Lexicography ASIALEX (2016) 3:39–62
DOI 10.1007/s40607-016-0026-y

ORIGINAL PAPER

A morpheme-based analysis of lexical bundles in Korean: an interface between corpus-driven approach and lexicography

Kilim Nam[1] · Hyun-ju Song[2] · Jun Choi[1]

Received: 15 March 2016 / Accepted: 29 June 2016 / Published online: 11 August 2016

Abstract This study proposes a new methodology for morpheme-based analysis designed to identify multi-word patterns in Korean, which is a typical example of agglutinative languages. The need for a new approach in corpus linguistics, which takes language typological characteristics into consideration, is also a crucial point of this paper. In Korean, functional words like prepositions or conjunctions are realized as bound morphemes (*emi* or *cosa*) that function as 'minimal grammatical units'. When formulaic expressions in Korean are analyzed according to the morpheme unit, as it is the case in our study, the findings yielded show significant differences from those of previous studies. Based on this methodology, our results provide supporting evidence for the following: (1) lexical bundles are prevalent in Korean, just as in English; (2) computer-defined formulaicity might be language-universal; (3) finally, differences in distributions or discourse functions of morphemic bundles in various genres or registers can be language-specific. The external and internal language factors that may influence these differences are discussed.

Keywords Lexical bundle · Morphemic bundle · Morpheme-based analysis · Korean · *emi* · *cosa*

Abbreviations

ADNZ	Adnomializer (suffix)
ADVZ	Adverbializer (suffix)
AUX	Auxiliary (verb/adjective)
CCM	Complement case marker

✉ Jun Choi
c-juni@daum.net

[1] Kyungpook National University, 80 Daehak-ro, Buk-gu, Daegu 41566, Korea

[2] Keimyung University, 1095 Dalgubeol-daero, Dalseo-gu, Daegu 42601, Korea

CONJ	Conjunctive (particle)
DCL	Declarative (ending)
GER	Gerund
HON	Honorific (form)
INF	Infinitive (suffix)
NOMZ	Nominalizer
OM	Objective marker
OBLG	Obligative
PST	Past tense (suffix)
POSS	Possessive marker
QU	Quotative (particle)
RET	Retrospective mood (suffix)
SM	Subjective marker
TOP	Topic marker
VOL	Volitional mood (suffix)

1 Introduction

Over the past few decades, corpus linguistic approach to multi-word sequences has significantly improved in the methodology of multi-word unit identification and classification. Many studies on ‘patterns’ (for example, Hunston and Francis 2000), ‘lexical bundles’ (Biber et al. 1999, 2004; Hyland 2008), and ‘formulaic language’ (Schmitt 2005; Wray and Perkins 2000; Ellis et al. 2008) have proved that the basic semantic unit of human communication consists not of individual words but of sequences of words, i.e., multi-word units (hereafter, MWU). It has also been reported that the sequences are differentially distributed across various registers or genres, such as written and spoken discourse or academic and non-academic discourse.

Multi-word sequences are said to occupy approximately 30–50 % of a language (Foster 2001; Erman and Warren 2000), though the exact number differs from register to register. Moreover, the corpus-based formulaicity is reported to have a “psycholinguistic validity” in language processes of L1 speakers and L2 learners (Ellis et al. 2008). Despite all the findings and implications discussed in previous studies, however, research on lexical bundles hitherto has mainly focused on English and a few other languages from the same inflectional language family. The question whether the same methodology can be applied to languages other than English, such as Korean or Japanese which both belong to the agglutinative language family, has not been thoroughly examined or discussed, and neither has the question whether agglutinative and inflectional languages show similar patterns in the frequency and distribution of MWUs. The need to develop a new methodology to analyze languages other than English and similar languages according to their typological features has become imperative to evidence the universality or individuality (particularity) of formulaic expressions.

To date, Korean lexical bundles have been analyzed in terms of ‘word unit’, that is a unit divided by spaces before and after, just as in English and other inflectional languages. Biber et al. (2010: 87–88) noted that lexical bundles are relatively rare in Korean compared to English or Spanish and mentioned that such rarity might be related to the typological features of the language. Since grammatical functions in Korean are realized as morphological inflections that are attached to a word stem, it seems reasonable, when analyzing multi-word expressions in Korean, to reflect on possible differences that may lie in the status of ‘words’ in each language type (Biber et al. 2010: 91–92). Lee (2012) has also pointed out that the Korean language necessitates, for the identification and analysis of lexical bundles, considering different type of unit that accords with its own typological features and not those of Indo-European languages. Although these problems have been raised in previous works, yet no new methodologies have been designed to analyze agglutinative languages.

In this study, we present a new methodology of ‘morpheme-based analysis’ for the Korean language as an alternative to the ‘word-based analysis’ used for identifying lexical bundles in English and other inflectional languages. Also, we discuss whether morpheme-based lexical bundles can form the basis for an ‘ultimate dictionary’ as Sinclair has defined (Sinclair et al. 2004: xxiv), which would comprehensively describe Korean multi-word units. Korean can be considered as representative of agglutinative languages in which function words, such as prepositions and conjunctions in English, are realized as bound morphemes, i.e., particles (*cosa*) or inflectional endings (*emi*). These particles or endings in agglutinative languages are systematically combined with noun or verb stems. Therefore, we will use the term ‘morphemic bundle’, instead of a ‘lexical bundle’, which we define as a multi-word unit that is identified through the analysis of morpheme-based languages. The research findings are expected to provide supporting evidence for the language-universal prevalence of lexical bundles. The results suggest that lexical bundles are also extremely frequent in Korean, when taking morphemes as units. This is in direct contradiction to the findings presented in previous studies, which adopted a word-based approach to identify lexical bundles in Korean (Kim 2009; Biber et al. 2010). We will also address various issues relating to lexical bundles from a typological perspective. On the one hand, the prevalence of lexical bundles can be considered as a linguistic universal in that speakers of any language type tend to use multi-word sequences in accordance with the economy principle. Thus, a Korean speaker will (naturally) tend to use prefabricated phrases rather than make up new combinations of words and/or morphemes when speaking. On the other hand, the distribution or frequency of the bundles across various registers may be language-specific in that these depend on the demands or cultural conventions of different language communities.

This paper is organized as follows. Section 2 introduces the linguistic features of Korean as an agglutinative language and discusses the need for a morpheme-based approach. In Sect. 3, we briefly explain how the corpus used in this study is designed and suggest a new methodology for identifying morphemic bundles. In Sect. 4, Korean morphemic bundles are investigated quantitatively and qualitatively within the classificatory parameters predefined for this study. This enables us to examine the frequency and distribution of morphemic bundles and establish a functional taxonomy of these morphemic bundles. Section 5 explains the implications carried by such a morpheme-based approach in Korean lexicography. Indeed, lists of morpheme-based lexical bundles extracted from large-scale corpora can be utilized as fundamental lexicographical resources, especially in selecting Korean MWUs as headwords and overcoming the shortcomings within the dictionary entries. Finally, in Sect. 6, we discuss the validity and applicability of this analysis to other areas of applied linguistics, such as Korean lexicography and Korean language teaching.

