referential, and special conversational functions) to classify the Key-LBs, with an ultimate aim to find out the structural and functional differences in the use of LBs between Hawkes and Yang.

# 4. Findings

Our study finds that even though Yang yields a higher TTR of LBs, only 57 of them were exclusively used by Yang himself; Hawkes, on the other hand, shows a lower TTR of LBs but 139 among them were exclusively used by Hawkes himself (see Table 2). This reveals that quite a number of Yang's LBs were also found in Hawkes' translation, but not as many Hawkes' LBs were equally found in Yang's translation. Their distinctive use of LBs differs not only in number but also in structures and functions – while both Hawkes' and Yang's Key-LBs are mostly VP-based (i.e. consisting of a verb component), Hawkes has a higher proportion of VP-based Key-LBs (75.54%) than Yang (61.40%) which is closer to Conrad and Biber's (2005) finding that 90% of the lexical bundles used in spoken British English involve verb components. On the other hand, a higher proportion of PP-based Key-LBs (i.e. bundles starting with a preposition) (17.54%) is found in Yang, which is more in line with the findings in Conrad and Biber (2005) that prepositional phrases are common in academic prose instead. In terms of functions, Hawkes' Key-LBs are prominently making a *stance* (47.48%) while many of Yang's serve as *referentials* (36.84%). We argue that both the structural and the functional differences between their Key-LBs are manifestations of Yang's L1 transfer (i.e. Chinese), which will be explained in detail below.

Table 2: Structural and functional classifications of Yang's and Hawkes' Key-LBs

| | Yang | | Hawkes | |
|---|---|---|---|---|
| Structural classifications | Key-LB types | % | Key-LB types | % |
| NP-based | 9 | 15.79 | 21 | 15.11 |
| VP-based | 35 | 61.4 | 105 | 75.54 |
| PP-based | 10 | 17.54 | 11 | 7.91 |
| Others | 3 | 5.26 | 2 | 1.44 |
| Total | 57 | 100 | 139 | 100 |
| Functional classifications | Key-LB types | % | Key-LB types | % |
| Stance | 10 | 17.54 | 66 | 47.48 |
| Discourse organizers | 10 | 17.54 | 31 | 22.3 |
| Referential | 21 | 36.84 | 37 | 26.62 |
| Special conversational functions | 16 | 28.07 | 5 | 3.6 |
| Total | 57 | 100 | 139 | 100 |

A majority of Hawkes' and Yang's Key-LBs are verb-phrase based (VP-based) which involve at least one verb component; as mentioned, Hawkes has used a higher proportion of VP-based Key-LBs (75.54%) than Yang did (61.40%), so the paper proceeds to study their sub-patterns. Our finding shows that 40.95% of Hawkes' Key-LBs start with a personal pronoun (e.g. *I, you, she*), 29.52% start with a verb (e.g. *be, do, have*, modal and other verbs), and

Table 3: Statistics of VP-based Key-LBs in HD

| VP-based Key-LBs | Hawkes' key-LB types | % | Yang's key-LB types | % |
|---|---|---|---|---|
| Starting with personal pronouns: | 43 | 40.95 | 5 | 14.29 |
| Starting with verbs (including be, do, have, modal verbs, and other verbs): | 31 | 29.52 | 15 | 42.86 |
| Starting with conjunctions, that, to, or not to: | 22 | 20.95 | 6 | 17.14 |
| Starting with wh- words: | 5 | 4.76 | 8 | 22.86 |
| Starting with existential markers (including there and this): | 2 | 1.9 | 1 | 2.86 |
| Starting with an adjective: | 2 | 1.9 | 0 | 0 |
| Total | 105 | 100 | 35 | 100 |

20.95% start with either a conjunction or linking words like that and to (see Table 3). Since the most predominant form of VP-based key-LB starts with personal pronouns, such a section will further explain this pattern with an exemplar. The VP-based Key-LBs which start with a verb will be discussed, as this pattern is commonly found in both Hawkes' and Yang's dialogue translations.

Following Li and Thompson's (1976), Yip (1995, 74) distinguishes between topic-promin-ence and subject-prominence to describe the difference between Chinese and English, suggesting that Chinese is a prime example of a topic-prominent language in which topic-comment relation plays a larger role whereas English is more of a subject-prominent language in which subject-predicate relation prevails. In other words, Chinese speakers tend to provide comments based on a mutual topic, while English speakers tend to describe a subject. The latter is in line with our finding that many of Hawkes' Key-LBs and especially the most significant one (i.e. *I think you*) are headed by a personal pronoun. *I think you* is the most significantly different LB (BIC: 36.58, LL: 49.70) which can be considered exclusive to Hawkes' dialogue translation. The phrase usually appears at the beginning of a sentence and manifests the subject prominence in English. As we can see in the example pair (see Excerpt 1), the suggestion of paying someone a visit is expressed in the form *I think you should* (i.e. first personal pronoun + verb base + second personal pronoun) in Hawkes' version. Meanwhile, such subject-predicate relation is absent in Yang's version. Yang simply used the directive *Go* to express the character's permission of the visit, a topic which has already been introduced in the previous dialogue exchange. Yang prioritized the topic (*Go*) whereas Hawkes adhered to the English convention of subject-prominence (e.g. *she is, I think you*). This resonates

with our assumption that, even during translation, native English speakers (i.e. Hawkes in this case) still largely use subject-predicate structures (e.g. personal pronoun + verbs) which are significantly less frequent among their Chinese counterparts (i.e. Yang in this case).

