Are Phonetically Similar Korean Words Also Semantically Similar?

Hana Jee

Richard Shillcock

The University of Edinburgh, Edinburgh

Abstract

In connection with the recent research trend in sound-meaning regularity, this article takes several novel approaches. In an effort to extend the current scope to non-European languages, the study focuses on Korean, including its comparatively new orthography. By choosing 74 Swadesh words as its sample, it not only looks at fundamental aspect of the language, but also facilitates future research to compare various languages. To measure them more precisely, criteria for Korean phonology and orthography were created. For semantic distances, Latent Semantic Analysis was used considering its ability to capture authentic dynamics of language. A significant, negative correlation was found between sounds and meanings as well as orthography and meanings, particularly within the semantic space of younger readers. Distinct sounds to indicate potentially confusing objects are considered to be an adaptive mechanism for children to acquire language. The study also numerically revealed how well Korean orthography represents its sounds.

Keywords: Regularity, Korean orthography, Sound-meaning correlation, Euclidean distance, Latent Semantic Analysis

Introduction

The sound-meaning correlations have been investigated in the context of different aspects. Some were interested in the synaesthetic effects that a few pairs of vowels arouse (Sapir, 1929; Brown, Black, & Horowitz, 1955; Taylor & Taylor, 1962; Kim, 1977; Diffloth, 1994 cited in Tsur, 2006; Hinton, Nichols & Ohala, 1995; Tsur, 2006), which has been already applied to some practical fields (Parault & Schwanenflugel, 2006; Lowrey & Shrum, 2007; Coulter & Coulter, 2010; Baxter & Lowrey, 2011; Imai, Miyazaki, Yeung, Hidaka, Kantartzis, Okada, & Kita 2015). Some others focus on the prosodic effects on lexical categories that make it easier to distinguish nouns from verbs. (Cutler & Carter, 1987; Kelly & Bock, 1988; Durieux & Gillis, 2001; Monaghan, Charter, & Christiansen, 2005).

More recently, researchers demonstrated robust statistical correlations between sounds and meanings (Farmer, Christiansen, & Monaghan, 2006; Tamariz, 2008; Fitneva, Christiansen, & Monaghan, 2009; Monaghan, Shillcock, Christiansen, & Kirby, 2014). For example, Tamariz (2008) found that Spanish vowels negatively contribute to systematicity, while consonants and stress structure enhance it. Recently, Blasi, Wichmann, Hammarstrom, Stadler, and Christiansen (2016) showed that certain sounds are connected to basic meanings in the majority of the world languages. The regularity or systematicity found in a few languages seems to be understood from the evolutionary point of view, which is plausible in that it facilitates language acquisition, particularly in early age.

However, it should be pointed out that the major focus is still on European languages. In order either to support the universality of this phenomenon or to emphasise the uniqueness of certain languages, more effort should be put into the study of different types of language. Considering the potential influence of spelling systems on spoken

language and vice versa, (Perry, Ziegler, & Zorzi, 2013; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) languages with unique, independent writing systems such as Arabic, Chinese and Korean would be particularly useful for comparison purposes.

This article investigates the extent to which Korean words are arbitrary and systematic. With a focus on Korean monosyllables, it sheds light on whether there is any regularity among sounds, meanings and the orthography system. Do words with similar sounds have similar meanings? Also, do words with similar characters have similar meanings?

Studying the correlation between sound-distribution and meaning-distribution – and, going one step further, orthography-distribution – may be able to explain a part of Korean language in terms of either arbitrariness or regularity.

Experiment

Material

In total, 74 monosyllabic Korean words were chosen from the Swadesh list (Swadesh, 1952), for various reasons. The Swadesh list is a list of 215 words that are universally observed in (presumably) all languages on Earth. Thus, it provides an opportunity for this experiment to be replicated in other languages. It would be worth noting that these words include many pure Korean words contrasting to Sino-Korean words, which originated from Chinese language. Compared to Sino-Korean words that tend to be compositional and complex, pure Korean words are short and simple. This makes it possible to explain the basic and fundamental aspect of Korean vocabularies. Furthermore, by using the universally common words, the study could compensate for the fact that Latent Semantic Analysis

(LSA) – which was used to measure semantic distances – was not available in Korean.