2 Typological characteristics of Korean and morpheme-based approach

2.1 The status of words and morphemes in Korean

Korean is a typical agglutinative language with SOV word order, in which ‘bound morphemes’, rather than ‘words’, play a fundamental role in grammar. This property requires the knowledge of the functions of particles, called *cosa*, and inflectional endings, called *emi,* for further understanding of Korean. As Sohn (1999: 15) describes it, a typical Korean sentence consists of “a long chain of particles or suffixes in consistent forms and meanings that are attached to nominals or verb stems.” Due to such morphological characteristics, traditional Korean grammarians (Cwu 1910; Pak 1935) have regarded *cosa* and *emi* as minimal syntactic units, which are equivalent to ‘grammatical words’ in English but are not ‘words’ in the sense of ‘minimal free form’ as defined by Bloomfield (1935: 178, 184). From a functional point of view, *cosa* (particle) and *emi* (inflectional ending) function as ‘grammatical elements’ in Sapir’s terms (1921: 32–33, 93) and are similar to Martinet’s ‘moneme’ (1962), i.e., minimal syntactic unit. These are important subject matters in both syntactic and morphological studies of Korean. On the basis of these perspectives, we take the grammatical morphemes of *cosa* and *emi* as basic count units when retrieving n-grams through morpheme-based analysis. Consider the following examples.

(1)	대학은	창의적	인재를	양성하여야	한다.
	tayhakun	*changuycek*	*inceylul*	*yangsenghayeya*	*hanta*
	university.TOP	creative	talented.OM	foster.OBLG	AUX.DCL
	‘University must foster creative and talented human resources.’				
(1’)	대학-은	창의적	인재-를	양성하-여야	하-ㄴ다.
	tayhak-un	*changuycek*	*incey-lul*	*yangsengha-yeya*	*ha-nta*
	university-TOP	creative	talented-OM	foster-OBLG	AUX-DCL
	‘University must foster creative and talented human resources.’				

In (1), the sentence consists of five *ecel* (units divided by spaces) but contains nine morphemic units (1′). The hyphen(-) used in the examples above denotes morphemic boundaries. The Korean grammatical morphemes, namely *cosa* and *emi*, have been isolated in (1′). This does create a considerable difference when calculating n-grams. These *cosa* (‘은 *un* (TOP)’ and ‘를 *lul* (OM)’) and *emi* (‘-여야 –*yeya* (OBLG)’ and ‘-ㄴ다 –*nta* (DCL)’) have several features that distinguish them from functional morphemes in Indo-European languages.

First, the bound morphemes *cosa* and *emi* are perceived to be separable units by Korean native speakers in their language use. Taking the following phrase as an example, ‘표를 요약하면 *phyolul yoyakhamyen* (‘as shown in the table’)’, Korean native speakers would perceive it as a 4-morpheme sequence: ‘표 *phyo* (‘table’) + 를 *lul* (object marking *cosa*) + 요약하 *yoyakha* (‘summarize’) + 면 *myen* (*emi* meaning ‘if’)’, even though it is composed of two spaced units, i.e., two *ecel*. In other words, *cosa* and *emi* hold the same status as independent words in English, which can each be taken as n-gram unit in Korean.

Second, it has been shown that the particles ‘*cosa*’ and inflectional endings ‘*emi*’ in Korean are very often translated into independent words in other languages. Sohn (1999: 15) states that *emi* and *cosa* may be realized in English as conjunctions and adverbs, respectively.

(2)	a. 가-고	b. 가-면	c. 가-니까
	ka-ko	*ka-myen*	*ka-nikka*
	go-CONJ	go-ADVZ	go-ADVZ
	‘go and’	‘if … go’	‘because…go’
(3)	a. 너-의	b. 너-만	c. 너-도
	ne-uy	*ne-man*	*ne-to*
	you-POSS	you-only	you-also
	‘your’	‘you … only’	‘you … also’

In examples (2) and (3), bound morphemes ‘-*ko* (*emi*)’ (2a) and ‘-*man* (*cosa*)’ (3b) are translated into ‘and’ and ‘only’, respectively. This proves that Korean particles (*cosa*) and inflectional endings (*emi*) are semantically independent in sentences.

Last, each type of inflection in Korean represents a single grammatical category, unlike English or German inflections which can express multiple grammatical categories; for example, the suffix ‘s’ in English can either signal the third person singular or express plural nouns, whereas the inflectional ending ‘-*myen*’ (2b) can only express the conditional, and the particle *‘-uy’* (3a) can only express the possessive case. Furthermore, there are many more types of *emi* (over 220 items) and *cosa* (over 400 items) in Korean, and when two or more are combined with a word stem, they follow a strictly fixed order, as shown in example (4). This combinatory system of grammatical morphemes results in complex word structures.

(4) a. 너-마저-도
ne-mace-to
you-even-also
'even you'

b. 가-시-었-겠-더-라
ka-si-ess-keyss-te-la
go-HON-PST-VOL-RET-DCL
'might have gone'

Previous studies (Kim 2009: 144; Biber et al. 2010: 88) argued that lexical bundles were a rare phenomenon in Korean. In fact, they failed to consider the aforementioned complex structure of 'words' in their methodology. Without taking the complex system of inflections of *emi* and *cosa* into account, chances of identifying lexical bundles are slim in number of agglutinative languages, and meaningful units cannot be properly retrieved.

2.2 The basic unit of morpheme-based analysis

Korean lexical bundles can be analyzed according to four types of units: (a) *ecel* unit, (b) word unit, (c) morpheme unit 1 (inflectional level), and (d) morpheme unit 2 (derivational level). Let us now consider the sentence 'Universities must foster creative and talented human resources.' (example 5), which has been analyzed in accordance with the aforementioned unit types (5a–5d).

(5) a. *ecel*-unit analysis

대학은	창의적	인재를	양성하여야	한다.
tayhakun	*changuycek*	*inceylul*	*yangsenghayeya*	*hanta*
university.TOP	creative	talented.OM	foster.OBLG	AUX.DCL

b. word-unit analysis

대학-은	창의적	인재-를	양성하여야	한다.
tayhak-un	*changuycek*	*incey-lul*	*yangsenghayeya*	*hanta*
university-TOP	creative	talented-OM	foster.OBLG	AUX.DCL

c. morpheme-unit 1 analysis (inflectional level)

대학-은	창의적	인재-를	양성하-여야	하-ㄴ다.
tayhak-un	*changuycek*	*incey-lul*	*yangsengha-yeya*	*ha-nta*
university-TOP	creative	talented-OM	foster-OBLG	AUX-DCL

d. morpheme-unit 2 analysis (derivational level)

대학-은	창의-적	인재-를	양성-하-여야	하-ㄴ다.
tayhak-un	*changuy-cek*	*incey-lul*	*yangseng-ha-yeya*	*ha-nta*
university-TOP	creativeness-ADNZ	talented-OM	fostering-do-OBLG	AUX-DCL

Each unit is marked by spaces or hyphens. Thus, (5a) comprises five units, (5b) seven units, (5c) nine units and (5d) eleven units.

The first type of analysis (5a) is based on the *ecel* unit. This is the approach adopted by the previous studies on Korean lexical bundles (Kim 2009; Biber et al. 2010). The reason why in Korean natural language processing (NLP), *ecel* is equated to 'word' is that just as the English word, an *ecel* is a unit isolated by spaces. However, an *ecel* may be composed of independent words and bound morphemes and, therefore, cannot be considered as a ‘minimal free form’ (Bloomfield 1935: 178) even though it is a unit separated by spaces just as the English word (as a unit) is. *Ecel* is not a grammatical but rather a conventional and orthographical unit that is regulated by the word spacing rule in Korean grammar. An *ecel* can consist of a single word; however, it usually combines two words or even more.

