Exploring meaning-sound systematicity in Korean

Hana Jee

Korean Language and Applied Linguistics, School of Education, Language & Psychology

York St John University

Lord Mayor's Walk, York YO31 7EX, UK

h.jee@yorksj.ac.uk

Monica Tamariz

Psychology, School of Social Sciences

Heriot-Watt University

Edinburgh EH14 4AS, UK

M.Tamariz@hw.ac.uk

Richard Shillcock

Psychology, School of Philosophy, Psychology and Language Sciences

The University of Edinburgh

10 Crichton Street, Edinburgh EH8 9AB, UK

rcs@inf.ed.ac.uk

Abstract

Studies of word-level meaning-sound systematicity in English and four other European languages have shown that words that sound similar tend to have similar meanings. The term 'systematicity' in this research tradition is defined as statistically non-arbitrary relations between sub-domains of language, in contrast to the traditionally assumed Saussurian arbitrariness. We explore such systematicity in a typologically distinct language, Korean. We find a relatively high level of systematicity, which we attribute to the method of analysis where we applied Latent Semantic Analysis (LSA) based on *eo-jeols*—sequences of syllable-blocks bounded by spaces in an internet corpus of written Korean. Eo-jeols embody a psychologically realistic spectrum of linguistic structure and influence, compared with previous purely lexically based studies of systematicity. Systematicity was pervasive in our sample of the Korean lexicon—partitioned by word frequency, etymological origin, syllabic constituents (onset, vowel, coda, rhyme), syntactic categories, homonyms, onomatopoeia, and loanwords—suggesting a fundamental basis for systematicity. We explain meaning-sound systematicity in terms of related degrees of cognitive effort in speaking and listening.

Keywords: systematicity, meaning-sound mapping, Korean, least effort

Exploring meaning-sound systematicity in Korean

1. Background

Language and Systematicity

Language is a complex system that consists of interlocking subsystems including phonology, morphology, syntax, semantics and pragmatics (see Freeman & Cameron, 2008). Spoken language takes place in a vulnerable medium. It benefits from robust redundancy within and between the different subsystems: for instance, if noise interferes with the final segments of a word, they may be inferred or even perceptually restored on the basis of the lexicon and the linguistic context within and outside the word (see Clark, 2013). This systematicity has traditionally been studied in terms of explicit rules that capture the relations between *symbols*—phonological features, phonemes, morphemes, syllabic structures, syntactic categories, and the categories of formal semantics and pragmatics. These rules are mainly *within* a subsystem (e.g. phonology, syntax) and to a lesser degree *between* subsystems (e.g. morphosyntax).

More recently, researchers have investigated a second type of systematicity, which is statistically defined and may be just as universal a phenomenon as traditional, symbolic systematicity. In contrast, it is typically *graded*, *partial* and *weaker*, and it obtains between domains that are traditionally assumed to be only arbitrarily related, such as semantics and phonology (Saussure, 1916). These relations are typically discovered as statistical generalizations over the contents of the lexicon of an individual language.

Thus far, researchers have found significant systematic relations between phonology and syntax (see Fitneva, Christiansen, & Monaghan, 2009; Kelly, 1992; Kelly, Morgan, & Demuth, 1996; Monaghan & Christiansen, 2008; Morgan & Demuth, 1996; Reali, Christiansen, & Monaghan, 2003; Vendler, 1968). For example: a monosyllabic word containing the vowel schwa is likely to be a function word in English, Turkish and Mandarin (Shi, Morgan & Allopenna, 1998); verbs in English are more likely to contain front vowels (Sereno, 1994).

An *onomatopoeic* word phonetically reflects an essential feature of a real-world entity or event. For instance, ‘tick-tock’ in English and ‘kachi-kachi’ in Japanese represent the noise of a clock.

In *sound symbolism*, certain segments or phonological features tend to correlate with meanings. For example, bilabial stops (/p/ and /b/) are related to ‘fullness’, and nasal /n/ to ‘nose’ in many languages (Blasi, Wichmann, Hammarstrom, Stadler, & Christiansen, 2016). Such systematicity has observable behavioural consequences; people correctly guess above chance level the contrasting meanings of unknown foreign words (Brown, Black, & Horowitz, 1955; Klank, Huang, & Johnson, 1971; Kunihira, 1971).

Synaesthesia involves cross-modal perceptual effects and is mainly triggered by linguistic stimuli like letters, numbers or punctuation; thus, the letter ‘A’ may be associated with the colour red, ‘B’ with blue, and so on. It is only observed in a small percentage of the population, but non-synaesthetes may also agree on particular letter-colour associations (Simner, Ward, Lanz, Jansari, Noonan, Glover, & Oakley, 2005).

This paper is concerned with the most recent type of systematicity to be researched, the systematic relations between phonology and semantics (Blasi et al., 2016; Dautriche, Mahowald, Bibson, & Piantadosi, 2017; Monaghan, Christiansen, Farmer, & Fitneva, 2010; Monaghan, Shillcock, Christiansen, & Kirby, 2014; Shillcock, Kirby, McDonald, & Brew, 2001; Tamariz, 2008). It is a *distributed* systematicity that has been demonstrated to hold overrepresentative and chiefly monosyllabic, monomorphemic samples of the lexicon: words that sound similar within a language tend to have similar meanings. It exists as an overall correlation—typically very small but statistically significant—between all the pairwise phonological distances between words and all the corresponding pairwise semantic distances between words, for substantial numbers of

words. This differs from sound symbolism, which is concerned with relationships between specific sounds and the meanings of particular words.

Systematicity is thus widespread in one form or another within the representation and processing of language. In general, topographic or structure-preserving mapping is pervasive within the human brain (cf. Thivierge & Marcus, 2007) and may be directly implicated in language processing (cf. Ellison 2013). In the current study we further explore the general nature of meaning-sound systematicity, with specific reference to Korean, a unique language in terms of geography, etymology, and history.

The special case of Korean

We can expect differences in linguistic systematicity between languages. For instance, not all English phonoaesthemes are found in other languages and the vowel contrast /a/-/ɪ/ corresponding to size (/a/ for larger objects and /ɪ/ for smaller objects) differs from language to language depending on whether the distinction relies more on F0 or on F2 (Shinohara & Kawahara, 2010); it is reversed in Vietnamese (Diffloth, 1994).

Most research on systematicity in language has focused on the major international languages. Of the three large-scale studies (Blasi et al., 2016; Dautriche et al., 2017; Lupyan & Dale, 2010), none has included Korean. Korean used to be categorized as a Ural-Altaic language along with Mongolian and Turkish (Ramstedt & Kim, 1979). It is increasingly—but still controversially—considered a *language isolate* (Georg, Michalove, Ramer, & Sidwell, 1999). As an agglutinative language, Korean features polysyllabic words with a complex system of suffixes that express different nuances (Sampson, 1985).

Korean also exhibits vowel harmony with co-occurrence restrictions on certain combinations. While not as strictly applied as in Middle Korean (15th~16thC, Kwon, 2018),

vowel harmony is still observed in Modern Korean phonotactics, onomatopoeia, predicate suffixes (Sohn, 2001) and postpositions (Larsen & Heinz, 2012). Korean vowels are divided into three classes: light vowels like /a/ or /o/ connote ‘light’, ‘bright’, and ‘small’; dark vowels like /ʌ/ and /u/ connote ‘heavy’, ‘dark’, and ‘large’; vowels corresponding to neither, like /i/ and /ɯ/, are referred to as neutral vowels (Kim-Renaud, 1976; Larsen & Heinz, 2012). Within the domain of ideophones (Sung, 2018), the light-dark vowel contrast demonstrates a *language-specific* aspect of sound symbolism: /a/ for ‘small’ and /ʌ/ for ‘large’. In other lexical fields Korean shows the *cross-linguistic* trend as well: /ɪ/ for ‘small’ and /a/ for ‘large’ (Shinohara & Kawahara, 2010).

A final aspect of Korean relevant to this study is the substantial presence of Sino-Korean words—words with Korean pronunciation but originating from the use of Chinese characters. The Korean peninsula has been under Chinese influence for centuries. Although spoken Korean and Chinese are very different, belonging to different language families, written Chinese was used by literate Koreans until hangeul, the Korean orthography, was invented and promulgated after 1446. When Chinese words were introduced, their pronunciation was modified to suit Korean phonology, where neither tones nor the final sound /r/ exist. This process produced a number of homonyms (e.g. 21 different words — 司庫, 史庫, 四苦, 四顧, 死苦, 私考, 事故, 社告, 思考, 思顧, etc. —all pronounced the same: ‘사고’ [sɐ.go]).

In the current study we investigate the level of systematicity—if any—between semantics and phonology in Korean. How similar is Korean, in this respect, to the five European languages—English, Spanish, Dutch, German and French—which are, to our knowledge, the only ones to have been so far studied using a direct comparison between phonological[1] and

[1] Dautriche et al. employ orthographic representations as a proxy for phonology in their large crosslinguistic study.

semantic distances (Dautriche et al., 2017; Monaghan et al., 2014; Shillcock et al., 2001; Tamariz, 2008. See below for further discussion of the Dautriche et al. study.)? We hypothesize that there will be a small but significant correlation between corresponding inter-word distances in meaning and form across a representative part of the Korean lexicon. We then extend this study of meaning-form systematicity to different lexical subsets and to the structure of contemporary Korean. From this vantage point, we then advance an interpretation of the nature of phonological-semantic systematicity in general.

2. Procedure

When English is analysed for semantic-phonological systematicity, the procedure is typically to construct a sample from as many monosyllabic words as possible, starting with the most frequent. (Polysyllabic words frequently exhibit substantial systematicity as a result of shared morphemes and/or extensive etymological relations with other words and so are not such a transparent test of 'pure' systematicity.) The monosyllabic word sample is then vetted for polymorphemic words (e.g. plurals, past tenses) and perhaps for words derived from the same lexeme (e.g. 'saw' and 'seen', 'strong' and 'strength') and etymologically related words (e.g. 'glass' and 'glaze'). The resulting unrelated monomorphemic words ('and', 'of', 'dog', 'cat', 'swamp', 'strain'…) are then entered into the Mantel Test (Mantel, 1967) in which all the pairwise phonological distances between words are tested for a correlation with all the corresponding pairwise semantic distances. A significant correlation indicates the presence of systematicity.