Excerpt 1
"你看看就過去罷，那是侄兒媳婦。" [Source]

"Yes," "she is your nephew's wife. I think you should. Just look in for a moment, though, and then join the rest of us." [Hawkes]

"Go if you want, but don't be long," "Remember she's your nephew's wife." [Yang]

Similar contrast is observed in Key-LBs which begin with a verb. *Ought to be* is the most significant Key-LB which starts with a verb component (BIC: 22.89, LL: 36.02), so we use this to exemplify the different foci between the two translators. As we can see in Excerpt 2, *ought to be* follows the subject *You* in Hawkes' version. In his rendition, Hawkes converted the invitation with a sense of urging 請 (literal translation: please) into a subject (*You*) and its predicate (*ought to be getting back*…). Yang on the other hand was not influenced by the subject-predicate convention in English but retained the semantic meaning of "*請 please*" in the source text. Since *Please* has a near equivalent meaning of 請, Yang kept the literal translation in the same order as that of the source text. Subject is again omitted in Yang's version. Excerpts 1 and 2 are just two of the many examples contrasting Hawkes' and Yang's preferences for subject-predicate and topic-comment structures respectively. Overall, Hawkes' Key-LBs follow the spoken English convention that most of the LBs in spoken English involve verb components (Conrad & Biber 2005) in the form of personal pronouns + verb (Biber 2009).

Excerpt 2
"如今來回老祖宗，債主已去，不用躲了。已預備下希嫩的野鶏，請吃晚飯去，再遲一會子就老了。" [Source]
"So now your creditors have gone, you can come out of hiding. You ought to be getting back now in any case. You've got some nice, tender pheasant for dinner and if you leave it much longer it will spoil." [Hawkes]
"Now I've come to report to our Old Ancestress: Your duns have gone, you can come out of hiding. I've some very tender pheasant ready. So please come back for dinner. If you leave it any later, it'll be overcooked." [Yang]

However, this is not the case in Yang's translation. Although more than half of Yang's Key-LBs are still VP-based, this proportion is still fewer than that of Hawkes' because 17.54%

Table 4: Statistics of PP-based Key-LBs in YD

| PP-based Key-LBs | Yang's key-LB types | % |
|---|---|---|
| Starting with a preposition and a determiner: | 6 | 60.00 |
| Starting with two prepositions: | 1 | 10.00 |
| Starting with conjunction: | 3 | 30.00 |
| Total | 10 | 100 |

of Yang's actually belong to prepositional-phrase (PP-based) LBs. Meanwhile, only 7.91% of Hawkes' Key-LBs are PP-based. This indicates that Yang has used quite a number of PP-based LBs which were significantly underused by Hawkes when translating *Hongloumeng* (see Table 4). Yip (1995, 78) pinpoints that bare noun phrases are often placed in the beginning of a sentence to refer to a topic due to topic-prominence in Chinese, but such a syntactic structure (i.e. sentences beginning with a bare noun) is not allowed in English. Hence, Yip believes Chinese speakers strategically use prepositional phrases to encapsulate a bare noun phase when they need to first talk about a topic. Since Yang is a native Chinese speaker, he may also have extensively used prepositional phrases to safeguard the grammaticality of placing a noun phrase topic in the prominent position of a sentence. Using a prepositional noun phrase to start a sentence is, based on our finding, more often found in Yang's dialogue translation. For example, Yang used *If not for* (BIC: 16.52, LL: 29.64) significantly more frequent than Hawkes did. *If not for* is a typical prepositional phrase which consists of the conjunction *if*, the adverb *not*, and the preposition *for*. In Excerpt 3, we can see that the source text in Chinese is structured as 要不是 (*if not*) and 我 (*me*) which Yang directly translated into *If not for me*. As the focus is on the speaker holding back the other one from attacking people, Yang kept this topic in his translation and used the prepositional phrase *If not for* to topicalize the object *me*. The syntactic order of *If not for me* is equivalent to the dependent clause 要不是我 (Literally: *if not me*) in the source text. Instead of topicalizing the object *me* with prepositional phrases, Hawkes followed the subject-prominent convention by using a verb phrase to start the sentence. He used the verb-pronoun-verb clause *Suppose I hadn't been here* to describe an imaginative inaction of the subject. We can see from this example that Yang resorted to prepositional phrases when he needed to topicalize bare nouns, which is an interlanguage feature commonly found among L2 Chinese speakers.

Excerpt 3
"要不是我，你要傷了他的命，這會子可怎麼樣?" [Source]
"If not for me you might have killed her. What do you intend to do now?" [Yang]
"Suppose I hadn't been here to protect her and you really had done her an injury, what would you have had to say for yourself then, I wonder?" [Hawkes]

Yang also used prepositional phrases at the end of sentences. For example, he exten-

sively used *for no reason* to express the absurdity that something happened. *For no reason* is one of Yang's Key-LBs which is composed of a preposition, a determiner and a noun, and it yields a very high keyness (BIC: 18.17, LL: 31.29) which means Yang used it way more frequently than Hawkes did. PP-based LBs like *for no reason*, when placed at the end of sentences, often serve as adverbials. From Excerpt 4 we can see Yang used prepositional phrases to describe the unlikeliness that someone would offend those people. Yang not only used prepositional phrases to make noun phrase topics grammatically sound (e.g. Excerpt 3), but also used them to describe actions. However, no such substantial use of prepositional phrases is found in Hawkes' dialogue translation. Hawkes can use a variety of linguistic choices to achieve the same purpose so he opted for the word *possibly*, a one-word adverb which is simpler than prepositional phrases. So far, our study has found that there are more unique VP-based LBs in Hawkes' dialogue translation and more distinctive PP-based LBs in Yang's dialogue translation; this confirms extant literature that Chinese speakers often use prepositions to introduce noun topics while native English speakers more often use verb phrases to tell subject-predicate relations. This typological difference between Chinese and English, as suggested by Yip (1995), is also found between Chinese and native English translators.

