QUANTIFYING THE INFORMATION CARRIED IN TONAL CONTRASTS IN PHOM*

PHONGSHAK PHOM

Department of Phonetics and Spoken English, EFLU
Tarnaka
Hyderabad, TS 500007, India

CHARLES REDMON†

Department of Linguistics, University of Kansas
Rm. 427, 1541 Lilac Lane
Lawrence, KS 66045-3129, USA

ABSTRACT. Corpus measurements of tonal ambiguity in written Phom as a function of word size, N-gram context, and frequency are reported. A significant inverse relation between the number of syllables in a word and the tonal contrast size was found, while frequency of the preceding context was shown to be positively correlated with the size of the contrast comprising the ambiguity. N-gram context was also inversely related to the probability of disambiguation, but measures of relative entropy revealed this effect to be nonlinear; i.e. gains in information at the bigram were significantly greater than further gains in the trigram context.

KEYWORDS. Phom, tone, lexicon, entropy

1. INTRODUCTION. The resurgence in interest in earlier formulations of the phoneme as fundamentally a unit of lexical contrast (Martinet, 1938), notably in Surendran and Niyogi (2003) and more recently in the work of Oh and colleagues (2013; 2015), has placed new emphasis on the critical asymmetry in any given language's utilization of elements of its phonemic inventory to distinguish items in the lexicon. The present study applies this framework, with its foundations in information theory (Shannon, 1948; Hockett, 1967), to the lexical tone system of Phom.

Phom is a Tibeto-Burman language spoken in the state of Nagaland in Northeast India. It is closely related to Chang and Konyak, and is currently classified within a wider subgroup of languages referred to by Benedict (1976) as the Bodo-Konyak-Jinghpaw group and by Burling (2003) as the 'Sal' languages. While there is little available work on the language, Phom is generally analyzed as exhibiting a ternary lexical tone contrast (high, mid, low[1]), which is similar in inventory structure to languages like Dimasa and Rabha, though the acoustic manifestation of the three tones in Phom shows notable differences

*Presented at HLS22, IIT Guwahati, 2016; Preprint compiled on April 4, 2016
†Corresponding author: redmon@ku.edu
[1]Burling and Phom (1999) describe this as a falling tone, but for simplicity we will refer to it Low throughout the paper, keeping in mind this is a phonological designation rather than a description of the f0 characteristics of the tone.

(Burling and Phom, 1999; Sarmah, 2009; Phom, 2016). What is not known for Phom and many other languages in the region, however, is how tone is utilized in the maintenance of phonological contrast among items in the lexicon. To begin to fill this gap, we look to patterns of orthographic tone ambiguity among minimal pairs in Phom as a window on the behavior of the system.

2. METHODS. A digital corpus of written Phom (7618 tokens, 2635 types) was developed for the purposes of this study from selected chapters of P. Dako Phom's (2009) discourse on language, *Manshah*. Out of the 2635-word lexicon derived from the corpus, 521 tonal minimal pairs were identified by the first author, a native speaker of Phom[2] For each item the number of *tonal variants* (N_T) – alternative meanings of a given word based on tone (i.e. excluding homophones) – was recorded. This value also corresponds in a minimal sense to the *size* of the tonal contrast present in a given segmental word.[3] Words in this set were then presented to the first author in a randomized list with preceding context provided in stepwise fashion; namely, bigrams of each token of each minimal pair were presented, with new values of N_T recorded, and then for items which remained ambiguous the trigram context was presented.[4]

The set of minimal pairs was further analyzed for syllable count, context frequency, tone content (for monosyllables), and morphological structure (for disyllables). Contrast sizes (N_T) were analyzed as stochastic distributions over types (i.e. the derived lexicon) and compared across syllable count (mono-, di-, tri-) and context (unigram, bigram, trigram) conditions.[5] As a further quantification of the general information load on the tone system in Phom, the Shannon entropy of each distribution was computed and compared across conditions via the Kullback-Leibler divergence. All text processing operations were done in Python 3.5, with R 3.2 used for numerical and statistical computations .

3. RESULTS. The mean tonal ambiguity for isolated words in the corpus was 0.176 $(1 - E[P_T];\ P_T = 1/N_T)$, meaning that for a given word, in the absence of an orthographic indicator of tone, a Phom reader has approximately an 82.4% chance of identifying the intended meaning of the word. When types, not tokens, are considered, mean ambiguity was 0.114.