This procedure has different implications in the case of Korean. While our word-like unit of analysis is monosyllabic, it is also an 'eo-jeol': a single syllable-block bounded by a space on each side. In a text corpus, we might find 살 수 있어 ('I can buy it'), consisting of two

monosyllabic eo-jeols followed by one disyllabic eo-jeol. When encountered sequentially, syllable by syllable in the spoken language, the stand-alone progression will be interpreted as 살 (‘flesh’, or ‘arrow’), 살 수 (‘can live’), 살 수 있 (‘can live’), 살 수 있어 (‘I can buy it’)[2]. Corresponding issues of incremental interpretation can be seen in English, for instance, in terms of the Cohort Theory (Marslen-Wilson, 1987), word recognition after acoustic offset (Bard, Shillcock & Altmann, 1988) and the use of prosodic cues in word recognition (Grosjean, 1983). Thus, Korean has substantial homonymy coupled with morphological ambiguity. Languages vary in terms of how much complexity they display in the different linguistic subdomains of phonology, lexis, morphology and syntax; simplicity in one may trade-off against complexity in another. In Korean, we see substantial ambiguity at the level of the syllabic constituents, compromising the extraction of individual eo-jeols, as in the first eo-jeol in our example above. However, this ambiguity does not make our task impossible. Monomorphemic nouns can be extracted unambiguously (there is no plural morpheme, e.g. 개 kjɛ, 곁 kjʌt̄), for instance, as can adjectives and adverbs (e.g. 뒷 tyt̄, 맨 mɛn). Monosyllabic verbs on the other hand, may reflect inflectional categories (e.g. 봐 pʷɐ for 보다 poda ‘to see’, 뺄 ppɛl for 빼다 ppɛdɐ ‘to pull out’, 샌 sɛn for 새다 sɛdɐ ‘to leak’). This difference in fact gives us an interesting comparison between nouns and verbs, as we will see below.

Following previous research (Monaghan et al., 2014; Tamariz, 2008), we measured all the pairwise distances between the phonological representations of all the words in a representative subset of the lexicon, and all the pairwise distances between the corresponding semantic representations of those words. (These distances are respectively feature-edit distances and cosine distances between vectors; see below.) If form perfectly reflected meaning, these two

[2] The reader may confirm this on *Google translate.*

webs of pairwise distances would be isomorphic; we hypothesized that we will find a weak isomorphism, indicating non-arbitrariness. The correlation between these two very long lists of distances indicates the level of form-meaning correlation, or 'systematicity'; an arbitrary relation between word meaning and phonology would yield a non-significant correlation.

2.1. Corpus preparation

Using web scraping, we created a corpus based on Korean internet content reflecting authentic, contemporary language use, and including various styles—spoken and written, short comments and long narration. We collected the data on 22 July 2019 (Jee, *in prep*). The total number of word tokens was 28,858,796.

Although the corpus reflects users with computer access, it consists of spontaneous forms of language, both written and spoken (cf. Johannessen & Guevara, 2011). The subjects discussed tend to reflect casual, everyday life, compared with the special genres of more balanced corpora, such as the British National Corpus (BNC Consortium, 2007).

2.2. Sample

In this first investigation of phono-semantic systematicity in Korean, we extracted monosyllables only from the corpus. These monosyllables are defined as monosyllabic *written eo-jeols* with a space before and after them. Any case-inflected eo-jeols (e.g. 달이 ta.ɾi or 달은 ta.ɾɯn both glossed as 'moon is') were not included in the sample. The phonological form used in the sample was the stand-alone citation form.

Our corpus contained 966 phonologically unique monosyllabic words. We removed 254 items that were misspelled or had unclear meanings, leaving 712 monosyllabic words: 142 CV and 570 CVC. They consisted of various syntactic categories. Analysing only these monosyllables may not reflect the systematicity that is available in the larger vocabulary of

Korean–and generalising to the whole of the Korean vocabulary must be done with caution (although our assumption is that form-meaning systematicity—of morphologically- and syntactically-related kinds—only *increases* when extended beyond the monomorphemic kernel of the lexicon that we have studied). The total number of individual distances between word pairs was thus 253,116 (712 x 711 / 2). The correlation was measured between 253,116 phonological distances and 253,116 semantic distances.

Korean monosyllables account for only some 1% of the total number of word-types in the Korean lexicon, in one estimate (The National Institute of the Korean Language, https://stdict.korean.go.kr/statistic/dicStat.do, accessed on 25 February 2021). Are they representative of the spoken language? Such estimates are contingent on lexicographers' conventions about whether morphological variants should count as 'different words' and decisions as to whether to include technical vocabularies, acronyms, archaic usages and loanwords. However, a similar informal calculation for English yields 2.5%, comparable with the 1% in Korean given the noisiness of such calculations. Our 712-word monosyllabic sample was smaller than the total of 4,767 monosyllabic types in a standard Korean dictionary (The National Institute of the Koren Language, https://stdict.korean.go.kr, accessed on 25 February 2021), reflecting a frequency bias. Our sample size is still bigger than Hahm's (1962, cited in Byun, 2003) list of 200 monosyllables based on semantic frequency. Note, also, that most of our sample words were homonyms; some 95% of the Korean lexicon is ambiguous in this way (Kang, 2005).

A psychologically more relevant measure of the importance of our monosyllabic sample of 712 word-types is the proportion of tokens they account for in real speech. Monaghan et al. cite Baayen, Pipenbrock and Gulikers' (1995) estimate that monosyllabic words account for

some 70% of actual usage in English. Such an estimate is not available from our internet corpus for reasons that we discuss below, but Figure 1 shows the frequency profile of the 712 monosyllabic words in the first column; they occur disproportionately frequently in contemporary internet usage, as we would expect. (We discuss this graph further, below.) We assume, based on Zipf's (1949) principle of least effort, that monosyllabic words in Korean play a special role similar to the role such words play in English.

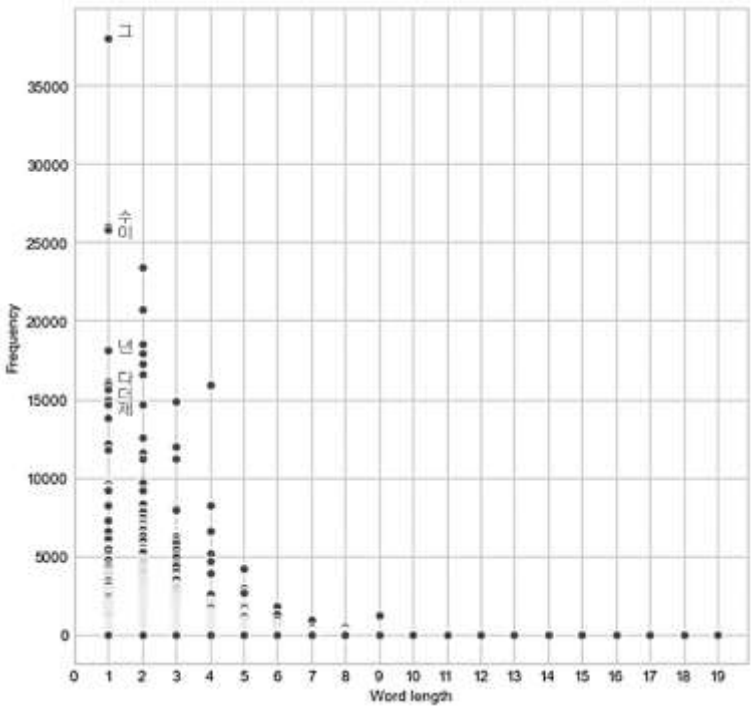
However, Korean text presents us with a special situation. Korean orthography consists of 'syllable-blocks', each standing for a single spoken syllable and containing the relevant phonemic representations. As in Thai, spaces are not obligatory. Legibility is not necessarily compromised by continuous strings of syllable-blocks—eo-jeols—consisting of adjacent words of one, two or more syllable-blocks, with perhaps only occasional punctuation and spaces to resolve particular ambiguities.

The first column in Figure 1 shows the frequency distribution, with examples, of the 712-word monosyllabic eo-jeols. The second column is the frequency profile of all the eo-jeols of two-syllable blocks in the text, the third column the three-syllable eo-jeols, and so on. Note that monosyllabic words can also appear in the multisyllabic eo-jeols. For example, 이삼년 'two to three years' is one eo-jeol but consists of three monosyllabic words: 이 i 'two', 삼 sɐm 'three', and 년 njʌn 'year'. Because our semantic measure is sensitive to the spaces within the written Korean text, such embedded monosyllables were not included in our sample.

Thus, each word-type in our 712-word sample is not represented by *all* of the tokens of that word-type in the corpus. Rather, it is represented by those tokens appearing as an eo-jeol, with a space to the right and left. We argue below that this subset of a word-type's tokens are more likely to be linguistically prominent in ways that we will define.

Fig 1.

The relation between word frequency and eo-jeol 'word length' in our corpus



Note. The minimum frequency = 5; the maximum length of eo-jeol = 19 syllable-blocks. See text for details.

2.3. Semantic distance between words

Words with similar meanings tend to occur in similar contexts (Firth, 1957). The key feature of corpus-based techniques that measure semantic distance is that they quantify the similarity between words by measuring their contextual similarity: if two words are interchangeable in any situation (i.e. absolute synonyms), the similarity between them will be 1, and the distance between them 0.