Excerpt 4
"誰可好好的得罪他?" [Source]
"Why should anyone offend them for no reason." [Yang] noun within prepositions
"Who could possibly have offended her?" [Hawkes]

Table 5: Statistics of stance Key-LBs in HD

| Stance functions: | Hawkes' key-LB types | % |
|---|---|---|
| Epistemic stance | 20 | 30.30 |
| Overall attitudinal/modality stance | 4 | 6.06 |
| Desire | 4 | 6.06 |
| Obligation/directive | 19 | 28.79 |
| Intention/prediction | 13 | 19.70 |
| Ability | 6 | 9.09 |
| Total | 66 | 100.00 |

This section moves on to discuss the functional differences between the two translators' Key-LBs. After manual classification, it is found that 47.48% of Hawkes' Key-LBs mainly express stances while 36.84% of Yang's Key-LBs mainly serve as referential bundles. This means almost half of Hawkes' unique LBs come from his use of stance markers. According to Biber and Barbieri (2007), the predominant function of LBs in all spoken university registers (i.e. teaching, class management, office, study groups, and service encounters) is to

express stance. Our finding is hence consistent with the findings of Biber and Barbieri (2007). It primarily supports the assumption that native English translators like Hawkes may keep stance-making in translation, which makes native English translators' use of LBs significantly different from non-native translators. To understand how the two translators exclusively used some bundles to achieve different communicative goals, this section will look into Hawkes' stance bundles and Yang's referential bundles. Among Hawkes' Key-LBs which are classified as stance, 30.30% construe an epistemic stance while 28.79% convey obligations/directives (see Table 5). The rest are distributed among intentions/predications, desire, ability, etc. This means quite a number of Hawkes' Key-LBs are either epistemic or directive. For instance, one of Hawkes' key-LB *I think I* (BIC: 20.40, LL: 33.52) is a very usual epistemic marker in conversational English. It indicates personal opinion and sometimes functions as a hedge to soften the illocutionary force of an assertion. In Excerpt 5, Hawkes added *I think I* before making the judgement of staying overnight. This self-acknowledgment of making the decision or hedging is, however, not mentioned in the source text. It is solely Hawkes' interpretation that a certain degree of hedging is required in this context. Such stance-makers are neither found in the source text nor in Yang's translation. Yang used shan't, the contraction form of shall not, to keep the formality and courtesy in the source text instead. On the other hand, Yang literally rendered the source text without adding any epistemic stances in relation to the context.

Excerpt 5

"有的是炕，只管睡。我是二爺使我送月銀的，交給了奶奶，我也不回去了。" [Source]

"There's plenty of room here for you to sleep. Make yourselves at home. Actually I came here to bring the mistress her monthly allowance. Now that I've given it to her, I think I shall spend the night here as well." [Hawkes]

"Well, there's plenty of room on the kang, just lie down as you like. Second Master sent me to bring the monthly allowance to the mistress, so I shan't be going back either." [Yang]

Apart from stating epistemic stances, Hawkes significantly used more LBs to perform a kind of speech act – directives. Among his stance Key-LBs, 28.79% assert obligation/directives. *You ought to* (BIC: 15.51, LL: 28.64) is one of the LBs with high keyness which functions as instructing people to do the thing conveyed in the subsequent sentence, regardless of context. This LB has a significantly higher frequency in Hawkes' dialogue translation, so we argue that Hawkes' version may contain more obligations and directives. Take a translation pair as an example (see Excerpt 6): in the source text 你細想去 (literal translation: you carefully think about) does not contain any sense of obligation. However, Hawkes added *you ought to be able to* in his translation, which signals an obligation for the listener to work things out themselves and no more hints would be given. This is not given in the source text, so Yang did not mention

any obligations in his translation but used the adverb *Just* to link the subject-less command *work it out yourself*. Overall, our finding has shown that Hawkes' translation has more stance LBs which are significantly more frequent than Yang's translation (i.e. 66 stance Key-LBs in HD and 10 in YD), which is consistent with previous literature (e.g. Biber & Barbieri 2007) that stance LBs are predominant in conversational English. Hence, we argue that Hawkes as a native English speaker tended to add stance LBs during translation while Yang as a non-native speaker used stance LBs to a lesser degree. And among these stance Key-LBs, Hawkes mainly used them to convey epistemic stances or obligation/directives and this has been exemplified in Excerpts 5 and 6.

Excerpt 6

"非也。我哥哥已經相准了，只等來年就下定了，也不必提出人來，我方才說你認不得娘，你細想去。" [Source]

"No, that's not the reason. It's because someone has already been chosen for my brother. We are only waiting for him to come home to make it public. I don't need to name names. If I tell you that you can't possibly become Mamma's god-daughter, you ought to be able to work it out for yourself." [Hawkes]

"No, it's because my brother has already set his mind on someone, and it'll be fixed up as soon as he returns. I needn't name any names. Why did I say you couldn't take her as your mother? Just work it out for yourself!" [Yang]

Table 6: Statistics of referential Key-LBs in YD

| Referential functions: | Yang's key-LB types | % |
|---|---|---|
| Identification/focus | 4 | 19.05 |
| Imprecision | 1 | 4.76 |
| Quantity/specification | 5 | 23.81 |
| Intangible framing attributes | 4 | 19.05 |
| Place reference | 1 | 4.76 |
| Time reference | 3 | 14.29 |
| Multi-functional reference | 3 | 14.29 |
| Total | 21 | 100 |

