Prosody perception by naïve listeners:

Evidence from a large multi-transcriber reliability study

Yoonsook Mo
Jennifer Cole
Eun-Kyung Lee
University of Illinois at Urbana-Champaign

Prosody perception

- How do ordinary listeners perceive prosody?
  - Are there differences across listeners in how they perceive the prosody for same utterance?
  - Are there differences in prosody perception based on the speaker?
- What properties of an utterance determine how prosody is perceived?
  - acoustic, phonological, syntactic, semantic, pragmatic…?

Why it matters

- Interpretation of prosody is important for speech comprehension
  - Prominence codes information status
  - Prosodic phrasing segments speech into chunks that cohere syntactically or semantically

Why it matters

- Prosody also conditions variation in the realization of consonants and vowels.
  - How does prosodically conditioned variation affect speech recognition?
  - …in spontaneous speech?

Methods in prosody research

- Determine the prosodic events in an utterance
  - Location and tune of prominences
  - Location, strength and tune of boundaries
- Determine how listeners perceive those prosodic events
- Determine the correlates of prosody in linguistic features at various levels of analysis.

Methods in prosody research

Q: How to?

- Determine the prosodic events in an utterance
  - Location and tune of prominences
  - Location, strength and tune of boundaries

A: Prosodic transcription

  - Is it reliable?
  - Is it feasible?

Prosody transcription studies

Transcriptions are judged to be reliable if independent transcribers agree on the location and type of prosodic events.

- High agreement rates between transcribers on the same utterance(s) indicate:
  - Speakers produce salient acoustic cues to prosody, and
  - Listeners perceive prosody similarly.

OR… Perceived perception is determined by "higher" level structure, and does not depend directly on acoustic cues.

Prosody transcription studies

- Limitations of prior studies
  - Materials: single, simple sentences or read speech (Streefkerk et al. 1997, 1998)
  - Transcribers: few prosodically trained (Yoon et al. 2004)
  - Procedure (Buhmann et al. 2002; Yoon et al. 2004)
    - Aided by visual inspection
    - Complex annotation scheme
    - Transcriber may choose to listen as many times as wanted
  - Analysis:
    - simple agreement scores --- don't model chance agreement
    - Cohen's inter-rater agreement scores --- only pairwise analysis

An alternative method

- Prosody transcription that is fast, reliable, and applicable for spontaneous speech.
- A coarse-grain transcription that locates prosodic events.
- A transcription that reflects inter-transcriber agreement through probabilistic prosody labels.

Naïve Prosody Transcription

- ***The transcribers***: large numbers of transcribers who are naïve with respect to prosodic theory and the goals of our research, i.e., “ordinary listeners”.
- ***The transcriptions***: locate prominence and boundary events, ignoring differences in *type* (i.e, tune, strength)
- ***The analysis***: evaluates variation in prosodic transcription across listeners, identifying regions of agreement, and assigning probabilistic prosody labels