We operationalised meaning as a context vector for each of our 712 monsyllables. Since Landauer and Dumais' (1997) Latent Semantic Analysis (LSA), new algorithms for estimating semantic distance have been developed, such as *Word2Vec* (Mikolov, Sutskever, Chen, Corrado, & Dean, 2013) and *FastText* (Joulin, Bojanowski, Douze, Jégou, & Mikolov, 2016; Mikolov, Grave, Bojanowski, Puhrsch, & Joulin, 2017), based on neural networks. *Word2Vec* is an open-source program that turns words into vectors, allowing us to quantify semantic distances between words. We trained our corpus through *genism.models.word2vec* with the skip-grams option, and calculated cosine similarity between every word pair to generate semantic distances between words. For example, semantic similarity between 닭 tɐk̚ 'chicken' and 쌀 ssʌl 'rice' is 0.233 whereas that between 쌀 ssʌl 'rice' and 술 sul 'liquor' is 0.985. This means *rice* and *liquor* share more similar semantic contexts than *chicken* and *rice* in our Korean corpus. The procedure was conducted on *Google Colab* due to the large size of the data.

Generating LSA context vectors from Korean internet text raised particular issues, which were also opportunities. The online authors of the corpus text were relatively insensitive to writing norms and frequently omitted spaces between words. In constructing context vectors from English text, a five-word window to the left and right of each token of a word-type simply contains five words separated by spaces; the spaces define lexical words, with only marginal ambiguities (e.g. 'they'll', 'high-school'). If any of those space-defined words belongs to the set of 'context words' chosen to define the semantic vectors, then its presence in the window is recorded in the developing vector for that word-type. The frequency of a context word determines how useful its contribution is; an extremely rare word like 'syzygy' will make virtually no contribution to defining any word-type's context-based semantics. In Korean, if many spaces are omitted the window to the left or right of any token of a monosyllabic eo-jeol

expands to contain five (space-separated) eo-jeols, each of which may contain one or more syllable-blocks. As Figure 1 shows, the maximum length of a Korean 'context word'—or eo-jeol—was 19 syllable-blocks.

What can our Korean 'context eo-jeols' tell us? First, Figure 1 shows that the longer eo-jeols tell us increasingly little, as they quickly become very rare. They are very diverse entities, ending up in the long tail of the distribution. We confirmed this conclusion by carrying out systematicity calculations (see below) with different-sized context vectors in *Word2Vec* from 446 (cf. Monaghan et al., 2010) to 100. There was negligible variability in the level of systematicity.

Second, eo-jeols are very rich; they embody several types of linguistic information. We assume that a corpus of relatively informal internet text is in some ways intermediate between spoken language and formally composed text (cf. Lee, 2001), making it a plausible proxy for a speech corpus and any implications for language evolution/learning/processing, but also making it an interesting genre in its own right.

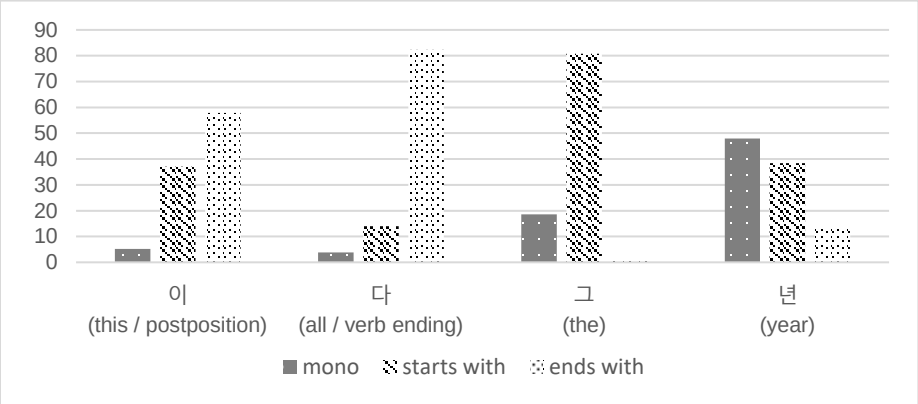
Figure 2 shows the four most frequent words in the corpus and the percentages of times that they stand alone ('mono'), have a space only to the left ('starts with'), and have a space only to the right ('ends with'). The clear differences indicate that the use of spaces in our corpus reflects syntactic function (which always has semantic implications) and will certainly correlate with prosodic performance structures. Omitting spaces signals a degree of legibility and/or predictability.

In summary, the context eo-jeols in our corpus supply a realistic mix of lexical, syntactic, semantic and pragmatic information of varying granularity; indeed, the separation of these

dimensions in computational psycholinguistics is itself artificial. We return, below, to a comparison of systematicity in Korean and in alphabetic languages.

Fig 2.

Percentage of occurrences of spaces and their locations for the four most frequent monosyllabic words



Note. 'Mono' = space before and after the character; 'starts with' = space before the character only; 'ends with' = space after character only. 이 seems to be used more as postposition than as determiner. 그 is mostly used as a determiner.

2.4. Phonological distance between words

We defined Korean consonants by place and manner of articulation and vowels by the position of the tongue-body and lip roundness (Appendix 1). By assigning 1 if a phoneme had the feature and 0 if it did not, each phoneme was represented by a binary vector (cf. Jee, Tamariz, & Shillcock, 2020).

The difference between two vectors was measured by feature-edit distance, counting how many different features there were between the two vectors. To calculate phonological distance between two monosyllables, the distance between the initial consonants, the distance between the

vowels, and the distance between the final consonants were added (cf. Monaghan, et al., 2010). For instance, the distance between /tɐk̚/ and /non/ is the sum of the distances between /t/ and /n/, /ɐ/ and /o/, and /k̚/ and /n/: 5 + 3 + 2 = 10. When we compared a CVC and a CV, we created an empty vector of 14 0s for the final consonant position of the CV. This decision reflects our assumption that there is a continuity of articulatory effort between a vector containing fewer 1s and an empty vector. For all phonological distance measures, textdistance 4.1.4 was used (Python 3.7.1)[3].

Assuming that linguistic systematicity evolved within the spoken language rather than the written language, we treated the syllables as they are pronounced, not as written. For example, /d/, /t/, /s/, /s̬/, /tɕ/ and /tɕʰ/ are all pronounced as /t̚/ in the syllable-final coda position[4].

3. Results

3.1. Meaning-Form Systematicity

We calculated the correlation, Pearson's r, between the corresponding form and meaning lists of distances. For the 712 Korean monosyllables: r = - .13 ($p < .001$). The negative value is due to the fact that we measured correlation between semantic *similarity* with phonological *differences* (Table 1). The result is understood as similar sounding words tending to have similar meanings.

[3] Textdistance 4.1.4 was imported from https://pypi.org/project/textdistance/ on 29 July 2019.

[4] Written language is typically an agreement among the members of a speech community. For example, writing /nɐt/ in different forms (i.e. 낫, 낮, 낯, 낟, and 낱) was an arbitrary decision of Korean linguists in 1933 to facilitate semantic distinction while reading (Sampson, 1985). If the 'Hangeul Simplification Plan' had been passed in parliament in 1953, we would probably write them all as 낫 (Lee, 2015).

Table 1 shows examples of phonological distances and semantic similarities. Although we cannot judge the overall correlation coefficient from a few pairs of distances, the tendency observed in Table 1 resembles the correlation we found. The level of phonological similarity agrees with the level of semantic similarity. For example, 술 ‘liquor’ is relatively similar to 쌀 ‘rice’ both phonologically and semantically, but is relatively dissimilar to 닭 ‘chicken’ on both counts. (N.B. A high score for semantic similarity means semantically similar.)

Table 1.

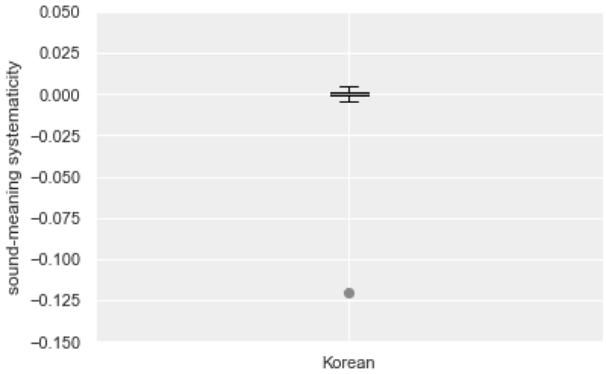
Examples of phonological distances and orthographical distances

First word	Second word	Phonological Distance	Semantic Similarity
닭 tɐk̚ ‘chicken’	술 sul ‘liquor’	12	0.217
닭 tɐk̚ ‘chicken’	쌀 ssʊl ‘rice’	10	0.233
술 sul ‘liquor’	쌀 ssʊl ‘rice’	5	0.985

We conducted a Monte Carlo permutation test to confirm the significance of the correlation, in line with the practice of other systematicity-in-language researchers. We randomly paired the semantic and phonological distances and calculated the correlation, repeating the procedure 10,000 times to collect a distribution of random correlation coefficients so as to calculate the statistical significance of the veridical correlation of -0.13 (Fig. 3). It is located far outside the distribution, confirming the significance of the correlation.

Fig 3

Monte-Carlo permutation test



Note. A Monte-Carlo permutation test confirms the significance of the sound-meaning correlation in Korean. The box represents 25%-75% of the distribution of chance-level coefficients, which are very close to zero. The horizontal lines represent their range. The dot shows the veridical coefficient.

3.2. Lexical frequency

We divided the 712 monosyllabic words into four groups based on frequency of use (Jee, Tamariz, & Shillcock, *in prep*) for the corpus-derived frequency data; these frequency data are from a wider range of vocabulary than previous studies of lexical frequency in Korean (cf. Byun, 2003; Hahm, 1962). The least frequent monosyllables produced significantly weaker (but still significant) systematicity than the other frequency groups (first and third quartiles, $z = 7.14$, $p < .001$; second quartile, $z = 8.94$, $p < .001$) (Table 2).

Table 2

The meaning-sound correlation of each frequency quartile.