Unlike Hawkes, many of Yang's Key-LBs are referential and that means 36.84% of the LBs which are unusually frequent in Yang's translation were used for referring to different attributes. What makes Yang's use of LBs different from Hawkes is that Yang mainly used LBs to refer to quantities and qualities while Hawkes mainly used LBs to make a stance. Also unlike Hawkes' stance Key-LBs which are mostly epistemic and obligatory, Yang's referential Key-LBs are distributed across many subfunctions including identification/focus, imprecision, quantity/specification, intangible framing attributes, place, time, and multi-functional

reference (see Table 6). Since Yang's referential Key-LBs are evenly distributed across all subfunctions, we have selected two referential Key-LBs for detailed analysis based on the two LBs' exceptionally high keyness. The first one is *this is just* (BIC: 11.58, LL: 24.70) which functions as an identification/focus marker. Yang used *this is just* significantly more frequent than Hawkes (see Excerpt 7), Yang used *This is just what* while Hawkes used *this way of carrying on*. *This is just what* differs from *this way of carrying on* as the former just refers to an ambiguous subject matter (which readers can by no means infer what is being warned against) but the latter identifies the exact misbehavior. If we look at the corresponding source sentence 正爲勸你這些 (literal translation: just persuading you about these), the word 這 (literal translation: this) is exactly an identifier in Chinese. By starting sentences with identifiers like 這, Chinese speakers can easily follow the topic so they need not reintroduce the topic again and again. Therefore, Yang probably chose *this is just* instead of *this way of carrying on* because it is rather redundant for Chinese speakers to mention the topic (i.e. misconduct in this case). Yang's version kept the ambiguity in the source text; under the influence of Chinese, Yang also used identifiers (e.g. this) to substitute a complex topic which is already known to readers. Therefore, we hypothesize that non-native translators whose first language is Chinese will likely be satisfied with identifiers like *this* and thus undermine the importance to explicate topics. The reason why Yang significantly used more identification LBs is probably due to the Chinese convention that 這 (i.e. this) is a more economical way of substituting a complicated topic. Hawkes on the other hand felt the need to explain the topic clearly.

Excerpt 7
"好好的，正为劝你这些，更说的狠了。" [Source]
"This is just what I wanted to warn you against, yet here you go, talking more wildly than ever." [Yang]
"It's precisely this way of carrying on that I was going to talk to you about, and here you go, ranting away worse than ever!" [Hawkes]

Imprecision is also a function of Yang's Key LBs. *On like this* (BIC: 19.81, LL: 32.94) is another Key-LB with high keyness value which is significantly more frequent in Yang's dialogue translation. This LB does not specify what qualities it is referring to; instead it makes the circumstances off the record and leaves readers room for imagination. For example, in Excerpt 8, Yang used *on like this* to refer to the girl's poor situation which is not explicitly mentioned in the corresponding source sentence. The source text 這個形景 (literal translation: this situation) does not specify clearly what situation the girl is in. On the contrary, Hawkes refrained from using the imprecise LB *on like this* but the noun phrase *her outward behavior*. Again, it is solely Hawkes' personal judgement that 這個形景 (i.e. this situation) is indeed referring to the girl's outward behavior. All the contrasting uses of LBs between the two translators mentioned above reveal how Chinese conventions (e.g., topic-prominence, ambiguity) has been manifested in Yang's translation.

Excerpt 8

"這女孩子一定有什麼話說不出來的大心事，才這麼個形景。外面既是這個形景，心裏不知怎麼熬煎。看他的模樣兒這般單薄，心裏那裏還擱的住熬煎。可恨我不能替你分些過來。" [Source]

"She must have some secret anxiety preying on her mind to carry on like this, yet she looks too delicate to stand much anxiety. I wish I could share her troubles." [Yang]

"One can see from her outward behaviour how much she must be suffering inwardly. And she looks so frail. Too frail for suffering. I wish I could bear some of it for you, my dear!" [Hawkes]

## 5. Discussion and conclusion

This paper has applied keyword analysis to find out 3-word and 4-word lexical bundles (LBs) which are significantly more frequent in either translator's *Hongloumeng* translation. We have found that many of Hawkes' Key-LBs (i.e. LBs unusually frequent in Hawkes' dialogue translation but infrequent in Yang's version) are verb-phrases, while many of Yang's Key-LBs (i.e. unusually frequent in Yang's dialogue translation but infrequent in Hawkes' version) are prepositional phrases. This supports previous literature (e.g. Yip 1995; Biber & Barbieri 2007, 2009; Conrad & Biber 2005) that LBs in spoken English are mostly verb phrases and Chinese speakers tend to use prepositional phrases to topicalize the bare nouns or noun phrases when they speak English. In our study, Hawkes as a native English speaker is found to have used more verb-phrase LBs whereas Yang as a native Chinese speaker is found to have used more prepositional-phrase LBs, confirming our expectation that native English speakers are likely to adhere to English conventions in their translation while Chinese translators are likely to be influenced by L1 transfer. Hawkes' verb-phrase Key-LBs such as *I think you* and *ought to be* are manifestations of subject-prominence in English; Yang's prepositional-phrase Key-LBs such as *if not for* and *for no reason* are influenced by topic-prominence in which preposition phrases often serve as adverbials in Chinese. In addition, Hawkes' Key-LBs such as *I think I* and *you ought to* also resonate with the convention that the most prominent function of LBs in spoken English is stance-making: assert epistemic stance and give directives. Meanwhile, Yang's Key-LBs such as *this is just* and *on like this* reflect Chinese speakers' frequent use of identifiers to substitute complex topics off the record. Hence, Hawkes' frequent use of verb phrases and stance-making LBs, as well as Yang's frequent use of prepositional phrases and referential LBs, lend strong evidence to the hypothesis that translators' styles can be influenced by their respective first languages. As Wang (2016) praised Hawkes' version for its high readability and Yang's version for its linguistic faithfulness to the source text, we believe Hawkes' adherence to the English convention such as native speakers' frequent use of