Naïve Prosody Transcription

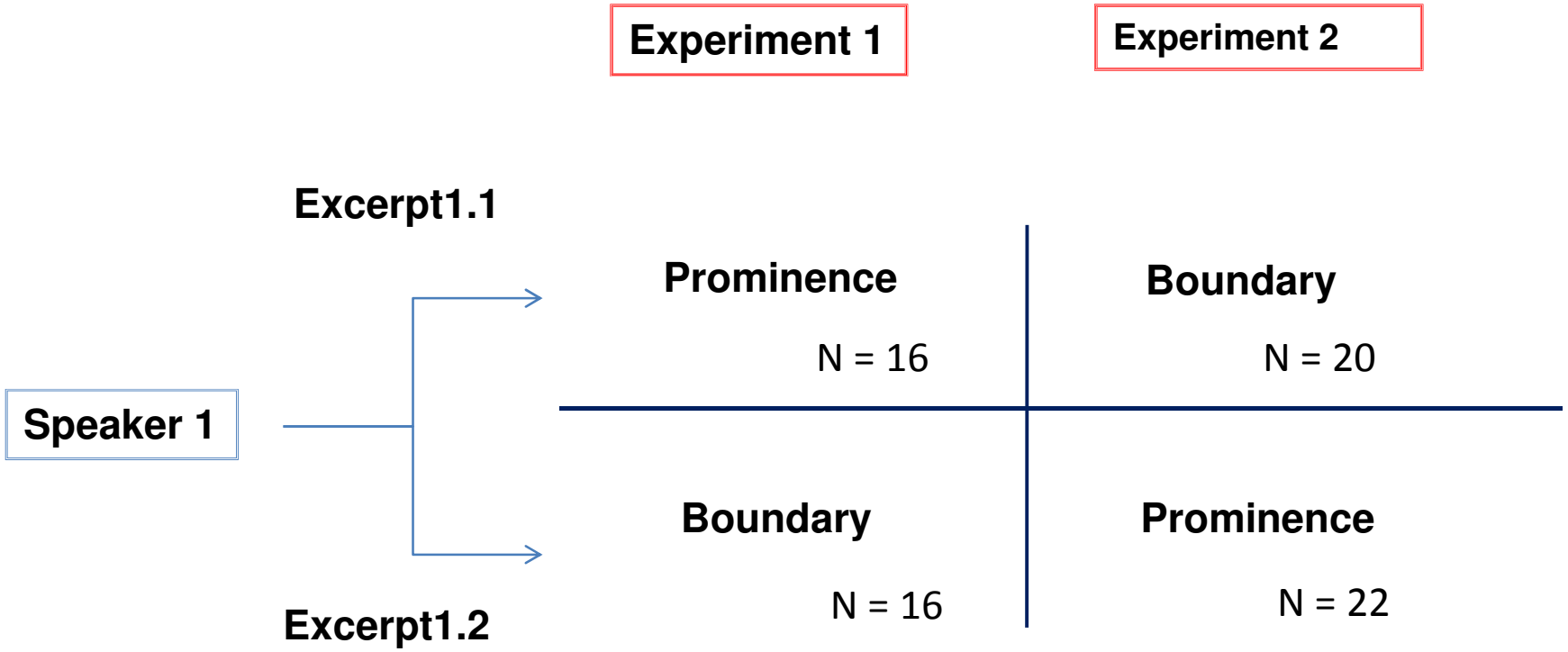
- ***Speed:*** Real time comprehension to diminish strategic analysis
- ***Reliability:*** Transcription reliability measured using Fleiss' Kappa statistic to calculate agreement rates for multiple (> 2) transcribers.

Present study

- Transcription of speech excerpts from the Buckeye Corpus of American English spontaneous speech (Pitt et al. 2007)
- A large number of naïve transcribers
  - 74 UIUC undergraduates…. and growing
- Real time transcription
- No visual inspection of speech display
- Simple annotation scheme

Materials

- 38 short excerpts (about 20 sec. each)
  - 19 speakers x 2 excerpts each



Annotation scheme

Definitions

- Prominence: words that “stand out” from other words
- Boundary: words that demarcate speech “chunks”

Prosodic mark-up on printed transcript of each excerpt:

- Prominence: word <u>word</u> word
- Boundary: word | word word…

– Subjects could make changes by crossing out markings.