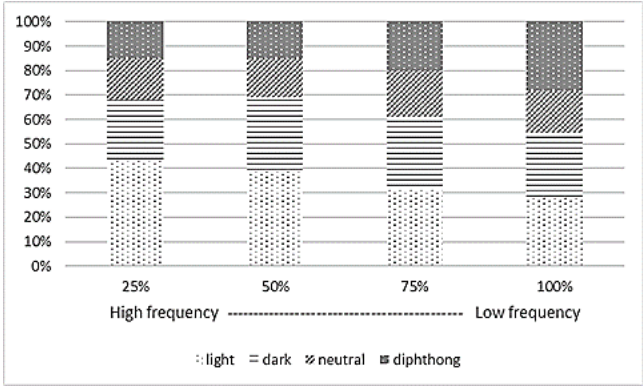
Total corpus = 712 words		N	*r*	*p*
Frequent . . Rare	25%	178	- .11	< .001
	25~50%	178	- .13	< .001
	50~75%	178	- .11	< .001
	75~100%	178	- .03	< .001

Note. N = the number of sample words in each subgroup.

We further explored the different frequency quartiles in terms of light vowels (/ɛ/, /ø/, /a/, /o/), dark vowels (/e/, /y/, /ʌ/, /u/), neutral vowels (/i/, /ɯ/) (Kim-Renaud, 1976) and diphthongs (Fig 4). The least frequent quartile has the least skewed distribution of vowels.

Fig 4

Light vowels, dark vowels, neutral vowels and diphthongs in each frequency quartile



3.3. Syntactic analysis

We extracted subsets of the 712 monosyllables based on syntactic role: 418 nouns, 142 verbs mostly inflected, 73 adjectives and adverbs, and 56 onomatopoeic words. Words that mimic actions (e.g. 콱 kʰʷɐ̄k, 폭 p̄uk̄, 홱 hwɛk̄) were categorized as adverbs, those that mimic sounds (e.g. 꽝 kkʷɐŋ, 삐 ppiː, 슝 ɕjuŋ) as onomatopoeic words. The onomatopoeic words were considered more iconic than mimetic words. The adjectives and adverbs partially overlapped with onomatopoeic words.

Every syntactic group and phonological subgroup demonstrated the same effect as the whole sample of 712 words (Table 3): form-meaning correlations were all significant and final consonants contributed most to the systematicity (although see below for further analysis of syllabic constituents). Onomatopoeic words returned the highest systematicity despite being fewest. The differences between the onomatopoeic words and the other subgroups were significant: nouns ($z = 2.8$, $p = .003$), verbs ($z = 2.63$, $p = .004$), modifiers ($z = 3.19$, $p = .001$).

Table 3

Sound-meaning systematicity in each syntactic subgroup

	Nouns	Verb	Modifiers	Onomatopoeia
Sample size	418	142	73	56
Total correlation	-0.13	-0.13	-0.10	-0.20
Initial consonants	-0.05	-0.04	-0.04	-0.09
Vowels	-0.04	-0.06	-0.05	-0.10
Final consonants	-0.12	-0.12	-0.08	-0.15

Note. All p values < .001. Modifiers include adjectives and adverbs.

In Spanish, the role of consonants in systematicity was opposite to that of vowels (Tamariz, 2008). However, Tables 3 to 5 consistently show that Korean consonants and vowels both positively contribute to systematicity and that the latter part of the syllable contributes more than the onsets (see below for details).

3.4. Etymological analysis

The systematicity of the whole sample of 712 words was found in all four etymological subgroups: 463 pure Korean (with and without homonyms, e.g. 값 kɐ̄p, 개 kjɛ), 141 Sino Korean (with and without homonyms, e.g. 급 kɯ̄p, 죄 t͡ɕøː), 81 homonyms (e.g. 배 pɛ, 차 t͡ɕʰɐ), and 75 loanwords (e.g. 컵 kʰʌ̄p, 햄 hɛm) (Table 4). Some homonyms involve both native and Sino-

Korean meanings. Loanword systematicity was more variable in size and direction of the correlation. The loanwords in our corpus were mostly English (e.g. ‘game’, ‘goal’, ‘shop’, ‘rap’).

Table 4

Form-meaning systematicity of etymological and lexical categories

	Pure Korean		Sino-Korean		Homonyms	Loan words
Sample size	463	432	141	119	81	75
Total correlation	-0.13	-0.14	-0.13	-0.11	-0.13	-0.01 (p = .01)
Initial consonants	-0.04	-0.04	-0.05	-0.03	-0.04	0.02
Vowels	-0.05	-0.06	-0.06	-0.06	-0.03	-0.05
Final consonants	-0.12	-0.12	-0.11	-0.08	-0.13	0.01 (p = .01)

Note. All *p*-values < .001 except where stated. The two etymological categories are with and without homonyms, respectively.

3.5. Phonological segmentation

We analysed how each syllabic constituent contributed to systematicity; Tables 3 and 4 show how consonants and vowels contribute to systematicity. Consonants and vowels both contributed positively to the whole systematicity (Table 5). Final consonants contributed most, but when vowel and coda were combined to form the rhyme the correlation increased. All the differences between correlations were significant: between onset and rhyme (z = 28.66, $p < .001$), between onset and vowel (z = 3.56, $p < .001$) and between final consonant and rhyme (z = 3.61, $p < .001$).

Table 5

Meaning-sound correlation of each syllabic constituent

	r		*r*
Initial consonant	- .04	Onset	- .04

Vowel	- .05	Rhyme (vowel + coda)	- .12
Final consonant	- .11		

Note. All *p*-values < .001.

We further investigated whether particular segments contribute to syntactic and etymological identity (Figures 5 and 6, respectively). Paralinguistic words are sounds like ‘ah’, ‘uh’ and ‘oh’, and tend to begin with /tʰ/, /pʰ/, /kʰ/ and /h/ and end with /k̚/. A few qualitative researches (Martin, 1962; Sohn, 2005) also observed that many Korean onomatopoeic words ends with /k̚/. The syllables ending with /n/ and /l/ are likely to be inflected verbs (/n/ for past, e.g. 한 hɐn; /l/ for future, e.g. 할 hɐl).

Etymologically, the onset /g/ strongly suggests that the word is Sino-Korean (Fig 6a). The tensed onsets (/t͈/, /p͈/, /t͈ɕ/, /k͈/ and /s͈/) are clear indications of being native Korean, which agrees with Sohn (2005)’s analysis of Korean phoneme distribution. Meanwhile, the aspirated onsets are likely to be loanwords, and the Sino-Korean monosyllables hardly ever end with /t̚/.

Fig 5a.

Connection between the initial consonants and syntactic category.

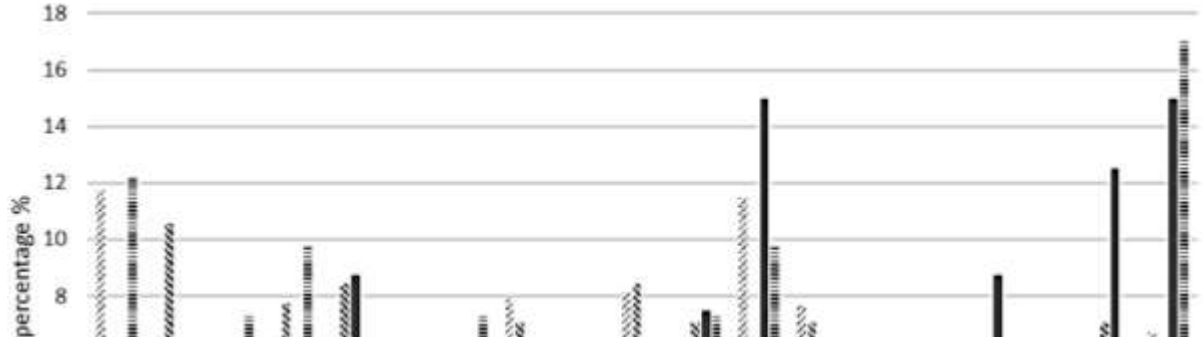
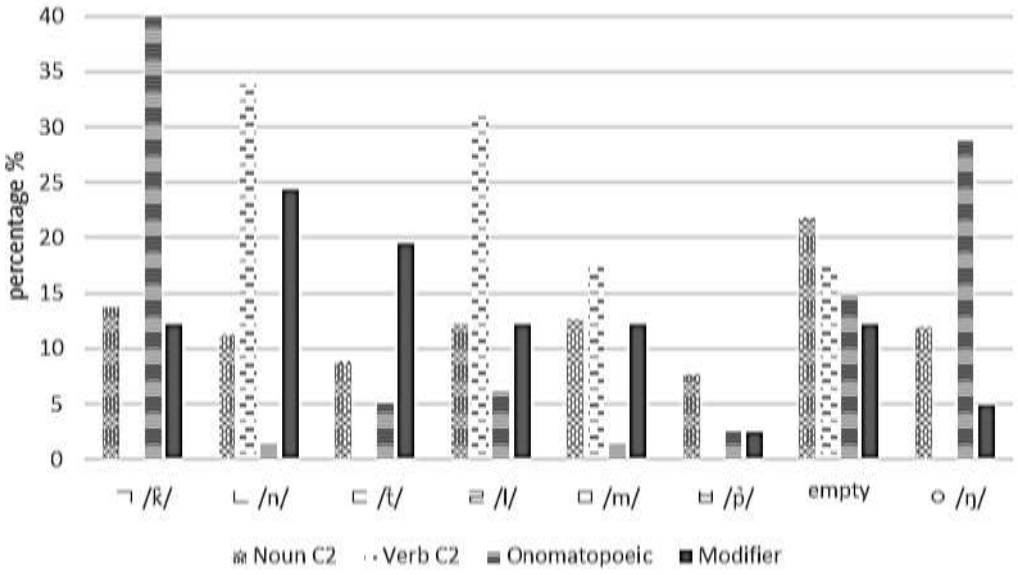