The Sino-Korean words were excluded. For example, ‘목’ means ‘neck’ in pure Korean but can also be understood as ‘tree’ as Sino-Korean for ‘木’. In addition, all homophones and homographs were omitted (‘눈’ means both ‘eye’ and ‘snow’). We will return to this issue below.

Also, it is worth mentioning that some words were translated to monosyllabic Korean even though this translation might not be optimal, in an effort to enlarge the sample size. For example, ‘all’ is better translated to ‘모든’, but instead the more informal ‘다’ was used. Also, ‘놈’ was chosen for ‘boy’ by the same token. Finally, words whose coda consists of two slots were excluded (‘홀’ and ‘맏다’, for example).

Eventually, 28 nouns plus 46 stems of verbs and adjectives were chosen, and phonological, orthographical and semantic distances between two words were calculated respectively: a total of 2701 (74 x 73 x 1/2) cases.

Phonological Distance

To establish phonological criteria, Harm and Seidenberg (1999) considered 11 features: sonorant, consonantal, voice, nasal, degree, labial, palatal, pharyngeal, round, tongue and radical. However, this does not perfectly fit Korean phonemes. Take ‘ㄷ’, for example, which is alveolar flap /ɾ/ or alveolar approximant /ɹ/, and might be simply perceived as /l/ in English but has a subtly softer sound than /l/. In addition, not including tensed consonants (ㅃ /p/, ㅈ /t/, ㄱ /k/, ㅊ /t͡ʃ/) in the template makes it difficult to efficiently represent the allophone sets (k - k̤ - kʰ, for instance). Last but not least, the position of half consonants such as /w/ and /j/ is not clear, because these are considered obvious vowels in Korean: ‘ㅜ’ and ‘ㅣ’, respectively, albeit different in intensity of the sound.

Therefore, more precise and effective criteria needed to be made and the phonetic template from the high school Korean language textbook was a good alternative. Unlike Harm and Seidenberg (1999), where both consonants and vowels were analysed by identical features, the Korean textbook strictly differentiates consonants from vowels.

As shown in Table 1, each feature was presented as either -1 or 1. One phoneme entails a vector of 20 numbers. The vector of ‘none’ for empty slots consists of twenty -1s. The distance between two words was determined following the method of Monaghan, Christiansen, Farmer, and Fitneva (2010): Euclidean distance. This is used to calculate the shortest distance between two points.

$$d(\mathbf{p}, \mathbf{q}) = d(\mathbf{q}, \mathbf{p}) = \sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + \dots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^n (q_i - p_i)^2}.$$

Table 1: The representation of an allophone set

		ㄱ[k]	ㄷ[k]	ㅋ[kʰ]
Articulation Location	Hard palatal	-1	-1	-1
	Soft palatal	1	1	1
	Labial	-1	-1	-1
	Dental	-1	-1	-1
	Throat	-1	-1	-1
Articulation Manner	Plosive	1	1	1
	Affricate	-1	-1	-1
	Fricative	-1	-1	-1
	Fortis	-1	1	-1
	Lenis	1	-1	-1
	Aspirated	-1	-1	1
	Nasal	-1	-1	-1
	Flow	-1	-1	-1
Vowel Related	Front	-1	-1	-1
	Back	-1	-1	-1
	High	-1	-1	-1
	Middle	-1	-1	-1
	Low	-1	-1	-1
	Round	-1	-1	-1
	Non-round	-1	-1	-1

The allophone sets (shaded), the most apparent characteristic of Korean consonants, are clearly distinguished by this measurement.

Semantic Distance

‘Contextual Co-Occurrence Vector’ (Monaghan et al., 2014), is based on contextual similarity between words, and counts other words that frequently co-occur through statistical analysis of huge corpora (Landauer & Dumais, 1997; Lund & Burgess, 1996). It also includes the possibility of substitution of a word. By pre-processing the overall distribution of a word in a certain context before even analysing correlations between others, it improves the authenticity of results (Landauer, Foltz & Laham, 1998). Therefore, Latent Semantic Analysis (LSA) was chosen as a feasible and efficient way to investigate distances among words.