- word ~~<u>word</u>~~ word
- word ~~|~~ word word…

Procedure

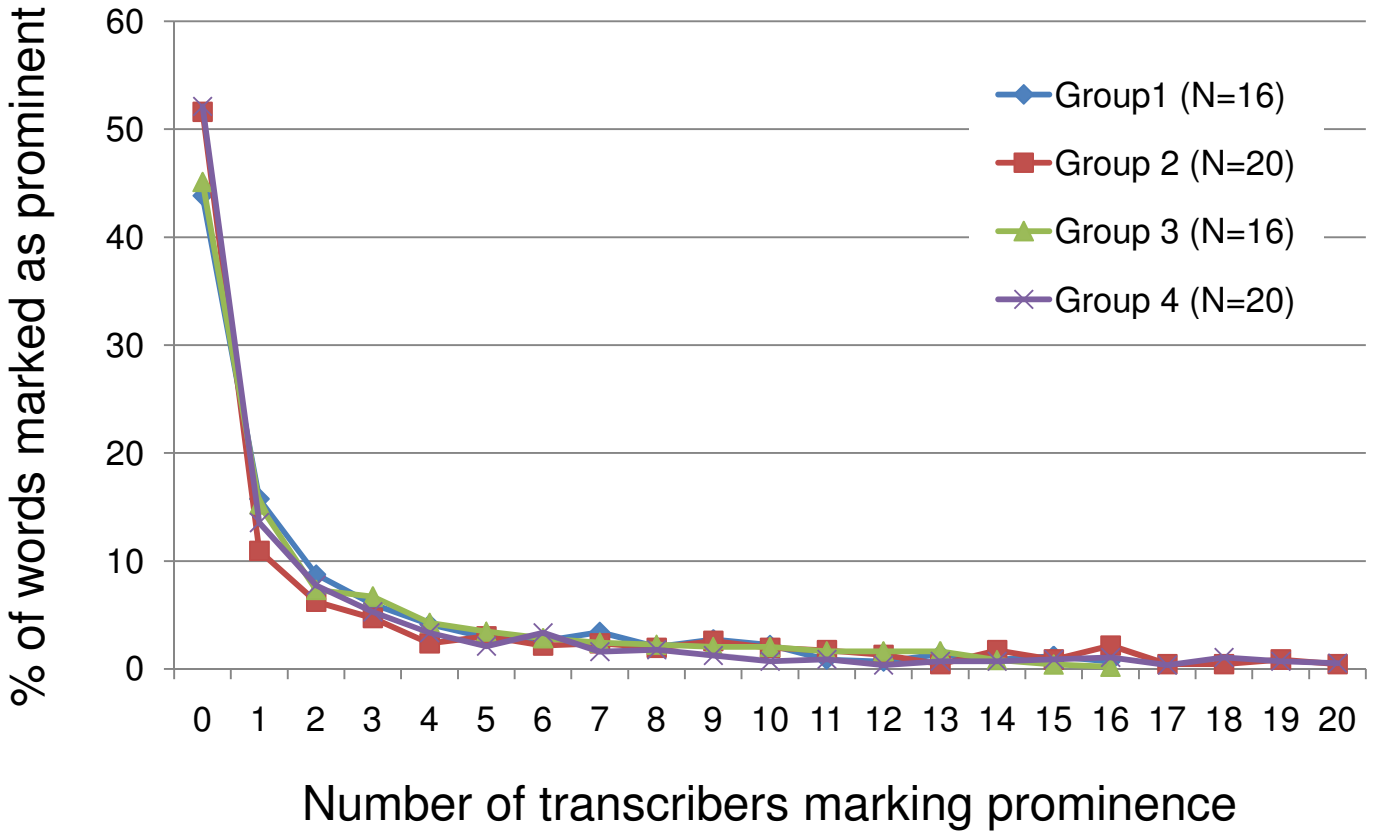
- Sound files played through headphones, ***no visual speech display***
- Transcription done in real time, with two listening passes
- Transcribers assigned to two groups.