Fig 5b

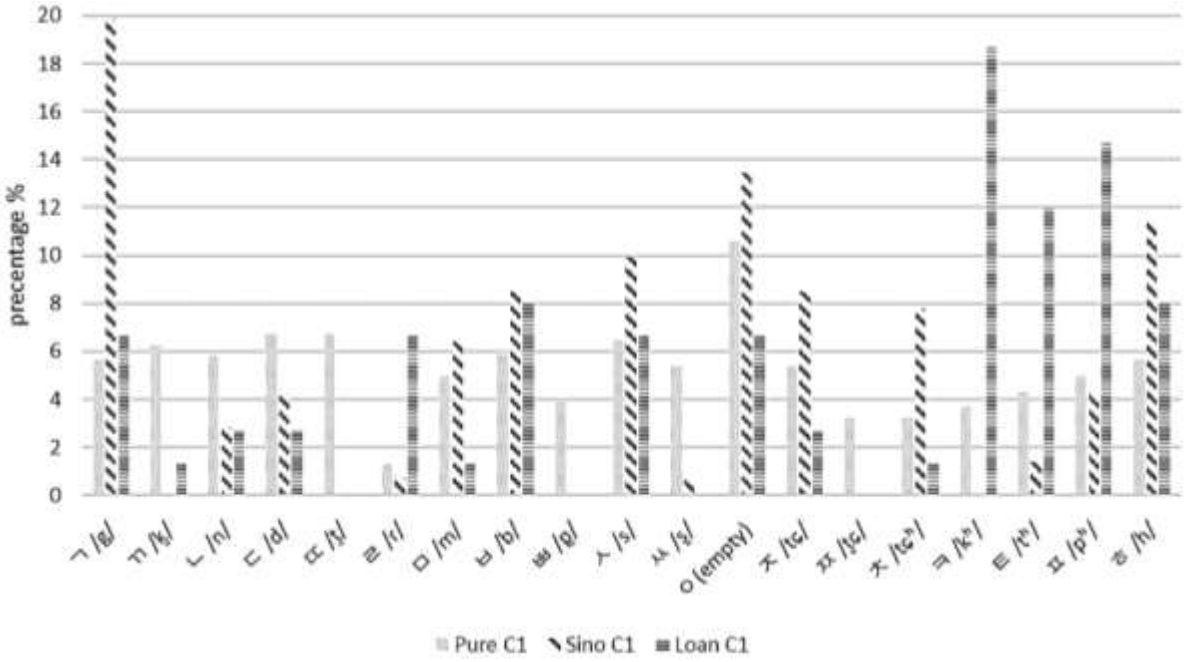
Connection between the final consonants and syntactic category.



Note. Korean phonology allows 7 consonants in the final position. The onomatopoeic words tend to end with /k̚/.

Fig 6a

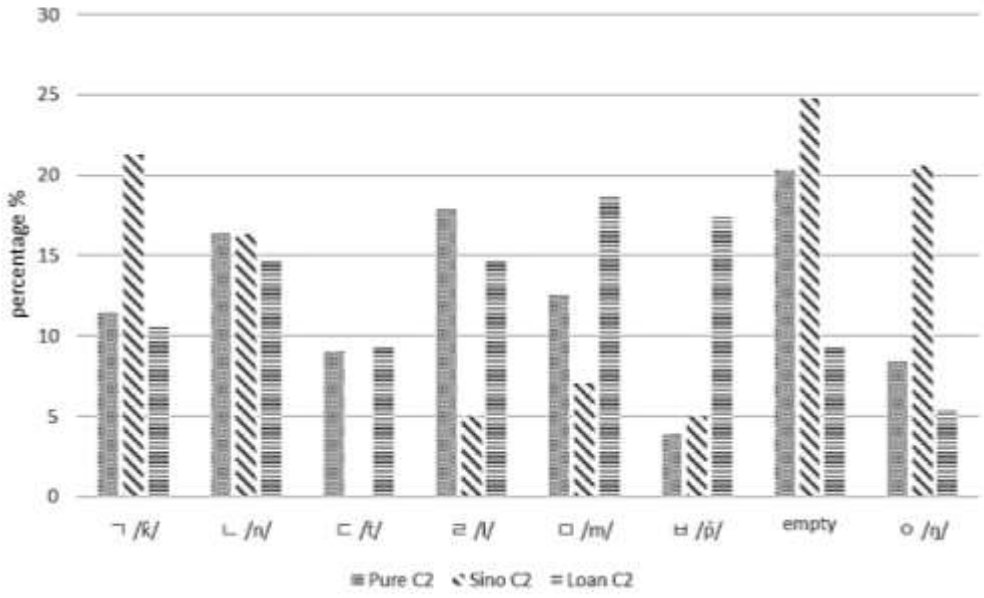
Connection between the initial consonants and etymological category.



Note. Loanwords tend to begin with aspirated phonemes.

Fig 6b

Connection between the final consonants and etymological category.



Note. Sino-Korean monosyllables do not end with /t̚/.

4. Discussion

We have successfully extended the exploration of word-level semantic-phonological systematicity to Korean. Previous studies have shown partial, segment-specific relations between phonology and meaning in many languages, and we have demonstrated similar such relations in Figures 5 and 6 for syntactic and etymological categories in our Korean data. However, the special significance of the studies by Monaghan et al. (2014) for English, by Tamariz (2008) for Spanish, and by Dautriche et al. (2017) for English, Dutch, German and French, is that they demonstrate a systematicity that is *distributed* across a substantial part of the lexicon at the whole-word level and not confined to particular word-internal symbolic relationships (as in phonaesthemes) between phonology and semantics (cf. Blasi et al., 2016).

We have found the same distributed systematicity in Korean, a language that is typologically very different from the previously studied European languages. Our study reinforces Dautriche et al.'s (2017) claim that this systematicity is universal[5]. Note that Dautriche et al.'s claim was made on the basis of using orthography as a proxy for phonology, which may have introduced etymological information preserved in the orthography and thereby inflated the observed systematicity in some languages: e.g. English orthography distinguishes the homophones "sight", "site" and "cite".

Our key *r* values—reaching as high as .13—were substantially larger than those in previous studies, which have typically been on the order of .05. Our measured systematicity accounts from 1.7% of the variance ($R^2 = .07$) and therefore begins to look like a regularity that a learner's brain might recognize. But why should there be this difference compared with previous studies, given that the methodologies were closely comparable in many respects?

[5] Dautriche et al. did not report data for Korean.

We suggest that our use of eo-jols as units of orthography in specifying the LSA context vectors explains the high levels of systematicity discovered in Korean. Our findings for Korean are text-based, in this sense, and are not directly comparable with reported levels of systematicity in other languages (e.g. Spanish) in which there is a different relationship between the text corpora and the generation of LSA contexts, leading to shorter, less rich LSA contexts in those other languages. We have explained above the necessity of adopting eo-jeols as units. Eo-jeols represent a rich combination of semantic, syntactic and pragmatic information specifying an extended context for a word. They include syntactic information by indicating whether a word is written independently or with marker(s). They also include pragmatic information by providing a writer with choices regarding adding spaces[6]. Note that Bullinaria and Levy (2007) show that larger LSA window sizes are more appropriate for semantic tasks.

We further suggest that this information will almost certainly be closely associated with prosodic information in speech. Thus, Korean has forced us to adopt what may be a very good way of characterizing the meaning of a word in terms of its context. The space-based, purely lexical segmentation of context used in previous explorations of phonosemantic systematicity thus perhaps underestimates the phenomenon by failing to allow a fuller role for other linguistic categories (e.g. syntax) which we know have a systematic relationship with phonology. We

[6] Eo-jeols combine semantic and syntactic information with various markers: 달 tɐl ‘moon’ acts as subject with the subject marker (e.g. 달이 tɐɾi) and as object with the object marker (e.g. 달을 tɐɾɯl). There are other markers: for indirect objective (e.g. 달에게 tɐɾɛgɛ or 달한테 tɐlhɐntɛ), for emphasizing (e.g. 달은 tɐɾɯn), and for additional semantic information (e.g. 달만 tɐlmɐn ‘the moon only’; 달도 tɐldo ‘the moon as well’; 까지 tɐlk̚ɐdʑi ‘until the moon’, etc.). They can also reveal pragmatic information by representing whether the writer prefers to put a space between them. Although influenced by education and training, the use of spaces indicates the writer’s units of semantic concepts.

suggest that the nearest analogue to the eo-jeol that might be applied to an English text corpus when quantifying phonosemantic systematicity is the ‘phonological phrase’ (Gee & Grosjean, 1983), which yields relatively flat, relatively symmetric tree structures that capture phrasing and pausing as well as other categories of linguistic structure. Phonological phrases yield a range of potential entities for an LSA analysis, reflecting the number of nodes in the tree structure dominating a space between words.

Analysis of a text corpus in terms of eo-jeols in Korean, or phonological phrases in English, produces a large, long-tailed distribution of entities to replace the context words in an LSA calculation. However, we have shown above that the semantic definitions produced are overwhelmingly dependent on the high frequency eo-jeols. Thus, the analysis is not only tractable but there is now no need for the arbitrary selection of LSA context words; instead, the whole corpus provides the context words.

Overall, however, the investigation of semantic-phonological systematicity is still in its early stages. We await a clearer picture of the implications of the necessary differences between the subsets of the lexicon chosen in studies of systematicity in different languages. For instance, the 712-word sample in the current study is smaller than the 2,572 monomorphemic word sample studied by Monaghan et al. (2014). Such differences between studies can reflect inherent differences between the languages involved.

There may, of course, exist language-specific differences in the size of semantic-phonological systematicity, beyond any differences between the details of the studies; languages are not necessarily equivalent in all aspects of their learnability and ease of use (cf. Tylén et al., 2019). One clear but partial influence on systematicity in Korean is the correlation between

phonology and etymological and syntactic categories (Figs. 5 and 6), to which we return below. The size of such correlations may also be language specific.

A further possibility in the present case is that Korean has been less multifariously influenced by other languages, compared with English, Spanish, German, Dutch or French, which have all had extensive interactions with other languages, involving large numbers of L2 speakers from diverse L1 backgrounds over prolonged periods of time. It may be that semantic-phonological systematicity best develops in relative isolation (cf. Lupyan & Dale, 2010).