Unfortunately, however, LSA is not yet available in Korean. Instead, the University of Colorado’s LSA website (<http://lsa.colorado.edu/>) was used, which provides a very convenient tool for measuring the distance between words. Among multiple semantic spaces, the following five were selected: General Reading up to 3rd, 6th, 9th, 12th grade, and 1st-year College. Assuming the Swadesh words are a subset of early-acquired words, this was to observe whether any interesting pattern had arisen according to age.

Orthographical Distance

As one of shallow orthography systems, Korean orthography that was established by the scholarly committee

and released in 1443 is deeply related to phonology. This is because the characters were designed to represent their phonetic characteristics: for example, ‘ㄱ[k]’ shows tongue touching soft palate. Therefore, it would be useful to look at the relation between writing distribution and meaning distribution, which will be discussed further below.

To estimate orthographical distance between two Korean words, a novel method was created: the Stroke Share Rate. Divided into consonants and vowels, this measurement aims to count the number of strokes or lines that two Korean letters share. As illustrated in Figure 1, there are 12 minimum paths that Korean consonants can take.

Figure 1: The stroke share rate: consonants

In the case where some lines always entail others, only one of them was taken into account. For example, since lines c-e and b-g always entail lines d-f and b-f, respectively. It was also considered that the tensed sounds (ㄱ, ㅋ, ㆁ, ㆅ, ㆆ) have basic consonants duplicated.

The Korean vowels in the study can have a maximum of three strokes, corresponding to the writing order. The first basic stroke is written in vertical (ㅣ) or horizontal (ㅡ), while the second shorter stroke can be added either on the left, right (ㅏ, ㅑ), upwards or downwards (ㅓ, ㅕ). The final strokes duplicate either the first stroke (ㅗ, ㅛ, ㅜ, ㅠ) or the second stroke (ㅛ, ㅝ, ㅞ, ㅟ). Although further additional strokes can be added (ㅚ, ㅜ, ㅟ, ㅟ), none of the sample had this type of vowel. Therefore, the minimum number of cases, eight, was taken into account.

Considering the structure of a Korean word, which consists of a consonant, a vowel and a final optional consonant (Figure 2), the number of orthographic features was 30 in total (12 + 8 + 12) which were represented as a single vector. The distances between two words were measured by Euclidean distance as well.

Figure 2: The structure of Korean monosyllables

Results

Sound-Meaning Relation

Table 2 illustrates the correlations between phonology and each semantic space. Importantly, slightly negative correlation was found in all relations, which implies that the more similar the words sound, the farther their meanings are from each other. Another interesting pattern is that as the space represents usage by younger learners, not only does the correlation ρ become higher, but the statistical significance also becomes stronger.

Table 2: The sound-meaning correlations

Semantic Space	Correlation ρ
3 rd grade	-0.06 **
6 th grade	-0.06 **
9 th grade	-0.05 *
12 th grade	-0.03
1 st year college	-0.03

Orthography-Meaning Relation

It was found that the relationship between orthographical and semantic distance is similar to that between phonological and semantic distance (Table 3). First, overall negative correlations are observed. Also, both correlation coefficients and statistical significance are stronger for lower reading age.

Table 3: The orthography-meaning correlations

Semantic Space	Correlation ρ
3 rd grade	-0.07 ***
6 th grade	-0.05 **
9 th grade	-0.04 *
12 th grade	-0.03
1 st year college	-0.02

Sound-Orthography Relation

The fact that sound-meaning relations and orthography-meaning relations showed similar patterns may be due to a very strong positive correlation between Korean phonology and orthography (all p-values < .001).

As Coltheart, et al. (2001) and Perry, Ziegler, and Zorzi (2013) pointed out, it should be considered that we cannot absolutely exclude orthographic influence from sound-meaning correlations and phonological influence from character-meaning correlations. Therefore, a partial correlation test was conducted to exclude these hidden effects from the above results. It was expected to reveal whether and the extent to which sound-orthography correlation affects the correlation between sound-meaning and orthography-meaning, respectively.

Interestingly, Table 4 shows that the hidden orthographical effect is larger than the hidden phonological effect. This implies that if the orthography influence is excluded, the systematicity in sound-meaning relation largely disappears. Therefore, it could be said that the contribution to systematicity in Korean monosyllables comes much more from orthography than from sounds themselves.