The second type of analysis (5b), which counts particles, i.e., *cosa*, as separate units, also raises a crucial issue. Korean school grammar has categorized as words all minimal free forms but also *cosa*, even though these are bound morphemes. Not only is this inclusion questionable, but also it poses the question why inflectional endings, namely *emi*, which are also easily separable units, are not similarly counted as independent units.

The third type of analysis (5c) breaks down the sentence into morphemic units, and more specifically, at the inflectional level; that is to say, the morpheme being the basic unit, *cosa* and *emi* are analyzed as individual units, whereas derivatives and compound words are no further analyzed. This is, in our view, the most appropriate method to identify lexical bundles in Korean. Indeed, the specificity of Korean lexical bundles is that they are composed of not only lexical morphemes but also functional morphemes (*cosa* and *emi*), which have been identified as independent units.

Finally, the morpheme-unit analysis at the derivational level (5d) is the most rigorous way to analyze the bundles from a morphological point of view. In this type of analysis, not only *cosa* and *emi*, but also derivational affixes are taken into account. Derivational affixes combine with nouns and verb stems to produce new words without following any specific pattern (the derivational suffix *–cek*, for instance, can be affixed to some nouns but not others without observing any particular rule), thereby complicating the analysis. Since derivational affixes are not considered as grammatical units and do not contribute to producing grammatical sequences, this method has been excluded from our study.

Most studies, up until now (Kim 2009: 144, Biber et al. 2010: 88), claimed that lexical bundles in Korean were rather rare compared to English. Furthermore, corpus-based Korean dictionaries, such as Yonsei Korean Dictionary (1998) and Korea University Korean Dictionary (2009), only feature typical idioms and proverbs in subentries and have failed to detect and include the various patterns of MWUs that appear with high frequency in spoken and written registers. This type of issues can be explained by the fact that previous research did not consider the complex structure of Korean words but, instead, based the extraction of MWUs on the English model. As an agglutinative language, Korean also counts complex inflectional units such as *emi* and *cosa* within the *ecel* unit. On the basis of orthographic space unit, it is not possible to extract MWUs which begin or end with such bound morphemes. Using a morpheme-based extraction methodology, this study argues that a corpus-driven methodology must be adopted, which takes into consideration the typological characteristics of the Korean language. Such an approach objectively allows the identification and determines the qualification of

MWU headwords, thereby establishing the nexus between corpus-driven research and Korean lexicography.

3 Methodology

3.1 Corpus

For the purpose of this study, we have used a morphologically annotated corpus rather than a raw corpus, although raw corpora present some advantage over annotated corpora, such as enabling researchers to examine language in use without prejudgments and utilizing larger data. Nevertheless, raw corpora fail to satisfy our present need that is to calculate high-frequency sequences of morphemic units with taking into consideration the complex word structures of the Korean language.[1]

The corpus used in this study comprises the Sejong Corpus (SC), compiled by the 21st Century Sejong Project and sponsored by the Korean government and the National Korean Language Institute as well as the Kyungpook National University Academic Corpus (KNUAC). The SC is the largest balanced corpus that has been compiled for the Korean language to date. As for the KNUAC, it is composed of academic articles of three disciplines: linguistics, literature, and education. These articles were collected from an academic journal approved by the National Research Foundation of Korea; for this reason, we believe that they are representative of (proper) academic discourse. We divided the corpus into four categories based on the registers described by Biber et al. (2004) so as to examine the universality and/or specificity of Korean lexical bundles. Table 1 below shows whether the texts composing the corpus belong to the spoken and/or academic discourse and indicates the number of *ecel*, the number of texts, the source of the texts by registers.

It can be noticed that the number of *ecel* varies quite considerably from one register to another. This is due to the lack of morphologically annotated spoken texts and academic articles covered by the SC and the KNUAC, respectively. We have seen earlier that using a morphologically annotated corpus was the most suitable option for the methodology we have adopted in this study. Since, ideally, the number of *ecel* should be more or less the same for each register, but the morphological annotation of spoken texts and academic articles would have required a practically unrealistic amount of time, frequencies have been, instead, standardized to solve this issue.

3.2 The identification of morphemic bundles: lexical bundles versus morphemic bundles

To discuss the universality and specificity of lexical bundles, cutoffs were implemented, as done in previous studies (Biber et al. 1999, 2004, 2010; Cortes

[1] The morphologically annotated corpus is the most frequently used version for the purposes of collecting word frequencies, extracting headwords for Korean dictionaries and selecting word lists for language education.

Table 1 Composition of the corpus

Registers	[+Spoken]	[+Academic]	No. of *ecel*	No. of texts	Sources
Conversation	+	−	293,837	48	SC
Classroom teaching	+	+	225,215	48	SC
Textbook	−	+	1,000,274	48	SC
Academic prose	−	+	364,451	96	KNUAC

Table 2 Cutoffs of extracting morphemic bundles

	Frequency	Text distribution	Length
Biber et al. (1999, 2004)	10/1 million words	More than 5 texts	4-word sequences
Cortes (2004)	20/1 million words	More than 5 texts	4-word sequences
Kim (2009)	20/1 million words	More than 5 texts	3-ecel sequences
Our study	20/1 million words	More than 5 texts	5-morpheme sequences

2004; Kim 2009). Table 2 compares the cutoffs determined for the frequency, distribution, and length of morphemic bundles in these previous studies and those determined in our study.

In this study, we have limited the length of morphemic bundles to 5-morpheme sequences, which seems to us the most appropriate length to identify meaningful units (Choi et al. 2010). In addition to this, we also take into account the quantitative relevance between words and morphemes of Korean, and we have taken into account the quantitative relationship between *ecel* and morphemes in Korean. Indeed, after performing a preliminary extraction of 4-, 5-, and 6-morpheme sequences, we analyzed these by comparing them first with each other and then to 3-*ecel* sequences. As a result, we found that in Korean, 5-morpheme sequences had an equivalent weight to 3-*ecel* sequences., A Korean *ecel* contains on average 1.71 morpheme; therefore, a 3-*ecel* sequence would contain 5.13 morphemes (3 *ecel*s × 1.71 morpheme = 5.13). Table 3 compares the 20 most frequent bundles of three *ecel* and five morphemes in academic prose.

The table shows that while 5-g morphemic bundles and 3-g lexical bundles have an equivalent weight, the borders of the sequences are not necessarily identical. Most of 5-morpheme sequences start with a bound morpheme (e.g. *–myen*), whereas 3-g word sequences always begin with an independent morpheme or a word. In fact, 13 out of 20 of the items above begin with bound morphemes, which are denoted by a hyphen (-). More specifically, the morphemic bundle no. 5 '*-myen taum-kwa kath-ha*' (if … as follows) and the bundle '*-l philyo-ka iss-ta*' (it is necessary to), where '*-myen*' and '*-l*' are combined to verbs, are both used in academic writing to present examples, tables, or figures and to state the reason for conducting a research, respectively. Neither of these bundles corresponds to word boundaries; nonetheless, both have a distinct meaning as a unit in academic writing.

