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

- Phenomenon of 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; Hockeit, 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 (Wosh 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 (Builing & Phon, 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 Manusah (Phom, 2005) 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

- Effects of WORD LENGTH were not consistent across contexts:
  - Unigram: $tri < di < mono$ in $N_T (p_{0.01} = 0.019, p_{0.001} = 0.001)$
  - Bigram: $tri < di < mono (p_{0.01} = 0.006, p_{0.001} = 0.012)$
  - Trigram: $tri < mono (p_{0.001} = 0.025)$

- 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 ($r = 0.30, p < 0.001$) and within word lengths ($0.06 < r < 0.14, ps < 0.01$).

- The inverse of contrast size ($Pr = 1/N_T$) was introduced as a measure of disambiguation probability (conceived as a lower bound on lexical predictability) for the study of tone system entropy.

Lexical Tone Distribution (monosyllables)

- 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 $N_T$.

Conclusions

- The contrast size 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

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