We can expect to see statistically larger estimates of systematicity as richer and psychologically more realistic measures of phonological and semantic distance are developed, and systematicity emerges as a more credible adaptation to infant language acquisition compared with the very small—but statistically significant—amount of variance captured by previously reported meaning-form correlations. Note that 'psychologically realistic' may even mean abandoning a symbolic paradigm in favour of physical measurements of units (see Torre, Luque, Lacasa, Kello, & Hernández-Fernández, 2019). Again, such improved metrics may well be language specific. In the current study, we have demonstrated the potential of the eo-jeol in these respects.

In summary, our study of Korean replicates and extends the existing research on semantic-phonological systematicity. Various factors may contribute to the observed systematicity being higher than reported for other languages, but our use of eo-jeols in defining contexts is the best candidate explanation.

However, as well as raising the issue of eo-jeols and similar ways of parameterising context, the study of Korean sheds light on the nature of phonosemantic systematicity itself. How does it come to exist in the first place? Korean allowed us to divide our sample of the lexicon in

ways not always reported in other studies: by frequency, syntactic category, etymological origin, homonyms, syllabic constituents, onomatopoeia and loanwords. The resulting analyses suggest a fundamental explanation of distributed systematicity in language behaviours.

Systematicity was least in the less frequent words (Table 2), reflecting Monaghan et al. (2014), who also point out that the higher systematicity in the culturally central frequent words may orient infants to the meaning-form relationship in the early stages of language acquisition. The picture that emerges is one in which rare words participate in less and less systematicity before dropping out of usage. *Symbolic* systematicity may help rare words stay in usage, as with English phonoaesthemes (*glint*, *gleam*, *glow … gloaming*), but it is an open question as to any similar effect being caused by the *distributed* systematicity that we report here for Korean.

When the sample was divided by syntactic category—nouns, verbs, modifiers, onomatopoeia—there was significant systematicity in each category, but it was highest in the onomatopoeia. What is special about onomatopoeia?

They possess particular phonological similarities—more than half have /k̚/ in their final position. But they are also less syntactically constrained than the other three categories, meaning that their lexical contexts tend more to resemble the general ('vanilla') profile of all the possible contexts in the corpus. Thus, onomatopoeia tend to sound the same and tend to 'mean' the same, according to their contexts—therefore they inherit high systematicity. Researchers have previously acknowledged the high proportion of onomatopoeic words in Korean and speculated on their benefits in terms of language acquisition in infants (Martin, 1962; Sohn, 2005).

Certain consonants tend to predict syntactic role or etymological roots, either in initial or final position in the syllable. Our 712 monosyllables included 142 inflected verbs, which is one clear source of systematicity, although their systematicity was still closely similar to that of

nouns (Table 3). It should be noted that there was a disparity in inventory size between onset and coda.

Tables 3, 4 and 5 show the importance of the final consonant and the rhyme; indeed, the greater role for the rhyme indicates the psychological reality of the rhyme as a category. Figures 5a and 5b reveal the wider, flatter profile of initial consonants compared with final consonants (as we also see in English), making the former more informative and facilitating incremental processing, but seemingly supporting systematicity less than the coda in our sample.

We also studied Korean's single greatest contact language—Chinese. We found no significant difference in systematicity between native Korean and Sino-Korean monosyllables (Table 4). Many Sino-Korean words are mistakenly considered as native Korean words due to their long history (e.g. 죽 tɕuk̚ 'porridge' or 수염 su.jʌm 'beard'). According to Sohn (2005), Chinese characters had been used since 194 B.C. and, as a result, most of the Sino-Korean words behaved as native words by the period of the Goryeo dynasty (935-1392). Native Korean and Sino-Korean both contribute to the overall systematicity in our sample. Considering the long history of co-existence of the two vocabularies, this result might indicate a close accommodation between them. Or it might indicate the same underlying explanation of systematicity. This analysis is the first exploration of semantic-phonological systematicity in Chinese vocabulary, as far as we are aware (c.f. Fulang & Kenstowicz, 2021; Starr, Yu, & Shih, 2018; Wong & Kang, 2019).

Korean's foreign loanwords constitute an extreme case of contact with another language; they are mostly English words and with a much shorter history of use in Korean, compared with Chinese words. These loanwords returned smaller, less consistent systematicity, as we might expect, but it was still significant even with the relatively small numbers of loanwords.

Loanword systematicity itself suggests a very basic mechanism—something outside the slower cultural evolution of the Korean lexicon.

Korean allows us to consider the role of homonyms in systematicity. Why are they such a pervasive feature of natural languages—who would design a language to use the *same* word to mean *different* things? Homonyms are the extreme end of a continuum of meaning differentiation; less extreme are differences of *sense*, as in 'newspaper' meaning physical pieces of paper compared with an institution such as '*The Times* of London'. This continuum can only be approximated by lexicographical criteria.

Table 4 shows a high level of systematicity within the homonym subset. There is a methodological perspective to this result. It is prohibitively time-consuming to check every occurrence of a homonym in a very large text corpus and to annotate it for its intended meaning. In reality, intended meanings of homonyms are heavily skewed towards the single most frequent meaning (Kang, 2005), but it remains the case that the semantic vector for a homonym in our study was necessarily an average of all of its different meanings (mediated by frequency in the corpus), thus tending towards the vanilla vector. The frequency of a homonym tends to be higher, *ceteris paribus*, given that it is the sum of the frequencies of all of its dictionary entries; it therefore tends to have the simpler phonology of more frequent words. Thus, a simpler phonology correlates with a fuller vector, indicating systematicity.

When we divided the 712-word sample in terms of the segments occupying different syllabic positions (Tables 3,4,5) we found pervasive, significant systematicity, even at the level of partitioning both by syntactic category and by syllabic position. Fig. 4 shows that the level of systematicity in the different frequency quartiles tracks the degree of skewedness of the distribution of vowels, with the light/dark/neutral vowels and the diphthongs being in more equal

proportions in the lowest frequency quartile, which has the lowest value for systematicity. Light vowels become more numerous with higher frequency, which has developmental implications: light vowels are known to connote something small, light, and bright and frequently appear in the texts for children (Cho, 2006; Kwon, 2018; Sohn, 2005).

We find the same direction of systematicity for all syllable positions: words that were similar at any position tended to be similar in meaning. This result contrasts with Tamariz's (2008) study of Spanish, in which words sharing vowels tended to have *different* meanings and words sharing consonants tended to have similar meanings. This difference may reflect wider differences concerning vowels in the two languages (e.g. vowel harmony in Korean); it is a subject for further investigation. Note that a negative correlation is still informative for the listener.

In summary, beyond the simple demonstration of systematicity in Korean, we have explored a more richly based systematicity, operating over a range of linguistic dimensions as reflected in eo-jeols, resulting in a level of systematicity capturing some 1.7% of the variance, compared with the less than 0.5% of previous studies. Moreover, the systematicity was *pervasive*; whichever way we partitioned our representative sample of the lexicon, we found significant systematicity, suggesting a general underlying explanation not exclusively concerned with whole-word phonology.

We now consider the nature of such an explanation, but we first make a point about *proxies* in causation. McDonald and Shillcock (2001a, b) observe that psycholinguists have conventionally seen word frequency as mediating behavioural responses, analogous to Hebbian learning (Hebb, 1949). However, the more frequently a word occurs, the more opportunity it has to acquire different contexts; in terms of a Latent Semantic Analysis-type context vector, the

frequent word ‘time’ can accrue more context-word hits within the window around each of its tokens in the corpus. Thus, word frequency may be seen as a *proxy* for *Contextual Distinctiveness* (CD), which represents how much information a word conveys about its contexts of use (see, also, Baayen, 2010). To illustrate: a contextually very constrained word like ‘amok’ has a very high CD score and attracts long behavioural response latencies. Words like ‘amok’ and ‘wreak’ may not initially even look like real words in isolation from their typical contexts ‘run —’ and ‘— havoc’. At the other extreme, the words with very low CD scores are mostly function words. The ‘word’ (in a spoken language corpus) with the lowest CD score is the filled pause ‘er’—it can appear in virtually any context in transcribed speech (this is similarly true of expletives). Therefore, word frequency is a *proxy* for a semantic distinction, and the context vector for a very high frequency word approximates the average, vanilla context vector available across the whole corpus.

When we try to understand cognitive processing, any activity can be a proxy for any other correlated activity. What activity has the most explanatory value? One answer is that it should be something that most effectively interacts in a material way with all of the other activities in the domain (spoken language, in this case), such that in its own way it is characteristic of what is happening in the whole domain (cf. Shillcock, 2014). We argue here that *human effort* is just such a ‘universal’ in the domain of communication (see, e.g., i Cancho and Solé, 2003). Zipf’s (1949) Law—that ordering the words of a corpus by decreasing frequency closely approximates a power function—was originally seen by Zipf in terms of least effort.

In a phonological version of the lexicon, there is a general effort-related gradient: if we proceed from lower frequency words (e.g. “preen”, in English) to higher frequency words (the filled pause, “er”) the words we encounter tend to become shorter (Zipf, 1935); their

phonological distinctiveness becomes less (Meylan & Griffiths, 2017); they contain segments that are easier to pronounce (see, e.g., Shi et al., 1998); they are more subject to phonological reduction when realized in fast speech (Gahl, Yao, & Johnson, 2012); their phonological neighbourhoods are denser, at least partly a result of words becoming shorter. One aspect of effort by the *speaker* means articulating longer, phonologically more complex words.

One aspect of effort by the *listener* can mean struggling to activate relatively atypical phonological forms that do not have the interactive support of phonological neighbours and/or struggling to suppress the more frequently occurring patterns of activation.