Table 3 3-*ecel* sequences and 5-morpheme sequences extracted from academic prose

	3-*ecel* sequences		5-Morpheme sequences	
	Form	Frequency	Form	Frequency
1	*hal swu issta* (할 수 있다) do/say.ADNZ way exist.DCL	1157.9	*ha-l swu iss-ta* (할 수 있다) do/say-ADNZ way exist-DCL	1157.9
2	*pol swu issta* (볼 수 있다) see/judge.ADNZ way exist.DCL	705.2	*po-l swu iss-ta* (볼 수 있다) see/judge-ADNZ way exist-DCL	705.2
3	*al swu issta* (알 수 있다) know.ADNZ way exist.DCL	647.5	*al-l swu iss-ta* (알 수 있다) know-ADNZ way exist-DCL	647.5
4	*swu issul kesita* (수 있을 것이다) way exist.ADNZthing.be.DCL	318.3	*ul al-l swu iss-* (을 알 수 있−) OM knou-ADNZ exist	628.3
5	*hwakinhal swu issta* (확인할 수 있다) identify.ADNZ way exist.DCL	257.9	*-myen taum-kwa kath-ta* (-면 다음과 같다) ADVZ next-with same-DCL	439.0
6	*swu issta i* (수 있다 이) way exist.DCL this	175.6	*iss-ul kes-i-ta* (있을 것이다) exist-ADNZ thing-be-DCL	378.7
7	*swu issta kulena* (수 있다 그러나) way exist.DCL however	142.7	*-i-la ha-l swu* (-이라 할 수) be-QU say-ADNZ way	375.9
8	*pol swu issnun* (볼 수 있는) see/judge.ADNZ way exist.ADNZ	134.4	*-la ha-l swu iss-* (-라 할 수 있-) QU say-ADNZ way exist	373.2
9	*tul swu issta* (들 수 있다) give.ADNZ way exist.DCL	129.0	*-l swu iss-ul kes* (-ㄹ 수 있을 것) ADNZ way exist-ADNZ thing	370.4
10	*kesul al swu* (것을 알 수) thing.OM know.ADNZ way	123.5	*swu iss-ul kes-i-* (수 있을 것이-) way exist-ADNZ thing-be	351.2
11	*issumul al swu* (있음을 알 수) exist.NOMZ know.ADNZ way	120.7	*-um-ul al-l swu* (- 을 알 수) NOMZ-OM know-ADNZ way	321.0
12	*swu issta ilehan* (수 있다 이러한) way exist.DCL be like this	109.8	*-i-lako ha-l swu* (-이라고 할 수) be-QU say-ADNZ way	321.0
13	*swu issta ttohan* (수 있다 또한) way exist.DCL also	101.5	*ul poi-e cwu-nta* (을 보여 준다) OM show-INF AUX-DCL	296.3
14	*swu issta ttalase* (수 있다 따라서) way exist.DCL thus	98.8	*-lako ha-l swu iss-* (-라고 할 수 있-) QU say-ADNZ way exist	293.6
15	*kesulo pol swu* (것으로 볼 수) thing.as see/judge.ADNZ way	96.0	*-l philyo-ka iss-ta* (-ㄹ 필요가 있다) ADNZ need-SM exist-DCL	282.6
16	*kesila hal swu* (것이라 할 수) thing.be.QUsay.ADNZ way	96.0	*iss-nun kes-i-ta* (있는 것이다) exist-ADNZ thing-be-DCL	268.9
17	*hal swu issnun* (할 수 있는) do/say way exist.ADNZ	93.3	*hwakinha-l swu iss-ta* (확인할 수 있다) identify-ADNZ way exist-DCL	257.9
18	*swu isski ttaymwunita* (수 있기 때문이다) way exist.NOMZreason.be.DCL	85.1	*-ess-ten kes-i-ta* (-었던 것이다) PST-ADNZ thing-be-DCL	255.2

Table 3 continued

	3-*ecel* sequences		5-Morpheme sequences	
	Form	Frequency	Form	Frequency
19	*swu issta thukhi* (수 있다 특히) way exist.DCL especially	82.3	*-ko iss-nun kes-i-* (-고 있는 것이-) GER exist-ADNZ thing-be	244.2
20	*cimcakhal swu issta* (짐작할 수 있다) guess.ADNZ way exist.DCL	82.3	*-l swu iss-ta i* (-ㄹ 수 있다 이) ADNZ way exist-DCL this	222.3

(6) −면 다음과 같다 *-myen taum-kwa kath-ta* [ADVZ next-with same-DCL]

a. 이를 표로 나타내면 다음과 같다.

This is shown in the following table.

b. 본문에서 예를 들면 다음과 같다.

This is shown in the following example.

c. 그 틀을 제시하면 다음과 같다.

The outline is as follows.

(7) −ㄹ 필요가 있다 *-l philyo-ka iss-ta* [ADNZ need-SM exist-DCL]

a. 이들 간의 관계는 더 정밀히 살펴볼 필요가 있다.

We need to have a closer look into the relation between these.

b. 여기서 우리는 시민사회가 무엇인지에 대해 살펴볼 필요가 있다.

We now need to examine what a civil society really means.

c. 슬픔의 종류를 이해하기 위해 감정의 변이들에 대해서도 파악할 필요가 있다.

In order to understand what sadness is, it is necessary to comprehend the emotional variations.

As shown in (6) and (7), morphemic bundles that begin with an inflectional ending (*emi*) must necessarily combine with verbs. Furthermore, it can be seen that these verbs share certain semantic similarities. The bundle '*-myen taum-kwa kath-ha*' (if … as follows) usually comes after verbs such as *nathanay-* (show), *ceysiha-* (present), (*yelul*) *tul-* (to take one example). These patterns intersect at the semantic node of 'showing' (an example, a table, a figure). As for the bundle '*-l philyo-ka iss-ta*' (it is necessary to), it usually follows verbs such as '*salphyebo-*' (examine), '*kemthoha-*' (investigate), '*phaakha-*' (comprehend), which express the idea of 'examination'.

Morphemic bundles may start or end with a bound morpheme. In this case, they play, as semantic units, a crucial role in determining the new meaning produced by their combination with the word coming before or after. These semantic patterns cannot be retrieved when performing *ecel*-based analyses.

4 A corpus-driven analysis of morphemic bundle

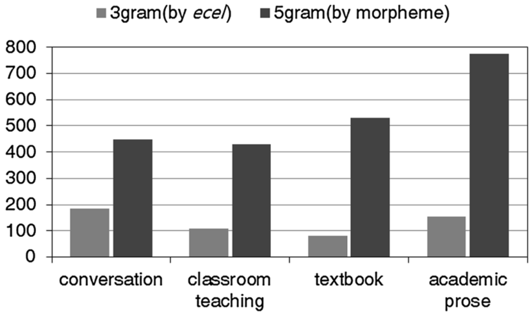
4.1 Quantitative analysis

4.1.1 Overall frequency

Figure 1 shows the number and distribution of 3-g lexical bundles (left-hand bar) and 5-g morphemic bundles (right-hand bar) extracted from the two corpora used for this study. As mentioned earlier, 3-g lexical bundles are identified by the *ecel* unit as applied in Biber et al. (2010) and Lee (2012), whereas 5-g morphemic bundles are identified by the morpheme unit. The figure shows a striking difference in numbers between *ecel*-based and morpheme-based analyses of bundles. On the basis of previous research (Biber et al. 2010), the number of 3-*ecel* (or 3-word) bundles by registers ranges from 80 to 185 items approximately, whereas that of 5-morpheme bundles, which corresponds to the results of our own investigation, ranges from 430 to 775 items. This proves that formulaic expressions are being used in spoken and written Korean approximately five times more than what was concluded in previous studies.

