Formulaic expressions of ‘thing’ in spoken and written registers: A corpus study of bound nouns *kes* and *ke* in Korean

Recent studies on corpus linguistics indicate that formulaic expressions (recurrent sequences of words or strings, e.g. *I don’t know if*) play an important role in both spoken and written discourse (Biber & Barbieri 2007). By using computational technologies, researchers have provided empirical evidence for the most common recurrent sequences and their discourse functions in a given register. However, few studies have explored formulaic sequences in agglutinative languages, where a string of affixes are attached to a root. The study of lexical bundles in Biber et al (2009) argues that compared to English, Korean has few recurrent sequences of words. However, the number of Korean bundles in Biber et al is based on orthographic words and does not represent morphological characteristics of Korean. For instance, many verbal suffixes in Korean which are expressed in words in English (e.g., *ka* ‘go’-*myen*, ‘if (he) goes’, *ka* ‘go’-*nikka* ‘because (he) goes’) are excluded from the bundles.[1] Given the rich inflectional morphology in Korean, it is vital to incorporate grammatical morphemes into formulaic sequences. Using a morpheme-based n-gram analysis, the present study examines the distribution and the use of formulaic sequences of bound nouns *kes* ‘thing’ and its spoken variant *ke*.[2]

The corpus for this study consists of 1 million words of written academic texts, 290,000 words of spoken academic corpus (e.g. lectures, public speech), and 220,000 words of conversation. The corpus is sampled from across 48 texts drawn from a sub-component of the 21st C. Sejong corpus. The frequency cut-off used to identify formulaic expressions for this study is based on 5-morpheme sequences (e.g. *iss-ul kes-i-ta*) occurring more than 30 times per million words drawn from at least 5 different texts. (A normalization is used to compare formulaic expressions across corpora of different size.) By employing computer programs which identify every 5–morpheme sequence incorporating *kes/ke* in the corpus, we compare and contrast the lexico-grammatical collocations of the two bound nouns across registers.

The findings show that the formulaic expressions of *kes* and *ke* are strikingly different in terms of their distribution and functions. First, as illustrated in Table 1, *kes* bundles are predominantly used in academic discourse, which is comprised of both written and spoken registers. In contrast, *ke* bundles are prevalent in spoken discourse,

[1] Space is placed in Korean before all words whereas all affixes are attached to the respective stems.

[2] The bound noun *kes*, the most frequent lexical item in Korean, is considered as the most flexible one among all bound nouns in terms of syntactic constructions (Hong 2006).

including conversation and spoken academics. There is not even a single token of *ke* bundle in written registers. Second, the overall type frequency of *ke* bundles (260 different types) is much higher than *kes* bundles (201 types), which indicates that *ke* bundles have more diverse lexico-grammatical collocations than *kes* bundles. Third, the formulaic expressions of *ke* are dominant in stance bundles which express attitudes or assessments of certainty. Stance bundles account for over 70% of the different bundles of *ke* and are predominantly used in sentence-final position (e.g. *-n ke-ya, -n ke-canha, -n ke-ci*). In contrast, *kes* bundles are often used in argument structures and function to signal a referential meaning.

The study indicates that the formulaic expressions of *kes* and *ke* develop to serve the most important communicative needs of a register (Biber & Barbieri 2007, Biber et al 2009). Academic written registers focus on informational communication. As such, *kes* bundles, prevalent in written register, retain the referential meaning of the bound noun (i.e. 'thing' or 'entities') and are often used in NP-based bundles for a referential framing. In sharp contrast, *ke* bundles in spoken registers become grammaticalized into stance markers along with the loss of a referential meaning and a phonetic reduction. In sum, morpheme-based bundles can provide a more detailed account of the distribution and functions of formulaic expressions in languages with inflectional morphology.

Table 1. Distribution of formulaic expressions of *kes* and *ke* across registers

	kes	*ke*
Academic written	160	0
Academic spoken	40	99
Conversation	1	161

(based on 5-morpheme sequences occurring more than 30 times per million words)

References

Biber, D. and F. Barbieri. 2007. Lexical bundles in university spoken and written registers. *English for Special Purposes* 26:263-286.

Biber, D., Y.-J. Kim, and N. Tracy-Ventura. 2009. A Corpus-driven approach to comparative phraseology: lexical bundles in English, Spanish, and Korean. Iwasaki et al (eds.) *Japanes/Korean Linguistics* 17:75-94, CSLI Publication.

Hong, S.-M. 2006. *Uycon myengsa 'kes'-uy sacek yenkwu* (A Diachronic study of the bound noun *kes*). *Emwunlonchong* 44:101-144.