verb phrases and stance-making LBs is one of the reasons why Hawkes' translation is better received among English readers. It is also found Yang frequently used prepositional phrases to literally translate Chinese sentences (especially those where bare noun phrases are topicalized). Besides, the source text involves a lot of referential markers (e.g. 這個 this) and Yang stayed as true to the source text as possible by frequently using referential equivalents such as *this is just* and *on like this* without adding interpretative descriptions. We argue that these are both conscious choices to stay faithful to the source text as well as a result of L1 transfer, i.e. Yang's first language being Chinese.

This paper sets out to compare native and non-native translators' use of multi-word expressions in terms of syntactic structures and functions, and have yielded positive results to support the argument that translators' styles are to a large extent influenced by their respective first languages. This study is, however, not without limitations. Only translation works by two translators (i.e. Hawkes and Yang) were sampled in the current study so the findings can only suggest, but not confirm, that translator's style is linked to the translator's first language. A large-scale comparative study with more translated texts by translators of diverse language background is thus recommended.

# References

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.

Biber, Douglas, Conrad, Susan, and Viviana Cortes. 2004. "If You Look At…: Lexical Bundles in University Teaching and Textbooks." *Applied Linguistics* 25(3): 371-405.

Bonsall, Bramwell Seaton. Trans. 2004. *The Red Chamber (Hongloumeng)*. The University of Hong Kong.https://lib.hku.hk/bonsall/hongloumeng/title.pdf

Cao, Xueqin, and E Gao. 1982. *Hong Lou Meng* [in Chinese]. People's Literature Press.

Chen, Yu-Hua, and Paul Baker. 2010. "Lexical Bundles in L1 and L2 Academic Writing." *Language Learning & Technology* 14(2): 30-49.

Conrad, Susan,. M., and Douglas Biber. 2005. "The Frequency and Use of Lexical Bundles in Conversation and Academic Prose". *Lexicographica* 20: 56-71. doi: 10.1515/9783484604674.56

Egbert, Jesse, and Michaela Mahlberg. 2020. "Fiction–one Register or Two? Speech and Narration in Novels." *Register Studies* 2(1): 72-101.

Li, Charles N. and Sandra A. Thompson. 1976. "Subject and Topic: A New Typology of Language." In *Subject and Topic*, ed. by Charles N. Li, 457-489, New York, NY: Academic Press.

Li, Defeng, Zhang, Chunling, and Kanglong Liu. 2011. "Translation Style and Ideology: A

Corpus-assisted Analysis of two English Translations of Hongloumeng." *Literary and Linguistic Computing* 26(2): 153-166.

Li, Xi, and Canzhong Wu. 2017. "Coherence in Hong Lou Meng and its English Translations: An Exploratory Investigation". *Functional Linguistics* 4(1): 1-14.

Liu, Kanglong, and Muhammad Afzaal. forthcoming. "Translator's Style through Lexical Bundles: A Corpus-driven Analysis of Two English Translations of *Hongloumeng*." *Frontiers in Psychology* 12.

Mastropierro, Lorenzo. 2018. "Key Clusters as Indicators of Translator Style." Target 30(2): 240-259.

Munday, Jeremy. 2012. *Evaluation in Translation: Critical Points of Translator Decision making*. London and New York: Routledge.

Pan, Fan, and Chen Liu. 2019. "Comparing L1-L2 Differences in Lexical Bundles in Student and Expert Writing." *Southern African Linguistics and Applied Language Studies* 37(2): 142-157.

Ran, Shiyang, and Ping Yang. 2013. "Breaking through the Bottleneck: A Comparative Investigation into the Chinese-English Translation Studies of 'Hong Lou Meng' [in Chinese]." *China Publishing Journal* 12: 61-63.

Scott, Mike. 2020. *WordSmith Tools Version 8*. Stroud: Lexical Analysis Software.

Wei, Naixing. 2007. "Phraseological Characteristics of Chinese Learners' Spoken English: Evidence of Lexical Chunks from COLSEC." *Modern Foreign Languages* 30(3): 281-291.

Wei, Yaoyu, and Lei Lei. 2011. "Lexical Bundles in the Academic Writing of Advanced Chinese EFL Learners." *RELC Journal* 42(2): 155-166.

Yip, Virginia. 1995. *Interlanguage and Learnability: From Chinese to English*. John Benjamins Publishing.

**Appendix A. Yang's 3-word and 4-word Key-LBs**

| Key-LBs | Freq. | BIC | Log-likelihood | Log-ratio | P-value |
|---|---|---|---|---|---|
| A FEW CUPS | 10 | 3.34 | 16.47 | 1,059.58 | 0.0000494857 |
| ARE WE TO | 11 | 4.99 | 18.11 | 1,059.71 | 0.0000207998 |
| AS THE PROVERB | 19 | 4.11 | 17.23 | 3.02 | 0.0000331040 |
| AS THE PROVERB SAYS | 18 | 2.76 | 15.88 | 2.94 | 0.0000673279 |
| BOUND TO BE | 20 | 19.81 | 32.94 | 1,060.58 | 0.0000000066 |
| BUT MIND YOU | 10 | 3.34 | 16.47 | 1,059.58 | 0.0000494857 |
| CARRY ON LIKE | 12 | 6.64 | 19.76 | 1,059.84 | 0.0000087720 |

