Quantifying the information carried in tonal contrasts in Phom

Phongshak Phom
English and Foreign Languages University
Hyderabad, TS, India

Charles Redmon
University of Kansas
Lawrence, KS, United States

The 22nd Himalayan Languages Symposium: IIT Guwahati, 2016

Background

- Phonemic contrasts are commonly treated as representing equivalent distinctions among abstract units in the *inventory* of a language, but at their conception contrasts were fundamentally *lexical* (Martinet, 1938)
- Within information theory, elements of the code are assumed to be utilized asymmetrically in conveying messages (Shannon, 1948; Hockett, 1967)
- Recent work has applied this perspective to cross-linguistic comparisons of contrast structures (Surendran & Niyogi, 2003; Oh et al., 2013, 2015), and to simulations of language change (Wedel et al., 2013)
- But the *role* of contrasts is rarely studied in under-documented languages
- Phom is one such case: a Tibeto-Burman language spoken in Nagaland with a ternary (high, mid, low) lexical tone contrast (Burling & Phom, 1999)
- Written Phom does not mark tones, making orthographic ambiguity one window on the unique information contributed by the tone system

Methods

- A 7,618-word corpus of written Phom based on selected chapters from *Manshah* (Phom, 2009) was developed for this study
- 521 tonal minimal pairs were identified in the 2,635-word corpus-derived lexicon (all data were processed in Python 3.5 and analyzed in R 3.2)
- The contrast size N_T – the number of different lexical items (excluding homophones) represented by a given orthographic word – was recorded for each *token* ($n = 2,222$) of the 521 minimal pair *types* in the corpus
- Values of N_T were then updated as the first author was given the context in which these words occurred in the corpus in the following stepwise procedure:
  - *Unigram* (no context) → *Bigram* (preceding word) → *Trigram* (preceding 2 words)
- Contrast size estimates defined a probability distribution from which effects of Word Length (syllable count, *mono–tri*) and Context (N-gram size) on tonal disambiguation could be measured

Results

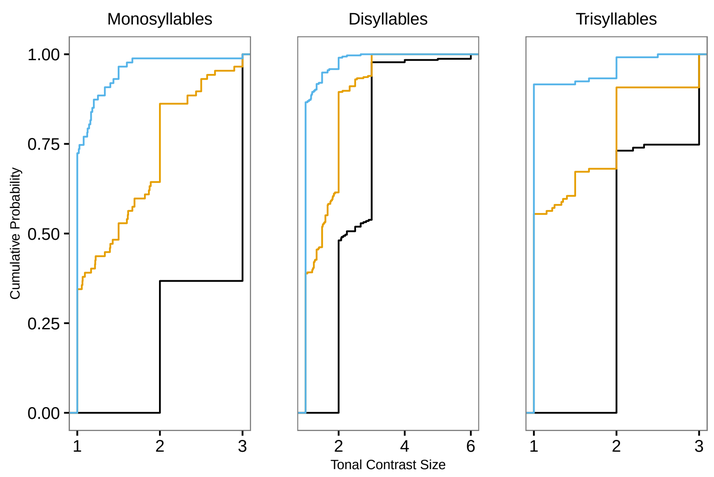

Figure 1: CDFs of contrast size in unigram (black), bigram (orange), and trigram (blue) contexts.

- Kolmogorov-Smirnov tests revealed a significant effect of Context on ambiguity: $trigram < bigram < unigram$ ($p < 0.001$)
- Effects of Word Length were not consistent across contexts
  - Unigram: $tri < di < mono$ in N_T ($p_{m-d} = 0.019$; $p_{d-t} < 0.001$)
  - Bigram: $tri < di = mono$ ($p_{d-t} = 0.006$; $p_{m-t} = 0.012$)
  - Trigram: $tri < mono$ ($p_{m-t} = 0.025$)
- Information gain ($N_{T(n-1)} - N_{T(n)}$) with context was also studied as a stochastic process, with gain at the bigram significantly greater than gain at the trigram ($p < 0.001$)
  - This result was consistent across word lengths ($p < 0.001$)
  - Trisyllables showed smaller gains relative to di/monosyllables at both bigram ($p < 0.05$) and trigram ($p < 0.01$) contexts, though this result is primarily due to their lower overall starting values of N_T
- Context frequency in bigrams (i.e. frequency of the preceding word) was also shown to significantly modulate contrast size, with positive Kendall correlations between frequency and N_T found both overall ($\tau = 0.10$, $p < 0.001$) and within word lengths ($0.06 < \tau < 0.14$, ps < 0.01)
- The inverse of contrast size ($P_T = 1/N_T$) was introduced as a measure of disambiguation probability (conceived as a lower bound on lexical predictability) for the study of tone system entropy