  Group 1: Prominence – Boundary

  Group 2: Boundary - Prominence

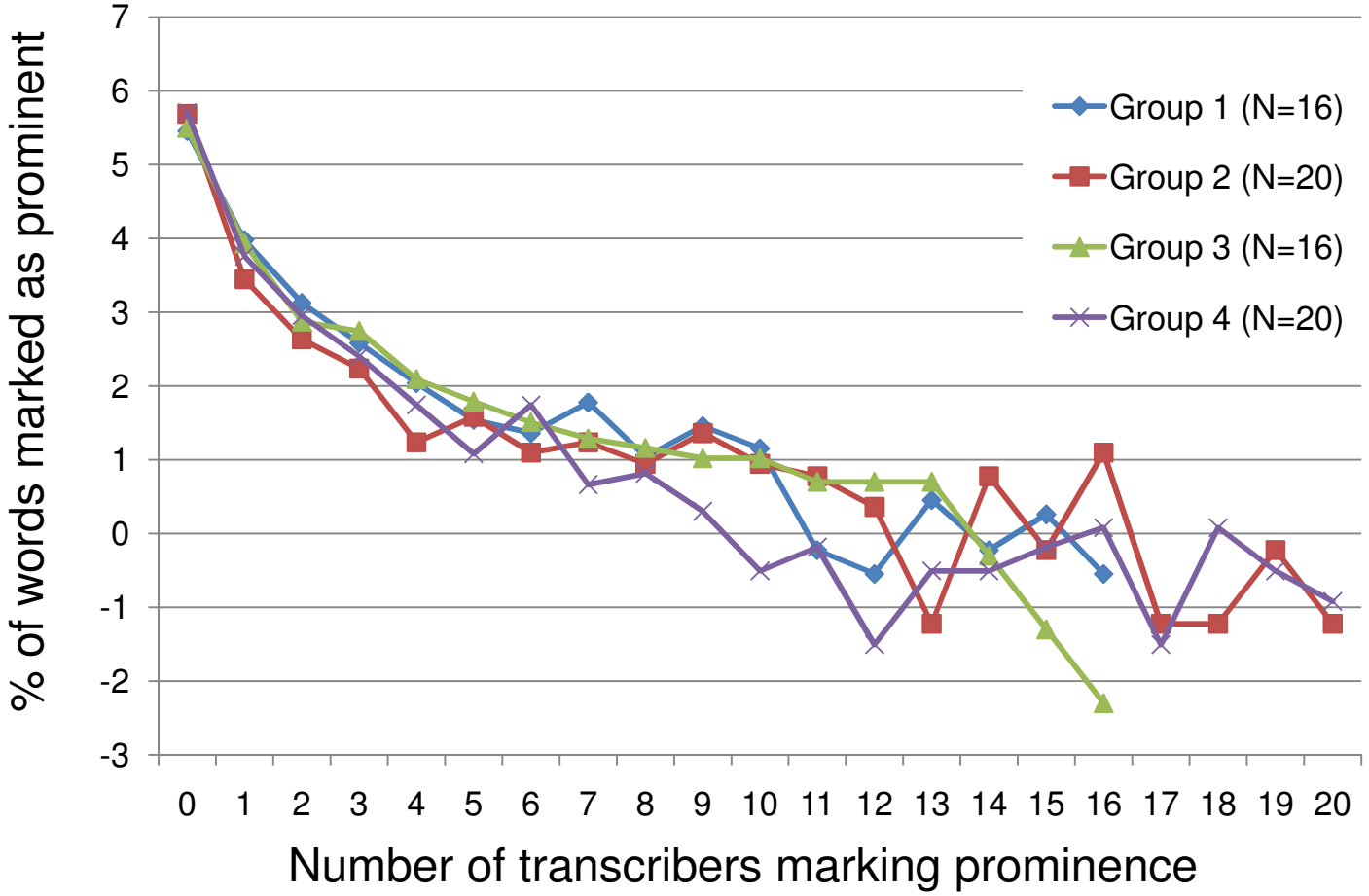
Results by listener: Prominence