| | | | | | |
|---|---|---|---|---|---|
| COULD IT BE | 10 | 3.34 | 16.47 | 1,059.58 | 0.0000494857 |
| COUPLE OF DAYS | 30 | 6.41 | 19.54 | 2.26 | 0.0000098684 |
| DO SUCH A | 10 | 3.34 | 16.47 | 1,059.58 | 0.0000494857 |
| DO YOU EXPECT | 11 | 4.99 | 18.11 | 1,059.71 | 0.0000207998 |
| DOES IT MATTER | 12 | 6.64 | 19.76 | 1,059.84 | 0.0000087720 |
| DON'T YOU KNOW | 11 | 4.99 | 18.11 | 1,059.71 | 0.0000207998 |
| EVEN IF HE | 10 | 3.34 | 16.47 | 1,059.58 | 0.0000494857 |
| FOR A COUPLE | 18 | 2.76 | 15.88 | 2.94 | 0.0000673279 |
| FOR A COUPLE OF | 18 | 2.76 | 15.88 | 2.94 | 0.0000673279 |
| FOR A STROLL | 12 | 6.64 | 19.76 | 1,059.84 | 0.0000087720 |
| FOR A WHILE | 29 | 7.26 | 20.38 | 2.40 | 0.0000063471 |
| FOR NO REASON | 19 | 18.17 | 31.29 | 1,060.50 | 0.0000000193 |
| HAVE SUCH A | 10 | 3.34 | 16.47 | 1,059.58 | 0.0000494857 |
| HAVE THE SAME | 12 | 6.64 | 19.76 | 1,059.84 | 0.0000087720 |
| HIGH AND LOW | 11 | 4.99 | 18.11 | 1,059.71 | 0.0000207998 |
| HOW CAN I | 36 | 13.68 | 26.81 | 2.52 | 0.0000002220 |
| HOW CAN WE | 25 | 28.05 | 41.17 | 1,060.90 | 0.0000000000 |
| HOW CAN YOU | 61 | 29.16 | 42.29 | 2.38 | 0.0000000000 |
| HOW COULD I | 20 | 5.47 | 18.59 | 3.09 | 0.0000161950 |
| HOW IT IS | 14 | 9.93 | 23.05 | 1,060.06 | 0.0000015717 |
| HURRY UP AND | 37 | 2.67 | 15.79 | 1.66 | 0.0000708355 |
| I MEANT TO | 17 | 14.87 | 27.99 | 1,060.34 | 0.0000001187 |
| I'D NO IDEA | 10 | 3.34 | 16.47 | 1,059.58 | 0.0000494857 |
| IF NOT FOR | 18 | 16.52 | 29.64 | 1,060.43 | 0.0000000491 |
| IT'S NO USE | 24 | 26.4 | 39.52 | 1,060.84 | 0.0000000000 |
| IT'S NOT THAT | 10 | 3.34 | 16.47 | 1,059.58 | 0.0000494857 |
| JUST WHAT I | 11 | 4.99 | 18.11 | 1,059.71 | 0.0000207998 |
| MUCH THE BETTER | 12 | 6.64 | 19.76 | 1,059.84 | 0.0000087720 |
| MY ADVICE AND | 12 | 6.64 | 19.76 | 1,059.84 | 0.0000087720 |
| NOTHING BUT A | 11 | 4.99 | 18.11 | 1,059.71 | 0.0000207998 |
| ON LIKE THIS | 20 | 19.81 | 32.94 | 1,060.58 | 0.0000000066 |
| ON THE SLY | 14 | 9.93 | 23.05 | 1,060.06 | 0.0000015717 |

| | | | | | |
|---|---|---|---|---|---|
| SAY ONE WORD | 10 | 3.34 | 16.47 | 1,059.58 | 0.0000494857 |
| SO AS TO | 30 | 19.64 | 32.76 | 3.68 | 0.0000000075 |
| SO HOW CAN | 13 | 8.28 | 21.41 | 1,059.96 | 0.0000037095 |
| SO LONG AS | 15 | 11.58 | 24.70 | 1,060.16 | 0.0000006664 |
| SO MUCH THE BETTER | 12 | 6.64 | 19.76 | 1,059.84 | 0.0000087720 |
| TAKE MY ADVICE | 13 | 8.28 | 21.41 | 1,059.96 | 0.0000037095 |
| TAKE MY ADVICE AND | 11 | 4.99 | 18.11 | 1,059.71 | 0.0000207998 |
| THE BLAME ON | 11 | 4.99 | 18.11 | 1,059.71 | 0.0000207998 |
| THIS CHANCE TO | 11 | 4.99 | 18.11 | 1,059.71 | 0.0000207998 |
| THIS IS JUST | 15 | 11.58 | 24.70 | 1,060.16 | 0.0000006664 |
| TO ASK FOR | 26 | 3.68 | 16.80 | 2.25 | 0.0000415054 |
| TO SEE TO | 13 | 8.28 | 21.41 | 1,059.96 | 0.0000037095 |
| TO SHOW MY | 10 | 3.34 | 16.47 | 1,059.58 | 0.0000494857 |
| WHAT DOES IT MATTER | 12 | 6.64 | 19.76 | 1,059.84 | 0.0000087720 |
| WHY NOT GO | 11 | 4.99 | 18.11 | 1,059.71 | 0.0000207998 |
| WHY SHOULD WE | 15 | 11.58 | 24.70 | 1,060.16 | 0.0000006664 |
| WOULDN'T THAT BE | 20 | 19.81 | 32.94 | 1,060.58 | 0.0000000066 |
| YOU DON'T UNDERSTAND | 18 | 16.52 | 29.64 | 1,060.43 | 0.0000000491 |

* Only Key-LBs with BIC>2.5 and log-likelihood>6.63 (for p-value<0.01) are listed here.