System Entropy

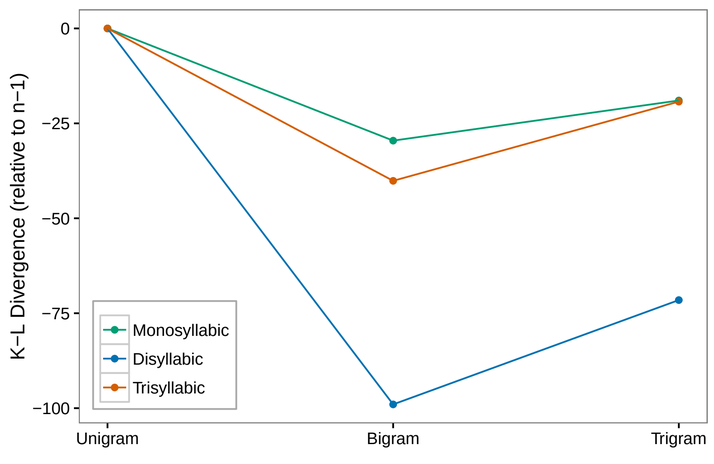

Figure 2: Relative entropy (the Kullback-Leibler divergence, $P_{T(n-1)}||P_{T(n)}$) by word length.

- Across word lengths, the Shannon entropy of P_T decreased nonlinearly from 266 to 126 to 22 bits with increasing context (unigram–trigram)
- Trisyllables showed the greatest asymmetry in information gain with context, with the ratio of relative entropy from bigram to trigram in the following relation: *di* (1.38) < *mono* (1.58) < *tri* (2.11)

Lexical Tone Distribution (monosyllables)

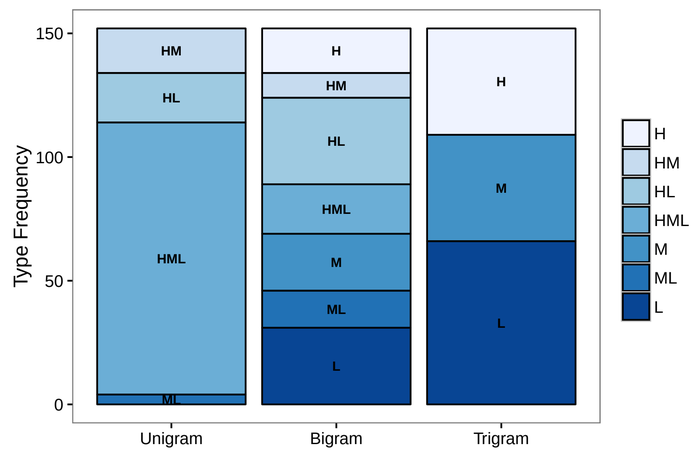

Figure 3: Distribution of potential tonal variants of monosyllables among unique paths (i.e. distinct contrasts: HM, HL, HML, ML) from ambiguous unigram to disambiguated bi/tri-gram.

- Binary contrasts involving the mid tone (HM, ML) were the least common, suggesting H and L might be more prominent under ambiguity
- Among disambiguated items, however, the three tones comprising the contrast were relatively evenly distributed: H – 30%; M – 30%; L – 40%

Morphology (disyllables)

- Disyllabic words formed via compounding exhibit asymmetries in the tonal variant of each constituent
- For example, the mid-tone variant of the second constituent in *yemshing* is more productive than the low ('stuffed up') or high ('to press') variants

yēm-shíng	*yém-shīng*	*yèm-shīng*
search-press	dry-place	animal-place
'to search for'	'a place to dry'	'a place for animals'

- For *vangdhum*, the first constituent is fixed, meaning the ambiguity is completely driven by the tone on the second constituent, *dhum*

váng-dhùm	*váng-dhúm*
rain-shade	rain-visit
'shade, shelter'	'rain visitation'

Conclusions

- The lexical role of the tone system in Phom, being more precisely quantified relative to effects of context (among others), may now serve as a reference for analyses of other tone systems in the region
- Future work should include syntactic and semantic constraints on N_T

References

References will be made available upon request.

Contact

E Phongshak Phom: phongshak@gmail.com
Charles Redmon: redmon@ku.edu
W http://redmonc.github.io