Modeling Neutral Vowels in Mongolian

Gallagher Flinn, University of Chicago

Keywords: probabilistic modeling, information theory, Mongolian, vowel harmony

This paper shows that using weighted finite state automata (wFSAs) to make generalizations about the vowel system in Khalkha Mongolian show fine-grained distinctions between what it means to be a neutral vowel in a vowel harmony system. /i/, which determines non-pharyngeal harmony when it occurs in the first syllable and is usually treated as transparent to harmony in word-medial syllables, shows weak pharyngeal harmony outside of the first syllable. Its long counterpart /i:/, however, behaves like a non-pharyngeal vowel regardless of its position.

Goldsmith and Riggle (2012) show that vowel harmony is detectable in Finnish based on distributional data by measuring the extent to which including information about long-distance interactions between vowels increases the probability assigned to the data by a weighted FSA. The model is constructed by calculating the extent to which the probability that a pair of vowels can be expected to co-occur based on their individual frequencies and their actual co-occurrence, the pointwise mutual information (PMI). Running a sequence of words from a corpus through the FSA and summing the weights of the arcs on each word's path assigns a probability to the corpus, and the success of the model is judged in terms of how much higher a probability it assigns to the data than a given competing model. This method has the advantage of being built from the bottom up; nothing needs to be assumed about the featural inventory of the language aside from the input of an alphabet of phonemes. It also has the ability to make generalizations about vowel neutrality, something which is not possible in other systems (Hayes and Wilson (2008)). Because they are defined by their lack of effect on the distribution of other vowels, neutral vowels are detected by the way they interfere with interactions between harmonic vowels, and their removal from the autosegmental harmony tier improves the efficiency of the model.

Khalkha Mongolian has a dual vowel harmony system where vowels must harmonize for both [± round] and [± pharyngeal] features, with the exception of /i/, which can co-occur with vowels of any type (Svantesson (2005)) with the exception that when /i/ is in the first syllable of a word, all following vowels must be non-pharygneal. Building on methods developed in Baker (2009) and Goldsmith and Riggle (2012), I build an intersected weighted FSA from conditional probabilities and PMI values gathered from independently collected corpus data. Surprisingly, /i/ does not behave as a truly neutral vowel would be expected to. Even though it determines non-pharyngeal harmony when it appears in the first syllable, /i/ shows positive PMI with pharyngeal vowels and moderately negative PMI non-pharyngeal vowels when it occurs word-medially. The reverse is true of its long counterpart /i:/, which patterns as expected, showing positive PMI with non-pharyngeal vowels and negative PMI with pharyngeal vowels.

This state of affairs implies several problems. First, what is a reasonable way to categorize /i/ given these measurements? Second, why would a long vowel behave differently than a short vowel which should otherwise have the same features? In order to resolve these questions I build a series of new models, first comparing the effects of excluding a given vowel from the vowel tier. These new models show that, although each fails to perform as well as a full vowel tier, the loss of /i/ represnents the lowest drop in the probability assigned to the corpus data. This suggests that, while it cannot be considered truly transparent to harmony, it is the weakest harmonic vowel. Characterizing /i/ is further complicated, however, by conflicting data from a second set of models ignoring individual distinctions between vowels. Instead, every possible 2-partition of the Mongolian vowel inventory is generated, and a new model built for each partition that is only aware of whether each vowel falls into one partition or the other. Of these new simplified models, it is shown that

the division between pharyngeal and non-pharyngeal is the most efficient distinction to make, and that in each of the most efficient models, /i/ appears in the subset with the true non-pharyngeal vowels, further complicating the issue of /i/'s neutrality.

References

Baker, A. C. (2009). Two bayesian approaches to finding vowel harmony. Technical report, Citeseer, http://home. uchicago. edu/ adamc/papers/vharmony.

Goldsmith, J. and Riggle, J. (2012). Information theoretic approaches to phonological structure: the case of finnish vowel harmony. *Natural Language & Linguistic Theory*, 30(3):859–896.

Hayes, B. and Wilson, C. (2008). A maximum entropy model of phonotactics and phonotactic learning. *Linguistic Inquiry*, 39(3):379–440.

Svantesson, J.-O. (2005). *The Phonology of Mongolian*. Oxford University Press.