**Appendix B. Hawkes' 3-word and 4-word Key-LBs**

| Key-LBs | Freq. | BIC | Log-likelihood | Log-ratio | P-value |
|---|---|---|---|---|---|
| A BIT AND | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| A BIT BETTER | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |
| A BIT OF | 67 | 36.24 | 49.36 | 3.39 | 0.0000000000 |
| A BIT TOO | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| A FEW MINUTES | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| A GOOD JOB | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |

| | | | | | |
|---|---|---|---|---|---|
| A LOT OF | 91 | 16.88 | 30.01 | 1.69 | 0.0000000401 |
| A MATTER OF | 33 | 3.66 | 16.78 | 2.37 | 0.0000418602 |
| A QUESTION OF | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| A THING LIKE THIS | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| A WORD WITH | 36 | 6.32 | 19.44 | 2.49 | 0.0000103774 |
| ABLE TO SEE | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| AND AFTER THAT | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| AND GET IT | 21 | 11.15 | 24.27 | 1,060.29 | 0.0000008329 |
| AND I DON'T | 18 | 7.68 | 20.81 | 1,060.07 | 0.0000050801 |
| AND IN ANY | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| AND IN ANY CASE | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| ARE GOING TO | 33 | 9.31 | 22.43 | 3.11 | 0.0000021743 |
| ARE IN THE | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| AS A MATTER | 29 | 2.61 | 15.73 | 2.50 | 0.0000729987 |
| AS A MATTER OF | 29 | 2.61 | 15.73 | 2.50 | 0.0000729987 |
| AWAY WITH IT | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| BE A BIT | 20 | 9.99 | 23.12 | 1,060.22 | 0.0000015211 |
| EXACTLY THE SAME | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| FOR A BIT | 35 | 8.13 | 21.25 | 2.78 | 0.0000040291 |
| GET ON WITH | 43 | 10.03 | 23.15 | 2.49 | 0.0000014962 |
| GOING TO BE | 47 | 19.75 | 32.87 | 3.20 | 0.0000000069 |
| GOING TO DO | 20 | 9.99 | 23.12 | 1,060.22 | 0.0000015211 |
| GOT TO HEAR | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| HAVEN'T GOT ANY | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| HEAR ABOUT IT | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| I AM AFRAID | 31 | 22.71 | 35.83 | 1,060.86 | 0.0000000001 |
| I AM NOT | 18 | 7.68 | 20.81 | 1,060.07 | 0.0000050801 |
| I AM SURE | 28 | 19.24 | 32.36 | 1,060.71 | 0.0000000099 |
| I DON'T KNOW WHY | 19 | 8.84 | 21.96 | 1,060.15 | 0.0000027787 |
| I DON'T THINK | 53 | 16.44 | 29.56 | 2.57 | 0.0000000513 |
| I HAVE BEEN | 20 | 9.99 | 23.120 | 1,060.22 | 0.0000015211 |

| | | | | | |
|---|---|---|---|---|---|
| I HOPE YOU | 23 | 13.46 | 26.59 | 1,060.42 | 0.0000002493 |
| I SHOULD HAVE | 52 | 6.74 | 19.87 | 1.89 | 0.0000082953 |
| I SHOULD LIKE | 28 | 19.24 | 32.36 | 1,060.71 | 0.0000000099 |
| I SHOULD LIKE TO | 21 | 11.15 | 24.27 | 1,060.29 | 0.0000008329 |
| I SHOULD THINK | 19 | 8.84 | 21.96 | 1,060.15 | 0.0000027787 |
| I THINK I | 29 | 20.40 | 33.52 | 1,060.76 | 0.0000000041 |
| I THINK IT | 25 | 15.77 | 28.90 | 1,060.55 | 0.0000000734 |
| I THINK IT'S | 19 | 8.84 | 21.96 | 1,060.15 | 0.0000027787 |
| I THINK WE | 27 | 18.09 | 31.21 | 1,060.66 | 0.0000000202 |
| I THINK WE OUGHT | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| I THINK YOU | 43 | 36.58 | 49.70 | 1,061.33 | 0.0000000000 |
| I THOUGHT I'D | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| I THOUGHT YOU | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| I WONDER IF | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| IF YOU ARE | 32 | 5.34 | 18.46 | 2.65 | 0.0000173559 |
| IF YOU ASK | 23 | 13.46 | 26.59 | 1,060.42 | 0.0000002493 |
| IF YOU ASK ME | 20 | 9.99 | 23.12 | 1,060.22 | 0.0000015211 |
| IF YOU WILL | 25 | 15.77 | 28.90 | 1,060.55 | 0.0000000734 |
| I'M AFRAID I | 25 | 15.77 | 28.90 | 1,060.55 | 0.0000000734 |
| I'M NOT SURPRISED | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| I'M SURE YOU | 19 | 8.84 | 21.96 | 1,060.15 | 0.0000027787 |
| IN ANY CASE | 69 | 23.09 | 36.21 | 2.43 | 0.0000000001 |
| IS A VERY | 20 | 9.99 | 23.12 | 1,060.22 | 0.0000015211 |
| IS GOING TO BE | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| IS SUCH A | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| IS THE ONE | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| IT MUST HAVE | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| IT SEEMS THAT | 23 | 13.46 | 26.59 | 1,060.42 | 0.0000002493 |
| I'VE JUST BEEN | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| KNOW WHAT THEY | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| LOOK AT YOU | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| ME ABOUT IT | 19 | 8.84 | 21.96 | 1,060.15 | 0.0000027787 |

