Advances in prosodic annotation:

A test of inter-coder reliability for the RaP (Rhythm and Pitch) and ToBI (Tones and Break Indices) transcription systems

Mara Breen (MIT), Laura Dilley (OSU), Edward Gibson (MIT),
Marti Bolivar (MIT), John Kraemer (MIT)
mbreen@mit.edu or dilley.28@osu.edu

INTRODUCTION

- The importance of prosodic factors in understanding and producing language is well-recognized by sentence processing researchers. However, the relationship between acoustic factors and the perception of prosodic events is complex (e.g., Pierrehumbert 1980). Therefore, a useful and practical means for investigating prosody and sentence processing is through annotation of prosodic information.
- This study investigates inter-coder reliability for two prosodic annotation systems:
  - The Tones and Break Indices, or ToBI, system (Silverman et al. 1992)
  - The Rhythm and Pitch, or RaP, system (Dilley and Brown 2005)
- This study addresses limitations on previous evaluations of ToBI (e.g. Pitrelli et al. 1994, Yoon et al. 2004), including small corpora and/or a small numbers of coders, while presenting the first evaluation of RaP.

METHOD

Training of Coders

- Training consisted of reading manuals and annotating digital practice files for ToBI and RaP.
- Initial training lasted 1-2 weeks and included one-on-one meetings with experts.
- Coders were then tested on 60-90 seconds of varied speech.
  - Annotations were evaluated by experts.
  - Trainees had to achieve a specified level of proficiency before beginning corpus annotation.
- Bi-weekly group-labeling (of non-corpus material) with expert coders continued throughout corpus labeling.

Corpus

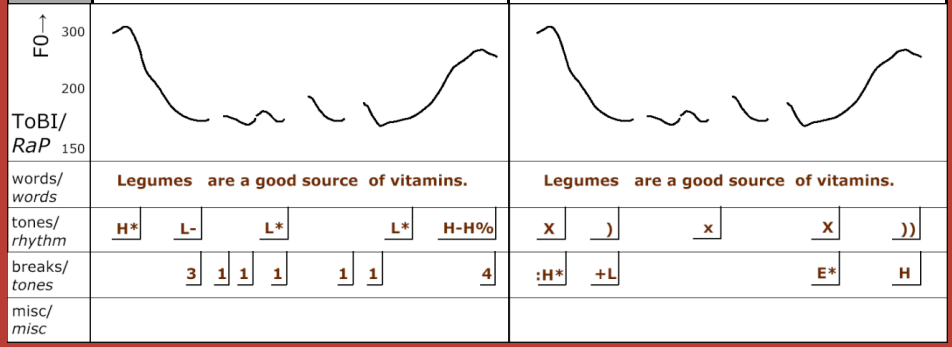
- Each file labeled by 2-5 coders; average 3.9
- Corpus consisted of both spontaneous (CallHome, LDC 1997) and read (Boston Radio News Corpus, Ostendorf, et al 1995) speech files
- ToBI: 44 minutes (22 spontaneous, 22 read)
- RaP: 22 minutes (10 spontaneous, 12 read)

Data analysis

- Agreement was determined using two metrics:
  - Coder-agreement-pairs per syllable (CAP/S): Total agreement corresponds to the number of pairwise comparisons between coders for a syllable which agree, divided by the total number of comparisons.
  - Kappa statistic (K): $K = (P_o - P_c)/(1 - P_c)$, where P_o is the percent agreement between coders and P_c is the percent agreement predicted by chance.
- The following agreement analyses were conducted:
  - *Beat presence (RaP only)*: Whether a syllable was a beat (X or x) or not a beat.
  - *Beat strength (RaP only)*: Whether a syllable was a strong beat (X), weaker. beat (x), or not a beat.
  - *Pitch accent presence*: Whether a syllable had a pitch accent or not.
  - *Pitch accent type*: Whether a syllable was a H*, L*, or unaccented (ignoring leading/trailing tones).
  - *Phrasal boundary presence*: Whether a phrasal boundary was present or not at a syllable juncture.
  - *Phrasal boundary strength*: Whether a phrase boundary is full (or big), intermediate (or small) boundary, or not present.