Biber et al. ((1999), Ch. 13) have found among 4-word lexical bundles recurring more than 10 times per 1 million words in English, over 450 items in conversation and more than 300 in academic prose. Despite the different cutoffs, if we compare their results with ours, we can argue that Korean speakers tend to use formulaic expressions as frequently as English speakers. The use of prefabricated expressions may vary from register to register in both languages; nonetheless, this quantitative similarity could indicate the universality of prefabricated chunks in language use. The reason why previous studies, such as Biber et al. (2010), could not find similar results, that is, a satisfactory amount of lexical bundles in Korean despite the use of relatively larger corpora, is that these works were based on word unit and 3-word sequences. As we demonstrated earlier, the Korean language consists of a complex agglutinative system in which the structures and forms of words vary

Fig. 1 Number of lexical (3-*ecel*) and morphemic (5-morpheme) bundles by registers



according to their sentence function; therefore, morphemes turned out to be the most appropriate unit to detect formulaic expressions.

4.1.2 Frequency across four registers (conversation, classroom teaching, textbook, academic prose)

It is generally agreed that formulaic expressions in English are more varied and more frequent in speech than in writing (Biber et al. 1999; Leech 2000; Ellis et al. 2008). This phenomenon has usually been explained from a cognitive perspective: according to Ellis et al. (2008: 376), the more frequent use of formulaic expressions in spoken English compared to written English results from the speaker's harder cognitive effort in speech as “speech is constructed in real time and this imposes greater working memory demands compared to writing, hence the greater need to rely on formulas”. As for Biber et al. (2004: 397), they have attributed the frequent occurrence of lexical bundles in classroom teaching to instructors' needs and limitations; more specifically, they “need to organize and structure discourse that is at once informational and involved, and is produced with real-time production constraints.”

To compare the number of lexical bundles in Korean with those in English across registers, we observed how the bundles are distributed in Korean and English in four registers, as demonstrated by Biber et al. (2004). To assess the use of lexical bundles in Korean, we have observed how they are distributed across four registers: conversation, classroom teaching, textbook, and academic prose (Fig. 2), which are those used by Biber et al. (2004) for their distribution analysis of English lexical bundles. The distribution of Korean morphemic bundles shown in Fig. 2 contrasts strikingly with the results yielded in studies on English. Generally speaking, the figure shows that lexical (morphemic?) bundles in Korean are more frequently found in writing than in speech. More specifically, the register that presents the highest frequency of morphemic bundles is the academic prose, though they are also relatively frequent in the written register of textbook. This is in sharp contrast with the results yielded by studies on English (Biber et al. 1999, ch. 13, 2004; Ellis et al. 2008: 376). In particular, Biber et al. (2004) have found that English lexical bundles occur more frequently in speech than in writing, and the descending order of

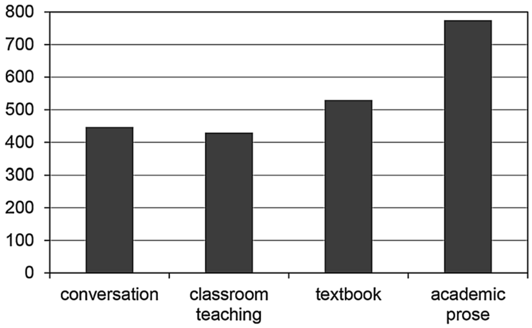


Fig. 2 Number of Korean morphemic bundles across registers

frequency by registers is as follows: classroom teaching > conversation > textbook > academic prose. Korean morphemic bundles rank in the contrary order, academic prose being the register with the most frequent occurrence of bundles and classroom teaching being the one with the lowest frequency.

To elucidate this distribution of morphemic bundles which is characteristic of the Korean language, it is necessary to analyze the factors influencing 'formulaicity' from various points of view. Indeed, we argue in this paper that the formulaicity of language is not merely determined by cognitive factors as claimed by Ellis et al. (2008) or Biber et al. (2004) but can also be influenced by factors such as the fixedness of word order and the formality of academic writing. The reason for such differences in distribution across registers between English and Korean can be explained by the determining factors in language formulaicity that differ from language to language.

Contributing factors to the formulaicity of language can be divided into two types, namely language-internal factors and language-external factors. Grammatical properties, such as the fixedness of word order and the morphemic sequences, constitute one of the most significant language-internal factors that influence formulaicity. For instance, Korean has, on one hand, a highly developed case marking system and, on the other, less restrictions on word order compared to inflectional languages; in consequence, relatively fewer bundles are found if the language is analyzed by the conventional word unit, i.e., a unit defined by boundary spaces. From a morphological perspective, being a typical agglutinative language, Korean comprises various bound morphemes that can combine with each other but only in a rigorously regular order. As a result of such a feature, morphemic bundles appear to be quite frequent, but only when taking the morpheme as a unit. As shown in example (8), the morphemic bundle '-음-에-도 불구하-고 *um-ey-to pulkuha-ko* ('despite')' consists of five bound (i.e., dependent) morphemes, the first three of which can only occur in this particular grammatical order, that is, 'nominalizer (음, *um*) + *cosa* 1 (에, *ey*) + *cosa* 2 (도, *to*)'. In example (9), likewise, the morphemic bundle '-었-음-을 보이-어 *ess-um-ul poi-e* ('shows that + past tense')' can only be constructed in this order: 'past tense marking *emi* + nominalizer + *cosa*'.

(8)	-음-에-도	불구하-고
	um-ey-to	*pulkuha-ko*
	NOMZ-at-also	regardless of-CONJ
	'in spite of –ing'	
(9)	-었-음-을	보이-어
	ess-um-ul	*poi-e*
	PST-NOMZ-OM	show-INF
	'show that (PST)…'	

Language-external factors contributing to formulaicity range from various sociocultural aspects to psychological aspects. From Fig. 2, for example, we have concluded that the sociocultural factor of 'academicity' (which can be represented by: [+academic]) plays a crucial role in the use of formulaic expressions in the

Korean language. The academic articles that we analyzed in the present study are the papers of scholarly writers and have been published in peer-reviewed journals; in other words, they follow the particular styles, form, and content required by the academic community. As a result, the level of fixedness in Korean academic prose is very high compared to the other registers (i.e., textbook and classroom teaching). Academicity in Korean is one of the most influential sociocultural factors that influences the use of formulaic expressions (nonetheless, it is necessary to investigate more registers to draw a more general conclusion). Furthermore, this can also indicate that sociocultural factors, such as academicity, are more influential in using formulaic expressions in Korean than they do in English, where psychological/cognitive factors are prevailing.

It appears that the most frequent formulaic expressions are generally used to convey the author's viewpoint or stance in a particular academic context. Typical examples are '알 수 있다 *al swu issta*', '볼 수 있다 *pol swu issta*', and '-이라고 할 수 있다 *-ilako hal swu issta*', which literally mean '(there is) a way to know', '(there is) a way to see', and '(there is) a way to say', respectively. These bundles suggest the writer's points of view and/or attitude when interpreting results and share a common functional meaning that expresses the possibility ('it could be realized/said/seen that').

Finally, it seems necessary to discuss the low frequency of morphemic bundles in spoken Korean. This is a distinguishing feature of Korean speech which is in total contrast with English and can be interpreted in two ways. First, on the basis of language-external factors, spoken Korean is subject to less psychological pressure and cognitive constraints in comparison with written Korean. Second, spoken Korean presents more cases of grammatical morphemes omission and heavily relies on the speaker's own speech style, while in written Korean, morphemic bundles usually consist of consistent grammatical morphemes. Therefore, adopting a morpheme-based approach can only result in the extraction of more varied and more frequent bundles in written registers than in spoken discourses.