Table 1 displays the mean contrast size (N_T) by syllable count and N-gram context. Two general trends are evident in the relation between N_T and the Syllable/Context conditions. First, in the absence of context there is a monotonic decrease in the size of the contrast with an increasing number of syllables in the word. Second, across syllable conditions there is a substantial decrease in the size of the contrast with greater context.

Results of Kolmogorov-Smirnov tests of stochastic relations between probability distributions revealed the expected disambiguating effect of context, with unigrams more ambiguous than bigrams $(p < 0.001)$ and bigrams more ambiguous than trigrams $(p < 0.001)$. Effects of syllable count were not found to be constant across contexts, however.

[2]These were words where the orthographic representation was ambiguous between 2 or more tonally distinct words.

[3]We emphasize this measure as a *minimal* estimate because we are not accounting for tonal information which does not represent the minimal difference between two words but nevertheless is encoded in the lexical entry.

[4]No further context was provided beyond the trigram as 95% of items were fully disambiguated at this point.

[5]Unambiguous unigrams are not included in the analysis as their effect is assumed to be constant across conditions.

	Unigram	Bigram	Trigram
Monosyllable	2.70 (2.63)	1.35 (1.58)	1.06 (1.09)
Disyllable	2.61 (2.54)	1.61 (1.58)	1.10 (1.08)
Trisyllable	2.30 (2.26)	1.45 (1.47)	1.09 (1.08)

Table 1: Mean token (and type) values of N_T by syllable count and N-gram context.

Figure 1 displays cumulative density functions (CDFs) of contrast size by context and syllable count. In the unigram context the relation $tri < di < mono$ in N_T was significant ($p_{m-d} = 0.019$; $p_{d-t} < 0.001$), but in the bigram context the relation was lost for mono- and disyllables; i.e. $tri < di = mono$. Only mono- and trisyllables were found to differ in N_T in the trigram context $(p = 0.025)$.

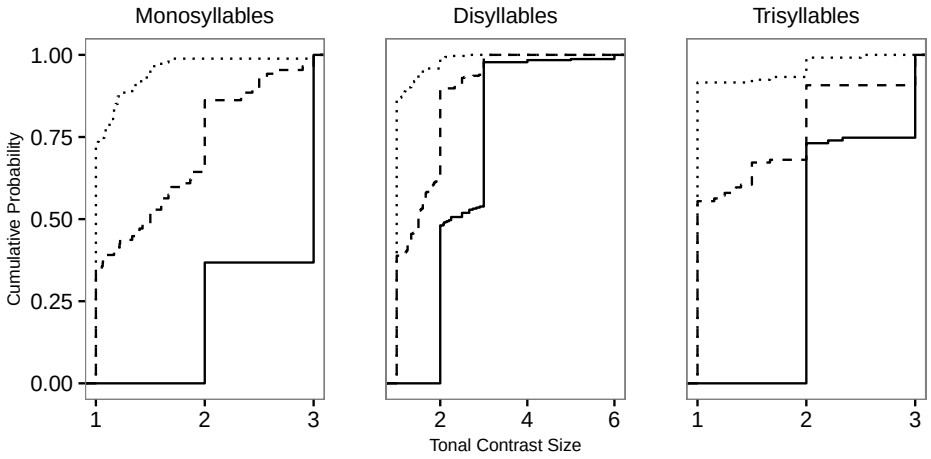

Figure 1: CDFs of contrast size N_T in unigram (solid), bigram (dashed), and trigram (dotted) contexts.

Of greater interest to this study, however, was the relative improvement in disambiguation with the addition of context. This effect was explored by means of the difference in contrast sizes between $N + 1$ and N-grams; e.g. unigram N_T – bigram N_T. Overall the gain in predictability of tone with the addition of a single preceding word (bigram context) was significantly greater than further gains in the trigram context $(p < 0.001)$, and this pattern was consistent across syllable counts $(p < 0.001)$.

Differences between syllable counts within context conditions, however, were not uniform. The gain in bigram context was lower for trisyllables relative to di- and monosyllables $(p < 0.05)$ though no significant difference was found between the latter two. Trigram gains showed the same $tri < di = mono$ pattern, though recorded differences were more robust than in the bigram context $(p < 0.01)$. This finding is not surprising given that most trisyllables were already fully disambiguated at the bigram, making further context in the

trigram relatively superfluous.