SYSTEM COMPARISON

Prosodic attribute		ToBI	RaP
	Rhythm	Does not capture rhythmic prominence	Captures three levels of rhythmic prominence X = strong beat; x = weaker beat; [no label] = not a beat
		Does not distinguish "rhythmic prominence" and "pitch accent"	Distinguishes "rhythmic prominence" and "pitch accent"
	Pitch accent	A pitch accent may be indicated with or without a pitch change	A pitch accent may be indicated only in the presence of a pitch change
		Does not distinguish levels of pitch accent strength	Distinguishes multiple levels of pitch accent strength
		Eight kinds of tonal labels: H*, L*, L+H*, L*+H, H+!H*, !H*, L+!H*, L*+!H	Six kinds of tonal labels: 1. H*, L*, E* = indicated on a rhythmically strong syllable 2. H, L, E = indicated on a rhythmically weak syllable (used with '+' notation)
		Distinctions among tonal labels are based on multiple perceptual and acoustic factors	Distinctions among tonal labels are based on perceived direction of pitch movement (rising, falling or level)
		Labeling is based on auditory perception + visual F0	Labeling is based on auditory perception only
		Discourse-relevant factors, e.g., size of pitch excursion, are captured implicitly or not at all	Discourse-relevant factors, e.g., size of pitch excursion, are captured explicitly
		Does not accommodate recent psycholinguistic and phonetic evidence about perceptual categories	Accommodates recent psycholinguistic and phonetic evidence about perceptual categories
	Phrasing	Redundancy and interdependency exists between phrasal boundary labels and tonal labels 1. Indicating a phrasal boundary requires indicating a tonal event at the same location and vice versa 2. Every phrasal constituent must contain a pitch accent	No redundancy or interdependency exists between phrasal boundary labels and tonal labels
		Annotating phrasal boundaries is *usually* based on perceived disjuncture	Annotating phrasal boundaries is *always* based on perceived disjuncture
		Three levels of disjuncture for phrasal boundaries: 1. [L-L%, H-H%, L-H%, H-L%] + 4 = big boundary 2. [H-, L-, !H-] + 3 = small boundary 3. [no tonal label] + [0, 1, or 2] = no boundary	Three levels of disjuncture for phrasal boundaries: 1.)) = big boundary 2.) = small boundary 3. [no label] = no boundary 4. H, L, E = optionally used singly or in sequence *if* there is accompanying tonal change
		Different tonal labels indicate pitch movement due to phrasal boundaries and pitch accents	The same tonal labels indicates pitch movement due to phasal boundaries and pitch accents
Theory		Pierrehumbert (1980), Beckman and Pierrehumbert (1986)	Dilley (2005)
Training		Training set includes a manual and digital audio files (Beckman and Ayers-Elam 1997)	Training set includes a manual and digital audio files (Dilley and Brown 2005)



RESULTS

Prosodic attribute		Agreement Type	Agreement CAP/S ToBI	Agreement CAP/S RaP	Agreement Kappa* ToBI	Agreement Kappa* RaP
	Rhythm	Beat presence	N/A	90%	N/A	0.80
		Beat strength	N/A	79%	N/A	0.65
	Pitch accent	Pitch accent presence	87%	86%	0.71	0.71
		Pitch accent type	80%	80%	0.68	0.65
	Phrasing	Phrasal boundary presence	88%	92%	0.66	0.74
		Phrasal boundary strength	76%	84%	0.40	0.61

*Agreement	Poor	Fair	Moderate	Substantial	~Perfect
Kappa	0-.2	.2-.4	.4-.6	.6-.8	.8-1.0

DISCUSSION

- RaP permits reliable coding of speech rhythm, while ToBI does not permit coding of speech rhythm.
- Agreement for coding phrasal boundaries is higher in RaP than in ToBI. This may be because in RaP, boundaries are based solely on perceived disjuncture, while in ToBI, boundaries are based on perceived disjuncture *and* tonal labels.
- Agreement levels for coding pitch accents are comparable in both systems.
- The RaP annotation system presents a viable alternative to ToBI for investigating prosody in sentence processing research.

REFERENCES

Beckman, M., and Ayers-Elam, G. (1994) "*Guidelines for ToBI Labelling.*" v. 2. Ohio State University. Available at http://www.ling.ohio-state.edu/research/phonetics/E_ToBI/.

Beckman, M. and Pierrehumbert, J. (1986) *Intonational structure in Japanese and English.* Phonology Yearbook, 3, 255-309.

Dilley, L. (2005) *The phonetics and phonology of tonal systems*, PhD thesis, MIT.

Dilley, L. and Brown, M. (2005) The RaP Labeling System, v. 1.0, MIT. Available at http://faculty.psy.ohio-state.edu/pitt/dilley/rap-system.htm.

Linguistic Data Consortium (1997) "CALLHOME American English Speech."

Ostendorf, M.F., Price P. J., Shattuck-Hufnagel S. (1995) *The Boston University Radio News Corpus.* Technical Report No. ECS-95-001, Boston University.

Pierrehumbert, J.B. (1980) *The phonology and phonetics of English intonation*, PhD thesis, MIT.

Pitrelli, J., Beckman, M. & Hirschberg, J. (1994) Evaluation of prosodic transcription labeling reliability in the ToBI framework. In *Proceedings of the International Conference on Spoken Language Processing*,123-126.

Silverman, K., Beckman, M., Pitrelli, J. Ostendorf, M. Wightman, C. Price, P., Pierrehumbert, J. & Hirschberg, J. (1992) ToBI: A standard for labeling English prosody. In *Proceedings of the International Conference on Spoken Language Processing*, 867-870.

Yoon, T., Chavarria, S., Cole, J., & Hasegawa-Johnson, M. (2004) Intertranscriber reliability of prosodic labeling on telephone conversation using ToBI. In *INTERSPEECH-2004*, 2729-2732.