As discussed so far, Korean morphemic bundles show considerable differences in usage from English lexical bundles. First of all, the linguistic unit (morpheme vs word) adopted for the analysis decisively affects the total frequency of the bundles retrieved. Also, the distribution of the bundles across registers results from both language-external and language-internal factors that are proper to the Korean language. Indeed, while multi-word sequences are as much used in Korean as in English, our research analysis has shown that the distribution of formulaic bundles by registers in Korean is markedly different from that in English.

4.2 Qualitative analysis: discourse functions

By analyzing quantitatively the Korean MWUs, we have tried to determine their language-universal features as well as language-specific characteristics. We now turn to their qualitative analysis by examining their discourse functions. English and Korean having different typological features, discourse functions in the two languages are expected to show distinguished characteristics. This, we believe, will support our methodological approach in the analysis of the Korean language.

Previous studies on lexical bundles (Biber et al. 2004, 2009; Cortes 2004, 2006; Wray and Perkins 2000; Schmitt 2005; Hyland 2008) have generally classified discourse functions into three or four categories. Although these categories are defined in various terms, some are commonly designated under the terms ‘discourse organizers’/‘text organizers’ and ‘stance features’/‘stance expressions’. Drawing on the categorizations made in previous research, we classify the discourse functions of Korean morphemic bundles into three types (1) stance bundles, (2) discourse organizers, and (3) content bundles. Examples of stance bundles are provided in example (10).

(10) a. 할 수 있다.
hal swu issta
‘be able to do’

b. 볼 수 있다.
pol swu issta
‘be able to see’

c. 알 수 있다.
al swu iss ta
‘be able to know’

d. 그럴 수도 있-
kulel swuto iss-
‘it could be’

e. -을 필요가 있다.
-ul philyoka issta
‘need/have to’

f. -인 줄 알았-
in cwul alass-
‘I thought perhaps’

Discourse organizers are mainly used to introduce new topics or to provide examples. In Korean, these are generally found at sentence boundaries, as demonstrated in example (11).

(11) a. –면 다음과 같다.
-myen taumkwa kathta
‘is as follows’

b. -는 것이 아니라
-nun kesi anila
‘it is not … but’

c. -다. 이와 같은
-ta. iwa kathun
‘… . Similarly’

d. -는 거야. 그래서
-nun keya. kulayse
‘it is… . So’

e. 같은 경우에는
kathun kyengwueynun
‘in this case’

f. 아니 그게 아니라
ani kukey anila
‘what I mean is

Content bundles are used to convey concrete lexical meanings and include words such as ‘보이다 poita (to be seen)’, ‘생각하다 sayngkakhata (to think)’, ‘때문 ttaymwun (because)’, ‘일반적 ilpanc*ek* (general)’, ‘경우 kyengwu (case)’, ‘하나 *hana* (one)’, as shown in example (12).

(12) a. 잘 보여 준다.
cal poye cwunta
‘be a mirror of’

b. 있기 때문이다.
isski ttaymwunita
‘it is because’

c. -는 것이 일반적이-
-nun kesi ilpanceki-
‘it generally is… ’

d. 생각을 했어.
sayngkakul haysse
‘I thought’

e. –는 경우가 많다.
-nun kyengwuka manhta
‘in many cases’

f. 중의 하나이다.
cunguy hanaita
‘it is one… among… ’

When a morphemic bundle has more than one discourse function, it falls under the category of the dominant function. For instance, the morphemic bundle ‘볼 수

있다 *pol swu issta*' can function either as a content bundle or as a stance bundle but falls under the stance category, because contexts showed that it is primarily used as a stance expression.

Not only do discourse functions of Korean morphemic bundles differ considerably from register to register, as shown in Fig. 3, but their distribution also shows substantial differences from those presented by Biber et al. (2004: 396) for English.

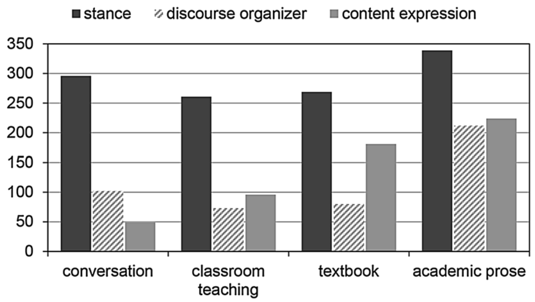
Figure 3 shows that overall, the highest frequency of stance bundles occurs in academic prose and conversation. This significantly differs from the case of English, where stance bundles are most common in classroom teaching and conversation (Biber et al. 2004: 26 and following).

Just as stance bundles, discourse organizers are also more frequently used in academic prose and conversation, with, nonetheless, a lower frequency rate. Although the discourse function categories show the highest frequencies of morphemic bundles in the same registers, these bundles present different patterns. For example, bundles containing the phrase of '이와 같은 *iwa kathun* ('like this')' mainly occur in academic prose, whereas bundles containing '-게 아니라 *key anila* ('it is not … but …')' or '-는 거야 *nun keya* ('it is…that…')' are mostly found in conversation.

The register that shows the highest frequency of content bundles is academic prose, followed by textbook. Content bundles usually contain words like '때문 *ttaymwun* ('because')', '대상 *taysang* ('object')', '보이다 *poita* ('to be seen')', '알려지다 *allyecita* ('to be known')', and '나타나다 *nathanata* ('to appear')'. These are used by the writer to provide supporting references and data for their arguments and to review or evaluate other studies.

There are two noteworthy features in Fig. 3. First of all, morphemic bundles functioning as stance expressions in Korean show the highest frequency in all registers. In contrast, stance bundles in English are the most frequent in spoken registers, while in written registers, referential bundles occur more often. Second, academic prose is the register that presents the most frequent use of morphemic bundles, regardless of their category. In Korean, linguistic formulaicity can be seen as a prerequisite of academic discourse, which greatly contributes to the use of morphemic bundles. This factor prevails over the cognitive factor that consists of real-time speech production constraints and results in higher frequencies of lexical

Fig. 3 Distribution of morphemic bundles across functional categories



bundles in classroom teaching and conversation, i.e., spoken registers, in the English language.

Figure 4 shows the percentage of the discourse functions within each register. Leaving aside the fact that they show the highest frequency in all four registers, the stance bundles' ratios can be ranked in the following descending order: conversation > classroom teaching > textbook > academic prose. This clearly shows that stance bundles tend to be used more in casual discourse and less in academic types of discourse. Approximately 70 % of morphemic bundles express a speaker's attitude in conversation, while only about 45 % convey a writer's opinion in academic prose, despite the fact that the stance bundles are the most frequent functional category. In fact, this is related to the stylistic features of academic prose, that is to say, discourse organizers and content expressions also play crucial roles in academic prose, thereby occurring more frequently in this register than the others. Academic writers need to use more discourse organizers and content expressions to advance strong and consistent ideas and tend to employ more technical and precise expressions in their writing. Consequently, the ratio of stance bundles decreases in academic prose in comparison with the other two categories.

Academic prose put aside, the ratios of content expressions, ranked in descending order, are as follows: textbook > classroom teaching > conversation. This means the more academic and written the register gets, the more the use of morphemic bundles, functioning as content expressions, increases. The [+written] and [+academic] features found in textbooks and academic prose are closely related to content bundles as they convey propositional meaning.

