Submitted to ISMIL 19

Probabilistic phonotactics in Indonesian

Diana Stojanovic

Phonotactics describes permissible sequences of phonemes in a given language. Languages differ in their phonotactic constraints, for instance, English allows longer stretches of consecutive consonants (consonant clusters), while Hawaiian allows only a single consonant between two vowels. Literature suggests that phonotactics influences speech perception, child language acquisition, and second language learning.

While phonotactics can be seen as a set of permissible syllable shapes, or a dichotomy between permissible and impermissible sequences, a more gradient approach distinguishes likely from unlikely sequences.

In this paper I present results of a probabilistic phonotactic analysis of Indonesian. By analyzing collection of stories of 10,000 words, I compare the likelihoods of consonant clusters in different positions in a word (initial, medial, and final) in two scenarios: 1) *words in text* (as a proxy for spoken communication), where frequent words contribute more to the pattern likelihood than rarely occurring words, and 2) *word list*, where each word contributes equally to the pattern likelihood.

Table 1 illustrates results for the likelihoods of word-initial clusters. We see that while both obstruent - sonorant sequences and sonorant-obstruent sequences occur in the language, obstruent-sonorant sequences are more likely (with frequency is used as a proxy for likelihood). Secondly, if frequencies in the lexicon (here, word list) are considered, sonorant-obstruent sequences are not very likely, adhering to the sonority principle. However, keeping in mind that the size of the sample is small, these sequences occur frequently in the text (as an approximation for occurrence in conversations) and possibly have higher weight on our perception and learning.

Table 1. Number of occurrences and percentage of all initial consonant clusters

	word list		**text**	
	occur	percent	occur	percent
obstruent-sonorant	11	84%	44	64%
sonorant-obstruent	1	8%	23	33%
obstruent-obstruent	1	8%	2	0.02%
mb	1	8%	23	33%
ks	1	8%	2	3%

These results are analyzed in view of sonority theory and consonant cluster patterns found in other languages. At the end, implications for language learning and language recognition are outlined.