In the semantics implied by lexical entries, we again see a general effort-related gradient. Proceeding again from lower to higher frequency words, we see increasing homonymy and polysemy (Morton, 1979; Jastrzembski, 1981), and the emergence of the qualitatively new category of function words. In context vector terms, very sparsely populated vectors progress to vectors that approximate more closely the vanilla vector that reflects the statistics of usage of the language. *Speaker* effort may mean struggling to make a rare, sparsely specified semantics activate the relevant phonology and suppressing potentially interfering relations between meaning and phonological form. Similarly, for the *listener*, effort may mean activating a sparsely specified meaning and struggling to suppress potential interference.

In this view, we are dealing with a complex system (cf. Van Orden & Stephen, 2012), in which everything interacts more or less with everything else. The default state is the *poise* within which the relation between any two elements is the average one in the past experience of the individual. It takes effort to move from that state to a state in which a particular constellation of phonological activity or a particular constellation of semantic/pragmatic activity is foregrounded. Somewhat counterintuitively, *less* effort is required to move from the default activity state to

deal with a word like ‘get’ (a frequent homonym, participating in common idioms, and also in transition to the functor lexicon) compared with a word like ‘preen’ (an infrequent word with a more elaborate pronunciation and a rarefied, sparsely specified usage).

Wray and Grace (2007) discuss the distinction between *esoteric* and *exoteric* languages, in the context of language evolution. Esoteric languages, such as Korean, are spoken among people with strong common cultural ground, are complex and heavily contextualized and they are more likely to be prototypical of human language. Exoteric languages, such as English, have undergone contact with strangers and with other languages and have involved L2 learning; they are characterized by greater structural simplicity and by a larger content word lexicon.

We propose that this basic esoteric/exoteric distinction exists within all current language use as a graded distinction operating even within an utterance. It is best understood in complex systems terms. Esoteric communication corresponds to default engagement of a wide range of heavily contextualized, mutually supportive interactivity within the system, operating over speech and gesture, and involving relatively little effort in creating small but informative departures from this default poise. Exoteric communication corresponds to the relatively effortful, larger departures from this wide range of activity to generate relatively sparse, ‘abstract’ constellations of activity that are also informative. Deploying the word ‘get’ (easy to pronounce and with easily accessed rich, multiple uses and contexts) as opposed to ‘trove’ (harder to pronounce and with harder to access specific uses and contexts) illustrates the difference. We suggest that the correspondence between the degree of effort involved is the basis of systematicity between form and meaning.

In conclusion, phonosemantic systematicity exists in Korean, as in other languages, but the specific qualities of Korean have allowed us to show a higher level of systematicity

compared with previous studies, suggesting directions for future research and also illuminating the nature of the systematicity. The best way to understand this systematicity is in complex system terms—a correlated reduction of effort in speakers and listeners in generating phonology and meaning along a dimension that corresponds to an esoteric/exoteric distinction.

Acknowledgments

We would like to thank six anonymous JEAL reviewers and Dr James Huang. Their comments have greatly improved the paper; any remaining errors are our own.

Appendix

Table A. 14 articulatory features for Korean consonant phonemes.

	ㅂ	ㄷ	ㄱ	ㅅ	ㅈ	ㅁ	ㄴ	ㅇ		ㄹ	ㅋ	ㅌ	ㅍ	ㅊ	ㅎ	ㄸ	ㄲ	ㅃ	ㅆ	ㅉ
	p	t	k	s	tɕ	m	n	ŋ	-	l	k^h	t^h	p^h	$tɕ^h$	h	$t\vec{t}$	$k\vec{k}$	$p\vec{p}$	s̬	t̬ɕ
palatal	0	0	0	0	1	0	0	1	0	0	0	0	0	1	0	0	0	0	0	1
velar	0	0	1	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0
labial	1	0	0	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	0	0
alveolar	0	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	0	0	1	0
dental	0	1	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0	0	0	0
glottal	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0
plosive	1	1	1	0	0	0	0	0	0	0	1	1	1	0	0	1	1	1	0	0
affricate	0	0	0	0	1	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1
fricative	0	0	0	1	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	0
fortis	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1
lenis	1	1	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
aspirated	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0	0	0	0	0	0
nasal	0	0	0	0	0	1	1	1	0	0	0	0	0	0	0	0	0	0	0	0
lateral	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

Note. The phonemes with diacritic ($\vec{k}$, $\vec{t}$, $\vec{p}$) indicates unreleased consonants in the final position. We treated these as same as the tensed ones (k̬, t̬, p̬) for our analysis.

Table B. 11 articulatory features for Korean vowel phonemes.

	ㅏ	ㅓ	ㅐ	ㅔ	ㅗ	ㅜ	ㅡ	ㅣ	ㅚ	ㅟ	ㅑ	ㅕ	ㅒ	ㅖ	ㅛ	ㅠ	ㅘ	ㅞ	ㅙ	ㅝ	ㅢ
	a	ʌ	ɛ	e	o	u	ɯ	i	ø	wi	ja	jʌ	jɛ	je	jo	ju	wa	we	wɛ	wʌ	ɯi
front	0	0	1	1	0	0	0	1	1	1	0	0	1	1	0	0	0	1	1	0	1
central	1	0	0	0	0	0	0	0	0	0	1	0	0	0	0	0	1	0	0	0	0
back	0	1	0	0	1	1	1	0	0	0	0	1	0	0	1	1	0	0	0	1	0
high	0	0	0	0	0	1	1	1	0	1	0	0	0	0	0	1	0	0	0	0	1
middle	0	1	0	1	1	0	0	0	1	0	0	1	0	1	1	0	0	1	0	1	0
low	1	0	1	0	0	0	0	0	0	0	1	0	1	0	0	0	1	0	1	0	0
roundness	0	0	0	0	1	1	0	0	1	1	0	0	0	0	1	1	0	0	0	0	0
diphthong	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1	1	1	1
/w/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1	1	1	1	0
/j/	0	0	0	0	0	0	0	0	0	0	1	1	1	1	1	1	0	0	0	0	0
/ɯ/	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	1

Note. ‘Diphthong’ represents whether the phoneme is or is not a diphthong. The bottom three indicates the shorter vowels of the diphthongs.

References

Baayen, R. Harald. "Demythologizing the word frequency effect: A discriminative learning perspective." *The Mental Lexicon* 5, no. 3 (2010): 436-461.

Baayen, R. Harald, Richard Piepenbrock, and Leon Gulikers. "The CELEX lexical database (cd-rom)." (1996).

Bard, Ellen Gurman, Richard C. Shillcock, and Gerry TM Altmann. "The recognition of words after their acoustic offsets in spontaneous speech: Effects of subsequent context." *Perception & Psychophysics* 44, no. 5 (1988): 395-408.

Blasi, Damián E., Søren Wichmann, Harald Hammarström, Peter F. Stadler, and Morten H. Christiansen. "Sound–meaning association biases evidenced across thousands of languages." *Proceedings of the National Academy of Sciences* 113, no. 39 (2016): 10818-10823.

BNC Consortium. "British national corpus." *Oxford Text Archive Core Collection* (2007).

Brown, Roger W., Abraham H. Black, and Arnold E. Horowitz. "Phonetic symbolism in natural languages." *The Journal of Abnormal and Social Psychology* 50, no. 3 (1955): 388.

Bullinaria, John A., and Joseph P. Levy. "Extracting semantic representations from word co-occurrence statistics: A computational study." *Behavior research methods* 39, no. 3 (2007): 510-526.

Byun, Sung Wan. "Frequencies of Korean syllables and the distribution of syllables of PB word list." *Korean Journal of Otolaryngology-Head and Neck Surgery* 46, no. 9 (2003): 737-741.

Chen, Fulang & Michael Kenstowicz. "Phonotactics of gender in Mandarin given names: patterns and constraints. " *The Annual Meeting on Phonology,* University of Toronto, 2021.

Cho, Young-mee Yu. "Sound symbolism in Korean." *Korean language in culture and society* (2006): 64-73.

Clark, Andy. "Whatever next? Predictive brains, situated agents, and the future of cognitive science." *Behavioral and brain sciences* 36, no. 3 (2013): 181-204.

Dautriche, Isabelle, Kyle Mahowald, Edward Gibson, and Steven T. Piantadosi. "Wordform similarity increases with semantic similarity: An analysis of 100 languages." *Cognitive science* 41, no. 8 (2017): 2149-2169.

Diffloth, Gérard. "i: big, a: small. Sound symbolism, ed. by Leanne Hinton, Johanna Nichols, and John J. Ohala, 107-14." (1994).

Ellison, T. Mark. "Categorisation as topographic mapping between uncorrelated spaces." In *Algorithmic Probability and Friends. Bayesian Prediction and Artificial Intelligence*, pp. 131-141. Springer, Berlin, Heidelberg, 2013.

Firth, John R. "A synopsis of linguistic theory, 1930-1955." *Studies in linguistic analysis* (1957).

Fitneva, Stanka A., Morten H. Christiansen, and Padraic Monaghan. "From sound to syntax: Phonological constraints on children's lexical categorization of new words." *Journal of child language* 36, no. 5 (2009): 967-997.

Freeman, Diane Larsen, and Lynne Cameron. "Research methodology on language development from a complex systems perspective." *The modern language journal* 92, no. 2 (2008): 200-213.

Gahl, Susanne, Yao Yao, and Keith Johnson. "Why reduce? Phonological neighborhood density and phonetic reduction in spontaneous speech." *Journal of memory and language* 66, no. 4 (2012): 789-806.

Gee, James Paul, and François Grosjean. "Performance structures: A psycholinguistic and linguistic appraisal." *Cognitive psychology* 15, no. 4 (1983): 411-458.

Georg, Stefan, Peter A. Michalove, Alexis Manaster Ramer, and Paul J. Sidwell. "Telling general linguists about Altaic." *Journal of Linguistics* 35, no. 1 (1999): 65-98.

Grosjean, François. "How long is the sentence? Prediction and prosody in the on-line processing of language." (1983): 501-530.