Table 4: The Partial correlations

Correlation between	Controlled	ρ
Phonology – Semantics	Orthography	-0.005
Orthography-Semantics	Phonology	-0.04 *

Discussion

Previous studies revealed systematicity with a large sample (Tamariz, 2008; Monaghan, et al., 2010; Monaghan, et al., 2014). Particularly, Tamariz (2008) and Monaghan, et al. (2014) showed the mixed result of negative and positive correlations. The current study is significant in the sense that it showed consistent negative sound-meaning and orthography-meaning correlations. With homonyms and homographs included, the result is expected to show even stronger negative-correlation. There may be the effect of the Swadesh words, which are mostly short and simple, so whose context is easily acquired in early age (Goldfield & Reznick, 1990). In this regard, using the greatest articulatory contrast would be a good strategy to enhance the probability of distinguishing spatially or semantically close words. By doing this, efficiency of distinction would be maximized, even better than arbitrariness. This may explain why the statistics were robust in the youngest semantic space.

The fact that sound-meaning and orthography-meaning relations showed similar patterns appears to be due to the very strong correlation between orthography and sound in Korean. One of the unique features of Hangeul (the Korean alphabet) is that the sets of characters corresponding to the allophones were designed to have a common property, just as the allophones share the same articulation point. Likewise, Korean written vowels also were designed in a way that an addition or omission of strokes reflects phonological additions or omissions. For instance, ㅏ /ia/

possesses one more stroke than ㅑ /a/, indicating there may be an additional phonetic feature. Whereas it has been emphasised that the Korean alphabet was created scientifically and systematically (Diamond, 1994), the current study might be the first effort to show this numerically.

It is also initially surprising that the sound-meaning relationship is much less correlated without orthographical effect. The fact that LSA is generated from textual corpora could partly explain why orthography maps much better on semantics than phonology does. It could be also understood in terms of language change. Considering sounds change easier and faster than writing system, we can make an educated guess that spoken Korean have been changed since 15th century while written forms and their meanings remained mostly the same. To find out mechanism behind this, however, is what further research can shed light on.

Conclusion

One of the main findings of this study was overwhelming negative sound-meaning correlations in contrast to overall positive correlations found in English (Monaghan et al., 2014) and Spanish (Tamariz, 2008). Considering small size of the sample, it is impressive to observe strong statistics. Whereas we agree that strong sound-meaning correlation is very adaptive for the sake of child learning, further research is required as to whether the correlation is negative or positive.

The negative sound-meaning relation could be even better than arbitrary sound-meaning relation to distinguish confusing words. This might be the reason why the stronger correlations were observed in younger semantic ages. Furthermore, it numerically showed that the scientific design of Korean orthography system and that orthographical influence on regularity is now greater than phonological influence.

If replicated with larger sample, the future study could also reveal whether this negative sound-meaning correlation is only a partial phenomenon or can be generally applied to Korean. Future studies could also deal with other languages, particularly those that have influenced or been influenced by the Korean language. The Swadesh words might be a good starting point, in this regard.

References

- Baxter, S., & Lowrey M. T. (2011). Phonetic Symbolism and Children's Brand Name Preferences. *Journal of Consumer Marketing*, 28, 516-523.
- Blasi, D. E., Wichmann, S., Hammarstrom, H., Stadler, P. F., & Christiansen, M. H. (2016). Sound-meaning association biases evidenced across thousands of languages. *PNAS*, 113, 10818-10823.
- Brown, W. R., Black, H. A., & Horowitz, E. A. (1955). Phonetic Symbolism in Natural Languages. *The Journal of Abnormal and Social Psychology*, 50, 388-393.

- Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108, 204-256.
- Coulter, S. K., & Coulter, A., R. (2010). Small Sounds, Big Deals: Phonetic Symbolism Effects in Pricing. *Journal of Consumer Research*, 37, 315-328.
- Cutler, A., & Carter, D. M. (1987). The predominance of strong initial syllables in the English vocabulary. *Computer Speech & Language*, 2, 133-142.
- Diamond, J. (1994). Writing Right. Discover, June. <http://discovermagazine.com/1994/jun/writingright384>
- Durieux, G., & Gillis, S. (2001). Predicting grammatical classes from phonological cues: An empirical test. In J. Weissenborn & B. Höhle (Eds.), *Approaches to bootstrapping: Phonological, lexical, syntactic and neurophysiological aspects of early language acquisition*. Amsterdam: John Benjamins.
- Farmer, T. A., Christiansen, M. H., & Monaghan, P. (2006). Phonological typicality influences on-line sentence comprehension. *Proceedings of the National Academy of Sciences*, 103, 12203-12208.
- Farmer, T. A., Monaghan, P., Misyak, J. B., & Christiansen, M. H., (2009). Phonological typicality influences sentence processing in predictive contexts: A reply to Staub et al. *Journal of Experimental Psychology: Learning, Memory and Cognition*, in press.
- Fitneva, S., Christiansen, M. H., & Monaghan, P. (2009). From sound to syntax: Phonological constraints on children's lexical categorization of new words. *Journal of Child Language*, 36, 967-997
- Goldfield, A. B. & Reznick, J. S. (1990). Early lexical acquisition: rate, content, and the vocabulary spurt. *Journal of Child Language*, 17, 171-183.
- Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading and dyslexia: Insights from connectionist models. *Psychological Review*, 106, 491-528.
- Hinton, L., Nichols, J., & Ohala, J. J. (1995). *Sound Symbolism*. Cambridge: Cambridge University press.
- Imai, M., Miyazaki, M., Yeung, H. H., Hidaka, S., Kantartzis, K., Okada, H., & Kita, S. (2015). Sound Symbolism Facilitates Word Learning in 14-month-olds. *PLoS One*, 10, 1-17.
- Kelly, M. H., & Bock, J. K. (1988) Stress in time. *Journal of Experimental Psychology: Human Perception and Performance*, 389-403.
- Kim, K. O. (1977). Sound Symbolism in Korean. *Journal of Linguistics*, 13, 67-75.
- Landauer, K. T., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104, 211-240.
- Landauer, K. T., Foltz, W. P., Laham, D., (1998). Latent semantic analysis: An introduction to latent semantic analysis. *Discourse Processes*, 25, 259-284.
- Lowrey, M. T., & Shrum, J. L. (2007). Phonetic Symbolism and Brand Name Preference. *Journal of Consumer Research*, 34, 406-414.
- Lund, K., & Burgess, C., (1996). Producing high-dimensional semantic spaces from lexical co-occurrence. *Behavior Research Methods, Instruments, & Computers*, 28, 203-208.
- Monaghan, P., Charter, N., & Christiansen, H. M., (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, 96, 143-182.
- Monaghan, P., Christiansen, H. M., Farmer, T. A., & Fitneva, A. S., (2010) Measures of phonological typicality: Robust coherence and psychological validity. *The Mental Lexicon*, 5, 281-299.
- Monaghan, P., Shillcock, C. R., Christiansen, H. M., & Kirby, S., (2014). How arbitrary is language. *Phil. Trans. R. Soc. B*, 369, 1-12.
- Parault, J. S., & Schwanenflugel, J. P. (2006). Sound-symbolism: A piece in the puzzle of word learning. *Journal of Psycholinguistic Research*, 35, 329-351.
- Perry, C., Ziegler, J. C., Zorzi, M. (2013). A computational and Empirical Investigation of Graphemes in Reading. *Cognitive Science*, 37, 800-828.
- Sapir, E. (1929). A Study in Phonetic Symbolism. *Journal of Experimental Psychology*, 12, 225-239.
- Swadesh, Morris. (1952). Lexicostatistic Dating of Prehistoric Ethnic Contacts. *Proceedings of the American Philosophical Society*, 96, 452-463.
- Tamariz, M., (2008). Exploring systematicity between phonological and context cooccurrence representations of the mental lexicon. *The Mental Lexicon*, 3, 259-278.
- Tsur, R. (2006). Size-sound Symbolism Revisited. *Journal of Pragmatics*, 38, 905-924.
- Taylor, I. K., & Taylor, M. M. (1962). Phonetic Symbolism in Four Unrelated Languages. *Canadian Journal of Psychology*, 16, 344-356.