In sum, the discourse functions of morphemic bundles across registers in Korean show considerable differences in frequency and distribution from those in English. In Korean, stance bundles are the most common among the three discourse functions in all registers. This differs drastically from the findings on English yielded by other research (Biber et al. 1999, 2004). In other words, morphemic bundles in Korean are mostly used to express the speaker's attitude or the writer's stance. In particular, Korean academic prose has its own unique way of using morphemic bundles, as it is extremely formal and requires using distinct discourse function patterns. Our analysis shows that lexical bundles may be common in all

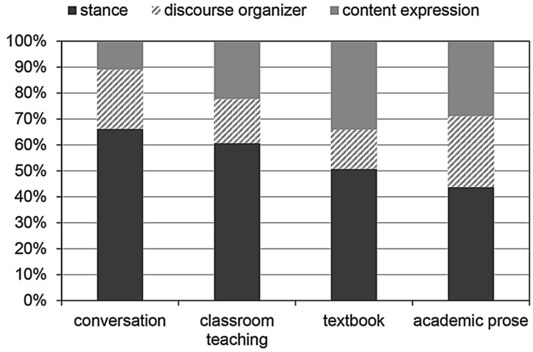


Fig. 4 Ratios of morphemic bundles for each functional category across registers

languages; their discourse functions are, nonetheless, language-specific as their use across registers depends on the conventions of a given language community.

5 Implication for Korean lexicography

Since the publication of Yonsei Korean Dictionary (1998) (hereafter, YKD), Korean lexicography has mostly produced corpus-based dictionaries and even later dictionaries, such as the Korea University Korean Dictionary (2009) (hereafter, KUKD), essentially consist in selecting headwords from large-scale corpora and arranging their various meanings according to frequency. It seems that the extraction of semantic units and dictionary compilation from a corpus-driven approach have not been sufficiently researched in Korean lexicography and corpus linguistics. In most cases, the MWUs that are greater than the phrasal unit are mainly described in subentries in Korean dictionaries, being subordinated to the main entry of the word unit. However, as Sinclair pointed out, "most normal text is largely constructed through the idiom principle" (1991: 113); in other words, the list of phrase units, which are widely used in communication, should thus be widely extracted. As a result, the extraction of "morpheme-based lexical bundles" in Korean can overcome two types of issues.

First, the inclusive extraction and the refinement of the morphemic bundle list mean that the lexicographer can select MWU headwords from an objective frequency list. Korean dictionaries feature as many idioms and collocations as lexicographers were able to collect. However, it is rather unlikely that they would include MWUs from both written and spoken Korean. For most dictionaries, lexicographers have used existing dictionaries and corpora to select and describe MWUs. Such a method greatly relies on their intuition and subjectivity. According to Nesselhauf, research on collocations can adopt either a "frequency-based approach" or a "phraseological approach" (2005: 12). If the former is entirely based on frequency and statistics, the latter approach rather focuses on semantic compositionality or grammatical features. Korean lexicography, thus far, has adopted the phraseological approach. However, by adopting a frequency-based approach, it would enable a greater inclusion of high-frequency MWUs in Korean dictionaries, in addition to the standard phrases selected on the basis of morpho-syntactic fixedness or semantic compositionality.

Second, as Korean dictionaries overlook the importance of MWUs' communicative function, MWUs are not given the status of headword and tend to be insufficiently, if not poorly, described. Once morpheme-based lexical bundles are extracted from texts of various genres and registers and granted the status of headword, they can be sufficiently defined as they become equally important as word unit. Semantic units that are greater than phrasal units are usually not given the status of headwords but, instead, are described in subentries, because Korean dictionaries, just as most language dictionaries, follow the 'one-word-one-headword principle' (Svensén 1993: 208). This principle implies that the grammatical category of a headword should be determined on the basis of a word unit, but it can, nonetheless, hinder the dictionary user's search as it depreciates the value of important MWUs to the benefit of less important words. The following examines a few representative cases.

The 3-g formulaic expression '-ㄹ 수 없- -*l swu eps*-' (cannot), which conveys the possibility/ability negation, occupies the 115th rank in the decreasing order of frequency in the Sejong Balanced Corpus, being used 4.7 times more than '못 *mos*' and 1.7 times more than '못하- *mosha*-' which both assume a similar function. Nevertheless, only *mos* and *mosha* are included in the dictionary as headwords, while the morphemic bundle '-*l swu eps*-' is merely described as a grammatical pattern within brackets under the headword '없- *eps*-' (be not, do not exist) in the 6th and 5th subentries of the online *Korean Standard Unabridged Dictionary* (1999) (hereafter, KSUD) and KUKD, respectively, as shown in the examples below.

(13) a. *epsta* (verbal adj.) 1. A state in which people, animals, things do not exist (…) 6. (mainly used in the '-*l swu epsta*' form) when something is not possible. I cannot believe what you said./ I cannot think anything anymore. (KSUD)

b. *epsta* (verbal adj.) 1. When (something, in a given place,) does not appear or exist (…) 5. Used in '-*ul swu(ka) epsta*' to express the impossibility to do something or the intention of not doing something. (KUKD)

Most dictionaries do not sufficiently explain the ambiguity of the morphemic bundle '-*l swu eps*-' (cannot), namely the negation of either the possibility or the ability to do something. Not only does its subordination(attachment) to the headword 'epsta' make it difficult for the dictionary user to look it up, but its definition also lacks detailed information on usage. For instance, the morphemic bundle '-*l swu eps*-' is much more common in written than spoken language. As a matter of fact, Nam (2015: 101) has analyzed a Korean balanced corpus and found that this morphemic bundle appears three times more often in writing than it does in speaking.

As demonstrated by Biber et al. (1999, 2004), morpheme-based lexical bundles are semantic units that differ across genres and registers. In that, it can be said that they are similar to words. Nevertheless, Sinclair has argued that the 'ultimate dictionary' (Sinclair et al. 2004: xxiv) does not regard words as basic semantic units but rather, meaning results from the combination of particular collocates (2004: 148). His argument remains consistent with the notions of 'lexical items' (Sinclair 1996), 'lexical approach' (Lewis 1993), 'idiom principle' (Sinclair 1991), and has been partly embraced in the lexicographical process since the COBUILD project which consists in corpus-assisted lexicography. However, even a comprehensive extraction of headwords which would reflect each particular language's typological characteristics still raises a number of issues. Frequency lists of MWUs have established that adopting a morpheme-based approach can constitute, combined with the frequency lists of the whole of Korean vocabulary, fundamental and valuable data in the selection of headwords as semantic units.

6 Conclusion: pedagogical applications and lexicography

This study has suggested a morpheme-based approach to identify formulaic expressions in Korean and has discussed the methodological differences compared to previous studies on English. We extracted 5-g morphemic sequences which include *cosa* (particles) and *emi* (inflectional endings). One important feature of our methodology is that we considered the bound morpheme *cosa* and *emi* as minimal syntactic units, which led us to rename lexical bundles 'morphemic bundles'. A key objective of our work was to determine, on one hand, whether the use of formulaic expressions is common in language families other than the Indo-European family and, on the other, whether the criteria for measuring formulaicity differ from language to language. If "all languages are patterned" (Hunston and Francis 1999: 14), as Sinclair and other corpus linguists have claimed, patterned MWUs should also be prevalent in Korean.