| | | | | | |
|---|---|---|---|---|---|
| NOT GOING TO | 38 | 10.97 | 24.09 | 2.89 | 0.0000009150 |
| OF THESE DAYS | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |
| ONE OF THESE DAYS | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |
| OUGHT NOT TO | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| OUGHT TO BE | 60 | 22.89 | 36.02 | 2.75 | 0.0000000001 |
| OUT OF HERE | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| SAY THAT I | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| SHALL BE ABLE | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| SHALL BE ABLE TO | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| SHE HAS BEEN | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| SHOULD LIKE TO | 22 | 12.31 | 25.43 | 1,060.36 | 0.0000004560 |
| SORT OF PERSON | 18 | 7.68 | 20.81 | 1,060.07 | 0.0000050801 |
| SORT OF THING | 45 | 14.62 | 27.74 | 2.82 | 0.0000001358 |
| SUPPOSED TO BE | 20 | 9.99 | 23.12 | 1,060.22 | 0.0000015211 |
| SURE TO BE | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| TALK TO YOU | 19 | 8.84 | 21.96 | 1,060.15 | 0.0000027787 |
| TELL HER THAT | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| TELL THEM THAT | 18 | 7.68 | 20.81 | 1,060.07 | 0.0000050801 |
| THAT I AM | 27 | 18.09 | 31.21 | 1,060.66 | 0.0000000202 |
| THAT I SHALL | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| THAT I WAS | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| THAT IF I | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |
| THAT IF YOU | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |
| THAT IT IS | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| THAT SORT OF | 34 | 26.18 | 39.30 | 1,060.99 | 0.0000000000 |
| THAT SORT OF THING | 18 | 7.68 | 20.81 | 1,060.07 | 0.0000050801 |
| THAT THEY ARE | 18 | 7.68 | 20.81 | 1,060.07 | 0.0000050801 |
| THAT WE SHOULD | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| THAT YOU ARE | 30 | 6.39 | 19.51 | 2.97 | 0.0000100017 |
| THAT YOU HAVE | 37 | 7.22 | 20.34 | 2.53 | 0.0000064828 |
| THE WAY I | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |

| | | | | | |
|---|---|---|---|---|---|
| THERE IS A | 27 | 3.52 | 16.64 | 2.82 | 0.0000450919 |
| THERE WOULD BE | 14 | 3.060 | 16.18 | 1,059.71 | 0.0000575279 |
| THING LIKE THAT | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |
| THINGS LIKE THAT | 19 | 8.84 | 21.96 | 1,060.15 | 0.0000027787 |
| THINK OF IT | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| THINK WE OUGHT | 19 | 8.84 | 21.96 | 1,060.15 | 0.0000027787 |
| THINK WE OUGHT TO | 19 | 8.84 | 21.96 | 1,060.15 | 0.0000027787 |
| THINK YOU OUGHT | 18 | 7.68 | 20.81 | 1,060.07 | 0.0000050801 |
| THINK YOU OUGHT TO | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |
| TO DO IS | 22 | 12.31 | 25.43 | 1,060.36 | 0.0000004560 |
| TO DO SOME-THING | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| TO HAVE BEEN | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |
| TO HEAR ABOUT | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| TO TALK TO YOU | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| TO TELL ME | 35 | 5.42 | 18.55 | 2.45 | 0.0000165668 |
| TO THINK THAT | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| TO YOU ABOUT | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| WANT TO GO | 30 | 21.55 | 34.68 | 1,060.81 | 0.0000000010 |
| WE OUGHT TO | 49 | 10.45 | 23.57 | 2.26 | 0.0000011994 |
| WHAT IT IS | 25 | 15.77 | 28.90 | 1,060.55 | 0.0000000734 |
| WHAT YOU ARE | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| WHAT YOU HAVE | 20 | 9.99 | 23.12 | 1,060.22 | 0.0000015211 |
| WHEN YOU ARE | 27 | 18.09 | 31.21 | 1,060.66 | 0.0000000202 |
| WHILE YOU ARE | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| YOU ARE GOING | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |
| YOU ARE GOING TO | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| YOU ARE NOT | 20 | 9.99 | 23.12 | 1,060.22 | 0.0000015211 |
| YOU ARE TOO | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |

| | | | | | |
|---|---|---|---|---|---|
| YOU DON'T NEED TO | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| YOU KNOW WHAT | 35 | 3.08 | 16.20 | 2.19 | 0.0000570292 |
| YOU OUGHT TO | 100 | 15.51 | 28.64 | 1.53 | 0.0000000844 |
| YOU OUGHT TO BE | 22 | 12.31 | 25.43 | 1,060.36 | 0.0000004560 |
| YOU THINK THAT | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |
| YOU WILL BE | 28 | 4.47 | 17.59 | 2.87 | 0.0000273610 |
| YOU WOULD BE | 17 | 6.53 | 19.65 | 1,059.99 | 0.0000092981 |
| YOU WOULD HAVE | 14 | 3.06 | 16.18 | 1,059.71 | 0.0000575279 |
| YOU'LL BE ABLE | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| YOU'LL BE ABLE TO | 15 | 4.22 | 17.34 | 1,059.81 | 0.0000312819 |
| YOU'VE GOT TO | 16 | 5.37 | 18.49 | 1,059.90 | 0.0000170411 |

* Only Key-LBs with BIC>2.5 and log-likelihood>6.63 (for p-value<0.01) are listed here.