Agreement patterns by word



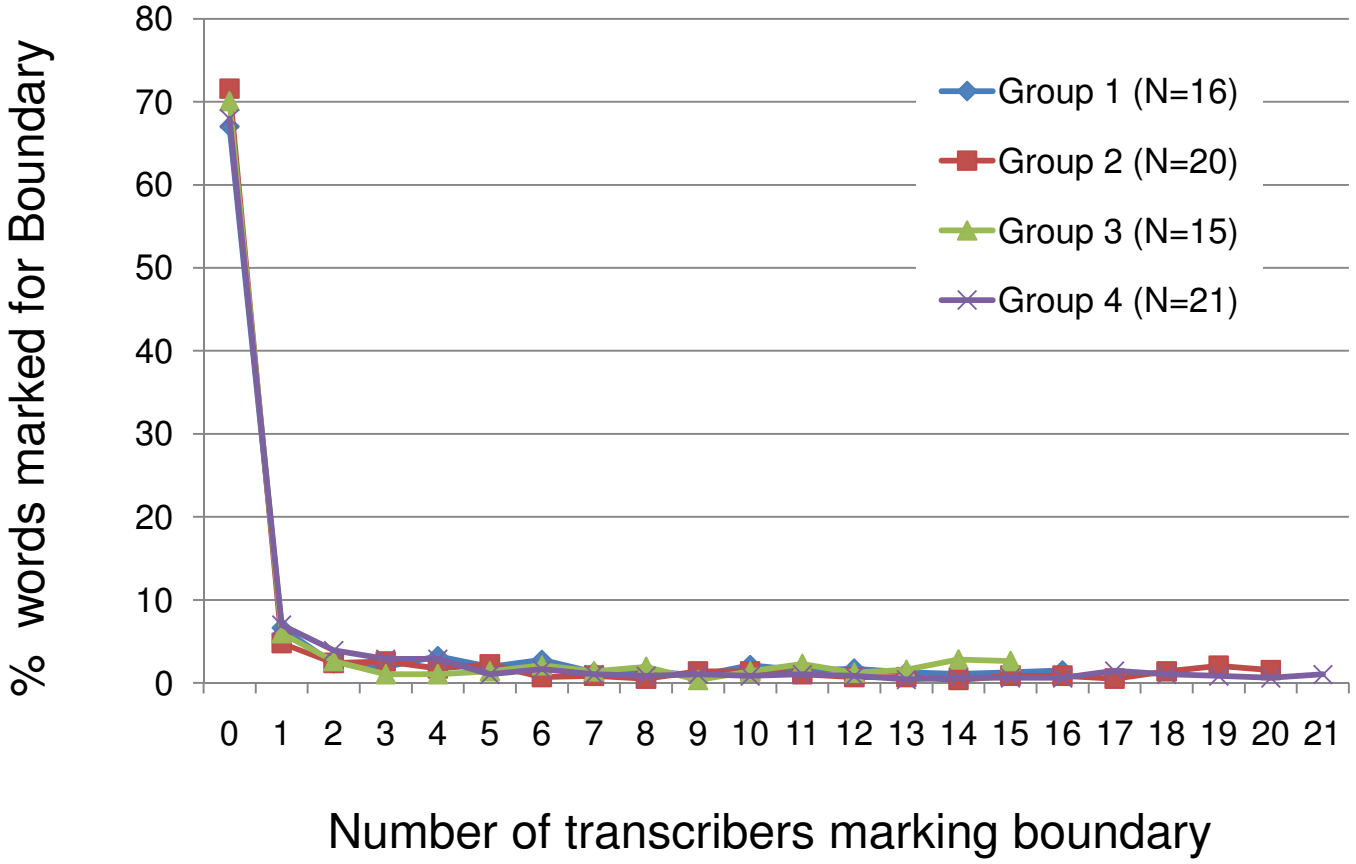
Results by listener: Prominence