This study has provided some interesting findings about the frequency and distribution of formulaic expressions in Korean and has presented a new methodology to identify morphemic bundles. The results show the prevalence of formulaic expressions in Korean, thus confirming that Korean is not an exception. Lexical bundles were rarely found in previous research on Korean. However, by applying a methodology based on morpheme unit, which is proper to agglutinative languages, it appears that Korean speakers and writers use a considerable number of formulaic bundles, almost as much as English speakers do. To discuss the formulaicity of a language which is not from the same language family as English, we have proposed a different approach which takes into account the typological characteristics of the language at issue. Linguistic formulaicity may be described from either morphemic or syntactic perspectives. The basic unit to compute n-grams can be a 'word', or a unit above or below a 'word', because each language has different grammatical characteristics, and the status of a word is language-specific. However, further investigation is required to determine which between the morphological and syntactical points of view is more appropriate to analyze language patterns.

In furtherance of our analysis, we have examined different patterns of formulaic expressions across registers in different languages. Frequencies and distributions of morphemic bundles extracted from four registers have showed interesting patterns. That is, morphemic bundles in Korean occur most frequently in academic prose and textbooks. This pattern of frequency and distribution in the registers may indicate that formulaic expressions in Korean are affected by the features of [+academic] and [+written]. This is in total contrast to the findings reported in previous studies on English (Biber et al. 1999; Leech 2000; Ellis et al. 2008). In discussing the factors that affect the use of formulaic language, we have observed that the linguistic conventions within the Korean academic community are highly influential, whereas in English, cognitive factors, such as real-time speech constraints, are more decisive in using bundle expressions. Also, these contributing factors to language formulaicity have been discussed from language-internal and language-external perspectives.

The study findings can be applied to various fields of applied linguistics. Traditionally, in Korean dictionaries and pedagogical materials for teaching Korean as a foreign language are present morpheme unit-based patterns as important grammatical patterns. These patterns generally begin with bound morphemes such as particles and inflectional endings, as seen in ‘N-에 관한 *N-ey kwanhan* (‘about N’), V-고자 한다 *V-koca handa* (‘try to V’), V-지 않으면 안 되- V-ci anhumyen an toy- (‘have to V’)’ etc. Korean morphemic bundles have already been proved to be significant elements, in compiling Korean learners’ dictionaries for foreigners (as sub-lemma or usage boxes for instance), and in ‘grammar focus/note’ (e.g., Yonsei University Institute, 2004) or ‘grammar and expressions’ (e.g., in Seoul National University Institute 2007) sections. Most of the high-frequency morphemic bundles identified in this study are supplied with useful lists of expressions in dictionaries and teaching materials. However, these patterns could not be identified properly by applying an approach based on word units, hence the need for a novel approach that is appropriate to the characteristics of Korean, i.e., a morpheme-based methodology.

Morphemic bundles do not only constitute educational resources for learners of Korean but can also be used in lexicography in the following two ways. First of all, these morphemic bundles should be included in the dictionary as headwords. Thus far, headwords have mainly consisted of words and grammatical morphemes; the only phrases that are defined in the dictionary as subentries are those whose overall meanings differ from the meanings of their individual components, as it is the case for proverbs and idioms. This study presented the cases of ‘-(*u*)*l philyoka issta*; -(*u*) *myen taumkwa kathta*’ which appear with high frequency in academic texts. This type of morphemic bundles should be included in the dictionary at least as subentries. Nonetheless, as semantic units, it is preferable to grant them the status of headwords. The inclusion in the dictionary of the MWUs that function as semantic units in particular genres can help learners of Korean understand and use appropriately the morphemic bundles; in other words, it can contribute to achieve fluency in the Korean language. Second, there should be more pragmatic information concerning morphemic bundles within the dictionary entries. Although some dictionaries do present information on the headword in discourse situation, this is far from being systematic. Rather, this type of information is considered optional. Furthermore, its content tends to depend on the lexicographer’s intuition rather than be based on corpus analysis. By analyzing the various aspects of morphemic bundles across genres and registers, it becomes possible to describe these units of communication objectively and thoroughly. On the basis of this study’s results, if morphemic bundles were included in the dictionary as headwords and their entries contained information on their usage frequencies and discourse functions according to registers, it will greatly help increase the language skills of dictionary users.

Acknowledgments This work was supported by the National Research Foundation of Korea Grant funded by the Korean Government (NRF-2014S1A5B4038405).

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

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

Biber, Douglas, Youjin Kim, and Nicole Tracy-ventura. 2010. A corpus-driven approach to comparative phraseology: lexical bundles in English, Spanish, and Korean. *Korean Linguistics* 17: 75–94.

Bloomfield, Leonard. 1935. *Language*. London: George Allen & Unwin.

Choi, Jun, Hyunju Song, and Kilim Nam. 2010. Formulaic expressions in Korean. *Discourse and Cognition* 16(3): 163–190.

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

Cortes, Viviana. 2006. Teaching lexical bundles in the disciplines: an example from a writing intensive history class. *Linguistics and Education* 17(4): 391–406.

Cwu, Sikyeng. 1910. *Kwuke Mwunpep (A Korean Grammar)*. Seoul: Pakmun Sekwan.

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

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

Foster, Pauline. 2001. Rules and routines: a consideration of their role in the task-based language production of native and non-native speakers. In *Researching pedagogic tasks: second language learning, teaching, and testing*, ed. Martin Bygate, Peter Skehan, and Merrill Swain, 75–97. Harlow: Longman.

Hunston, Susan, and Gill Francis. 2000. *Pattern grammar: a corpus-driven approach to the lexical grammar of English*. Amsterdam: John Benjamins.

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

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

Lee, Changsoo. 2012. Using lexical bundle analysis as discovery tool for corpus-based translation research. *Perspectives Studies in Translatiology*. doi:10.1080/0907676X.2012.657655.

Leech, Geoffrey. 2000. Grammars of spoken English: new outcomes of corpus-oriented research. *Language Learning* 50(4): 675–724.

Lewis, Michael. 1993. *The lexical approach*. Hove: Language Teaching Publications.

Martinet, André. 1962. *A functional view of language*. London: Oxford University Press.

Nam, Kilim. 2015. A corpus linguistical approach on the meanings and the discourse functions of '-l su eobs-'. *Textlinguistics* 38: 93–120.

Nesselhauf, Nadja. 2005. *Collocations in a learner corpus*. Amsterdam: John Benjamins.

Pak, Sungpin. 1935. *Cosenehak (Korean linguistics)*. Seoul: Chosen.e Yenkwuhoy.

Sapir, Edward. 1921. *Language*. New York: Harcourt, Brace & World Inc.

Schmitt, Norbert. 2005. Formulaic language: fixed and varied. *Estudios de Lingüística Inglesa Aplicada* 6: 13–39.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Sinclair, John. 1996. The search for units of meaning. *Textus* 9/1: 75–106.

Sinclair, John, Susan Jones, and Robert Daley. 2004. *English lexical studies: the OSTI report*. London: Continumm.

Svensén, Bo. 1993. *Practical lexicography: principles and methods of dictionary-making*. Oxford: Oxford University Press.

Sohn, Ho-Min. 1999. *The Korean language*. Cambridge: Cambridge University Press.

Wray, Alison, and Michael R. Perkins. 2000. The functions of formulaic language: an integrated model. *Language and Communication* 20(1): 1–28.

Dictionary

National Institute of Korean Language. 1999. *Phyo-cwun-kwuk-e-tay-sa-cen/Korean standard unabridged dictionary*. Seoul: Doosan Dong-A.

Research Institute of Korean Studies, Korea University. 2009. *Korea University Korean dictionary*. Seoul: Research Institute of Korean Studies, Korea University.

Yonsei Institute of Language and Information Studies. 1998. *Yonsei Korean Dictionary*. Seoul: Doosan Dong-A.