Hahm, Tae Young. "한국어음청력검사 어표와 명료도 검사 성적에 관한 연구." *대한이비인후학회지* 25, no. 2 (1962): 721-741.

Hebb, Donald Olding. *The organisation of behaviour: a neuropsychological theory*. New York: Science Editions, 1949.

i Cancho, Ramon Ferrer, and Ricard V. Solé. "Least effort and the origins of scaling in human language." *Proceedings of the National Academy of Sciences* 100, no. 3 (2003): 788-791.

Jastrzembski, James E. "Multiple meanings, number of related meanings, frequency of occurrence, and the lexicon." *Cognitive psychology* 13, no. 2 (1981): 278-305.

Jee, Hana, Monica Tamariz, & Richard Shillcock. “Quantifying sound-graphic systematicity; Application to multiple phonographic orthographies” In *In G21C 2020: Grapholinguistics in the 21st Century* (Vol. 4), June 17-19, 2020c, Fluxus Editions.

Johannessen, Janne Bondi, and Emiliano Raul Guevara. "What kind of corpus is a web corpus?." In *Proceedings of the 18th Nordic Conference of Computational Linguistics (NODALIDA 2011)*, pp. 122-129. 2011.

Joulin, Armand, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hérve Jégou, and Tomas Mikolov. "Fasttext. zip: Compressing text classification models." *arXiv preprint arXiv:1612.03651* (2016).

KANG, Beom-Mo. "Aspects of the use of homonyms. Language research 41 (1), 1-29." *Language research* 41, no. 1 (2005): 1-29.

Kelly, Michael H. "Using sound to solve syntactic problems: The role of phonology in grammatical category assignments." *Psychological review* 99, no. 2 (1992): 349.

Kelly, Michael H. "The Role of Phonology in Grammatical Category Assignments. Signal to Syntax: Bootstrapping from Speech to Grammar in Early Acquisitioned. by James L. Morgan & Katherine Demuth, 249–262." (1996).

Kim-Renaud, Y-K. "Semantic features in phonology: evidence from vowel harmony in Korean." In *Papers from the... Regional Meeting. Chicago Ling. Soc. Chicago, Ill*, no. 12, pp. 397-412. 1976.

Klank, Linda JK, Yau-Huang Huang, and Ronald C. Johnson. "Determinants of success in matching word pairs in tests of phonetic symbolism." *Journal of Verbal Learning and Verbal Behavior* 10, no. 2 (1971): 140-148.

Kunihira, Shirou. "Effects of the expressive voice on phonetic symbolism." *Journal of Verbal Learning and Verbal Behavior* 10, no. 4 (1971): 427-429.

Kwon, Nahyun. "Iconicity correlated with vowel harmony in Korean ideophones." *Laboratory Phonology: Journal of the Association for Laboratory Phonology* 9, no. 1 (2018).

Landauer, Thomas K., and Susan T. Dumais. "A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge." *Psychological review* 104, no. 2 (1997): 211.

Larsen, Darrell, and Jeffrey Heinz. "Neutral vowels in sound-symbolic vowel harmony in Korean." *Phonology* 29, no. 3 (2012): 433-464.

Lee, David YW. "Defining core vocabulary and tracking its distribution across spoken and written genres: Evidence of a gradience of variation from the British National Corpus." *Journal of English Linguistics* 29, no. 3 (2001): 250-278.

Lee, Sang Hyeok. "Phoneticism and Hangeul Simplification Plan viewed through Korean Linguistics History -With a Focus on the Conflict Related to the Draft for the Unification of Hangeul Orthography" *한성어문학 34*, (2015): 35-58.

Lupyan, Gary, and Rick Dale. "Language structure is partly determined by social structure." *PloS one* 5, no. 1 (2010): e8559.

Mantel, Nathan. "The detection of disease clustering and a generalized regression approach." *Cancer research* 27, no. 2 Part 1 (1967): 209-220.

Marslen-Wilson, William D. "Functional parallelism in spoken word-recognition." *Cognition* 25, no. 1-2 (1987): 71-102.

Martin, Samuel E. "Phonetic symbolism in Korean." *American studies in Altaic linguistics* 13 (1962): 177-189.

McDonald, Scott A., and Richard C. Shillcock. "Rethinking the word frequency effect: The neglected role of distributional information in lexical processing." *Language and Speech* 44, no. 3 (2001): 295-322.

McDonald, Scott, and Richard Shillcock. "Contextual Distinctiveness: a new lexical property computed from large corpora." *Behavior Research Methods, Instruments and Computers* (2001).

Meylan, Stephan C., and Thomas L. Griffiths. "Word forms-not just their lengths-are optimized for efficient communication." *arXiv preprint arXiv:1703.01694* (2017).

Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. "Distributed representations of words and phrases and their compositionality." In *Advances in neural information processing systems*, pp. 3111-3119. 2013.

Mikolov, Tomas, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. "Advances in pre-training distributed word representations." *arXiv preprint arXiv:1712.09405* (2017).

Monaghan, Padraic, and Morten H. Christiansen. "Integration of multiple probabilistic cues in syntax acquisition." *Trends in corpus research: Finding structure in data* (2008): 139-63.

Monaghan, Padraic, Morten H. Christiansen, Thomas A. Farmer, and Stanka A. Fitneva. "Measures of phonological typicality: Robust coherence and psychological validity." *The Mental Lexicon* 5, no. 3 (2010): 281-299.

Monaghan, Padraic, Richard C. Shillcock, Morten H. Christiansen, and Simon Kirby. "How arbitrary is language?." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369, no. 1651 (2014): 20130299.

Morgan, James L., and Katherine Demuth. "Signal to syntax: An overview." *Signal to syntax* (2014): 13-34.

Morton, John. "Word recognition." *Psycholinguistics: Series 2. Structures and processes* (1979): 107-156.

Ramstedt, Gustaf John, and Min-su Kim. *Remarks on the Korean language*. T'ap ch'ulp'ansa, 1979.

Reali, Florencia, Morten H. Christiansen, and Padraic Monaghan. "Phonological and distributional cues in syntax acquisition: Scaling up the connectionist approach to multiple-cue integration." In *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 25, no. 25. 2003.

Sampson, Geoffrey. "Writing systems." *London, UK: H utchinson* (1985).

De Saussure, Ferdinand. "Course in general linguistics.(Original publication, 1916)." (1916).

Sereno, Joan A. "Phonosyntactics." *Sound symbolism* (1994): 263-275.

Shi, Rushen, James L. Morgan, and Paul Allopenna. "Phonological and acoustic bases for earliest grammatical category assignment: A cross-linguistic perspective." *Journal of child language* 25, no. 1 (1998): 169-201.

Shillcock, Richard. "The concrete universal and cognitive science." *Axiomathes* 24, no. 1 (2014): 63-80.

Shillcock, Richard, Simon Kirby, Scott McDonald, and Chris Brew. "Filled pauses and their status in the mental lexicon." In *ISCA Tutorial and Research Workshop (ITRW) on Disfluency in Spontaneous Speech*. 2001.

Shinohara, Kazuko, and Shigeto Kawahara. "A cross-linguistic study of sound symbolism: The images of size." In *Annual Meeting of the Berkeley Linguistics Society*, vol. 36, no. 1, pp. 396-410. 2010.

Simner, Julia, Jamie Ward, Monika Lanz, Ashok Jansari, Krist Noonan, Louise Glover, and David A. Oakley. "Non-random associations of graphemes to colours in synaesthetic and non-synaesthetic populations." *Cognitive neuropsychology* 22, no. 8 (2005): 1069-1085.

Sohn, Ho-Min. *The korean language*. Cambridge University Press, 2001.

Sohn, Ho-min, ed. *Korean language in culture and society*. University of Hawaii press, 2005.

Starr, Rebecca Lurie, Alan CL Yu, and Stephanie S. Shih. "Sound symbolic effects in Mandarin and Cantonese personal names and Pokémon names." In *1st Conference on Pokémonastics, Keio University, Tokyo*. 2018.

Sung, Jae-Hyun. "Gradient harmony in Korean ideophones: A corpus-based study." *음성음운형태론연구* 24, no. 1 (2018): 29-50.

Tamariz, Monica. "Exploring systematicity between phonological and context-cooccurrence representations of the mental lexicon." *The Mental Lexicon* 3, no. 2 (2008): 259-278.

Thivierge, Jean-Philippe, and Gary F. Marcus. "The topographic brain: from neural connectivity to cognition." *Trends in neurosciences* 30, no. 6 (2007): 251-259.

Torre, Iván G., Bartolo Luque, Lucas Lacasa, Christopher T. Kello, and Antoni Hernández-Fernández. "On the physical origin of linguistic laws and lognormality in speech." *Royal Society open science* 6, no. 8 (2019): 191023.

Tylén, Kristian, Trecca Fabio, Christina Rejkjær Dideriksen, Byurakn Ishkhanyan, Ewa Dąbrowska, Riccardo Fusaroli, Anders Højen, Dorthe Bleses, Christer Johansson, Christiansen, Morten H. "The Puzzle of Danish" In *The Seventh Conference of the Scandinavian Association for Language and Cognition, SALC 7*. Aarhus, Denmark, 2019.

Van Orden, Guy, and Damian G. Stephen. "Is cognitive science usefully cast as complexity science?." *Topics in cognitive science* 4, no. 1 (2012): 3-6.

Vendler, Zeno. *Adjectives and nominalizations*. No. 5. Walter De Gruyter Incorporated, 1968.

Wong, Kristen Wing Yan, and Yoonjung Kang. "Sound symbolism of gender in Cantonese first names." In *Proceedings of ICPhS*, vol. 19, pp. 2129-2133. 2019.

Wray, Alison, and George W. Grace. "The consequences of talking to strangers: Evolutionary corollaries of socio-cultural influences on linguistic form." *Lingua* 117, no. 3 (2007): 543-578.

Zipf, George Kingsley. *The psycho-biology of language: An introduction to dynamic philology*. Routledge, 1935/2013.

Zipf, George Kingsley. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 1949/2016.