Comparing Learning Models for Korean Sound-symbolic Vowel Harmony

Background Goldsmith and Riggle (to appear) and Hayes and Wilson (2008) demonstrated that by allowing bigram models of learning to apply over a vowel tier, rather than over strictly adjacent sounds, certain cases of vowel harmony (VH) can be learned. Using the letters A and B to represent vowel classes of opposing harmony, a tier-based bigram model can be made to allow sequences of AA or BB while disallowing sequences of AB or BA. The bigram model, however, runs into problems when faced with intervening neutral (N) vowels. Since sequences AN, BN, NA, NB are all allowed (e.g. BNB, ANA), the bigram model incorrectly allows trigram sequences with mismatched harmony (e.g. BNA or ANB). A different learner developed by Heinz (under review), the Strictly 2-Piecewise (SP2) learner (see also Rogers et al., to appear), eliminates the need for separate tiers to capture VH while simultaneously accounting for instances of long-distance harmony, such as harmony across neutral vowels. This learner keeps track of precedence relations across a word and not simply bigram relations; thus, it correctly predicts a sequence such as BNA to be ungrammatical, as A cannot be preceded in a word by B.

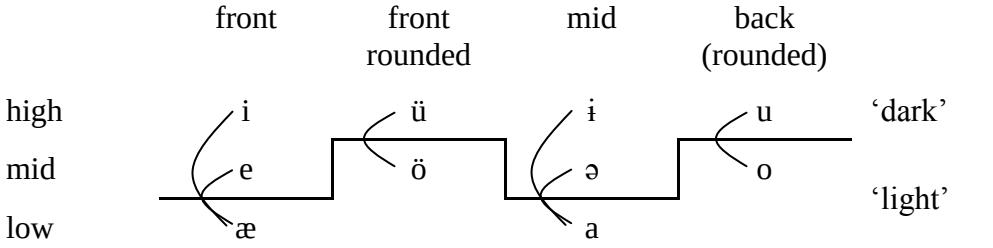
Issue Modern Korean presents an interesting case for VH and poses a difficulty for the above learners, including the SP2 learner. Korean sound-symbolic harmony—one of two VH types in Korean—divides vowels into 'light' and 'dark' vowels (1) (we are not concerned here with the features relevant for this distinction; see, for example, Cho 1994). The literature regularly notes that [i] and [ɨ] are bi-functional. In initial syllables, they behave as dark vowels, disallowing light vowels from occurring later in the morpheme. In noninitial syllables, they behave as neutral vowels, allowing both dark and light harmony to pass over them. Additionally, while these form a natural class [high, nonround], Cho (1994) has also noted that [u] behaves as a neutral vowel to some extent. To our knowledge, no literature has claimed that [ü] behaves as a neutral vowel, leaving the unnatural class [i ɨ u], exluding [ü].

Study We performed a corpus study of sound-symbolic morphemes in Korean to verify the claims about neutral vowels and to verify whether or not [i] and [ɨ] were the only neutral vowels. We made use of all phonetically-unique two- and three-syllable reduplicant bases in the corpus, totaling about 4,000 items. Whereas the literature on Korean VH typically focuses on vowel alternations within a morpheme—which leads to a very slightly different semantic meaning as in (2)—our study examined VH on a phonotactic basis; that is, we examined VH on surface forms and ignored dark-light alternations which lead to different semantic nuances.

Results By examining the data, we were able to empirically verify the following claims: i) VH remains a synchronic process in sound-symbolic morphemes in Modern Korean; ii) neutral vowels behave as dark vowels in initial syllables and as neutral vowels elsewhere; iii) the vowel [u] patterns together with the other neutral vowels; and iv) though the data is limited for [ü], the evidence suggests that it also patterns with neutral vowels. Because most authors (Kim-Renaud 1976, Park 1990, inter alia) have only considered [i] and [ɨ] as neutral vowels in Korean, these results shed light on a largely-unnoticed pattern, while maintaining a natural class for neutral vowels.

Learning Implications Neither tier-based bigram models nor the SP2 model are able to distinguish a word-initial neutral vowel—which functions as a dark vowel in this position— from a non-initial neutral vowel. For example, since word-internal sequences LNL are allowed (L=light, N=neutral), these learners incorrectly predict #NL to be acceptable. Although tier-based trigram models or more sophisticated SP models could describe this, the hypothesis space becomes much larger, making simple generalizations more complex to state and any generalization more difficult to find. Instead, what is needed is only to add to the SP2 model a mechanism for discovering a distinction between word-initial sounds from word-internal ones, and some possibilities are discussed.

1) Sound-symbolic Harmony. The dark-light 'pairs' for vowel alternations within a morpheme are given by curved lines.



(Based on Kim-Renaud (1976))

2) 'dark' vowels pənccək-pənccək 'sparkling, twinkling' (e.g. flash of light)

'light' vowels panccak-panccak 'sparkling, twinkling' (e.g. stars)

References:

Cho, Mi-Hui. (1994). "Vowel Harmony in Korean: A Grounded Phonology Approach." Ph.D. dissertation. Indiana University.

Goldsmith, John and Jason Riggle. (to appear). "Information theoretic approaches to phonological structure: the case of Finnish vowel harmony." In *Natural Language and Linguistic Theory*.

Hayes, Bruce and Colin Wilson. (2008). "A Maximum Entropy Model of Phonotactics and Phonotactic Learning." In *Linguistic Inquiry*, vol. 39, no. 3, pp. 379-440.

Heinz, Jeffrey. (under review). "Learning Long Distance Phonotactics." Manuscript. University of Delaware.

Kim-Renaud, Young-Key. (1976). "Semantic Features in Phonology: Evidence from Vowel Harmony in Korean." In *Chicago Linguistic Society* Vol. 12, pp. 397-412.

Park, Sayhyon. (1990). "Vowel Harmony in Korean." In *Ohak Yongu/Language Research.* Vol. 26, no 3, 469-499.

Rogers J., Heinz J., Bailey G., Visscher M., Wellcome D., Edlefsen M., and Wibel S. (to appear). "On Languages Piecewise Testable in the Strict Sense." In *Proceedings of the 11th Meeting of the Association of Mathematics of Language*.