Finally, the effect of context frequency on disambiguation was explored in the bigram condition by measuring Kendall's rank correlation between the log-transformed frequency of the preceding word and the tonal contrast size. An overall positive correlation between context frequency and N_T was found ($\tau = 0.10$, $p < 0.001$), with results consistent across syllable counts ($0.064 < \tau < 0.135$, $p < 0.01$). Thus tonal ambiguity was reduced when the target word was presented in a context where the preceding word was infrequent. This finding meets the general expectation that low-frequency words are likely to appear in a more restricted set of bigrams in the language, making their presence more informative when they do appear.

3.1. System entropy. As the above analysis is ultimately a measure of the degree of uncertainty in the mapping between segmental words and their tonally specified counterparts in the lexicon, the overall information contained in these contrasts can be estimated as the entropy of the probability distributions P_T. The Shannon entropy of P_T across syllable counts decreased with context as expected, from 266 to 126 to 22 bits ($D_{KL}(uni||bi) = -169$; $D_{KL}(bi||tri) = -110$). The asymmetry in information gain between bigram and trigram contexts, however, was greater for trisyllables than di- or monosyllables ($D_{KL}(uni||bi)$ / $D_{KL}(bi||tri) = 2.11$ for trisyllables, as compared with 1.58 and 1.38 for mono- and disyllables, respectively).

3.2. Lexical tone distribution. Thus far our examination of the tone system in Phom has been agnostic to the exact tonal categories participating in the contrast, focusing instead on the size of the contrast as a marker of the function of lexical tone more broadly. The degree to which each tone (a) is present in the source ambiguity, and (b) emerges as the disambiguated outcome, however, is relevant to the behavior of the tone system and thus was explored preliminarily for the monosyllabic minimal pairs. The distribution of tones on disambiguated lexical items was relatively balanced, with high-toned words comprising 30% of the set, low tones 40%, and mid tones 30%. At the source ambiguity, however, while all contrasts sets were represented (HML, HL, HM, ML), the ternary contrast was most common (66%), while the ML contrast was least utilized and far less common than the other two binary contrasts (4% as compared with 15% each for HL and HM).[6]

4. CONCLUSIONS. We have shown that while there are indeed asymmetries in the size and distribution of minimal tone contrasts across the Phom lexicon, these asymmetries are not unsystematic. The degree of ambiguity is predictably regulated by the size of the word and the context in which the word appears. In future work we aim to study other sources of information such as syntactic structure and target frequency to further test the bounds on the role of tone in lexical contrast.

ACKNOWLEDGMENTS. We would like to thank Dr. Indranil Dutta for providing helpful feedback on earlier versions of this work.

[6]Here we use the notation HML to indicate that a segmental word may be realized as high-, mid-, and low-toned lexical items, HL for high- and low-tone items, and so on. This is not to be confused with contour tone descriptions which sometimes use HL to indicate a falling tone.

REFERENCES

Benedict, P. (1976). Sino-Tibetan: Another look. *Journal of the American Oriental Society*, 96(2):167–197.

Burling, R. (2003). The Tibeto-Burman languages of Northeastern India. In Thurgood, G. and LaPolla, R., editors, *The Sino-Tibetan Languages*, Routledge Language Family Series, chapter 11, pages 169–191. Routledge, New York, NY.

Burling, R. and Phom, L. (1999). Phom phonology and word list. *Linguistics of the Tibeto-Burman Area*, 21(2):13–42.

Hockett, C. (1967). The quantification of functional load. *Word*, 23(1–3):300–320.

Martinet, A. (1938). La phonologie. *Le Français Moderne*, 6:131–146.

Oh, Y., Coupé, C., Marsico, E., and Pellegrino, F. (2015). Bridging phonological system and lexicon: Insights from a corpus study of functional load. *Journal of Phonetics*, 53:153–176.

Oh, Y., Pellegrino, F., Coupé, C., and Marsico, E. (2013). Cross-language comparison of functional load for vowels, consonants, and tones. In *Proceedings of Interspeech 2013*, pages 3032–3036.

Phom, P. (2016). *An acoustic study of lexical tones in Phom*. PhD thesis, The English and Foreign Languages University, Hyderabad, India. In preparation.

Phom, P. D. (2009). *Manshah*. Popular Printing Press, Dimapur, Nagaland.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria.

Sarmah, P. (2009). *Tone systems of Dimasa and Rabha: A phonetic and phonological study*. PhD thesis, University of Florida, Gainesville, FL.

Shannon, C. (1948). A mathematical theory of communication. *The Bell System Technical Journal*, 27:379–423.

Surendran, D. and Niyogi, P. (2003). Measuring the usefulness (functional load) of phonological contrasts. Technical Report TR-2003, Department of Computer Science, University of Chicago.