Log_2 of Agreement patterns by word



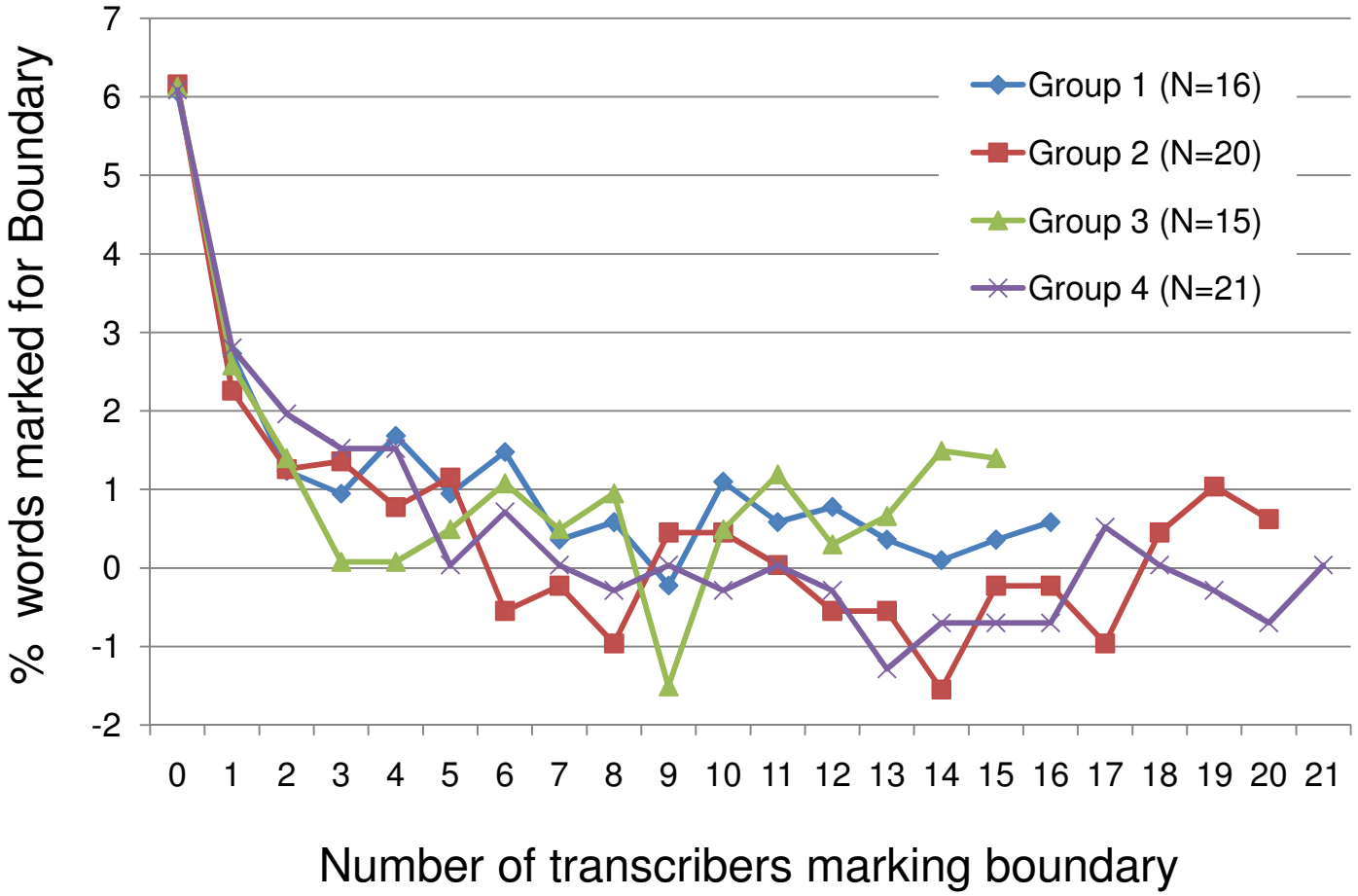
Results by listener: Boundary

Agreement patterns by word



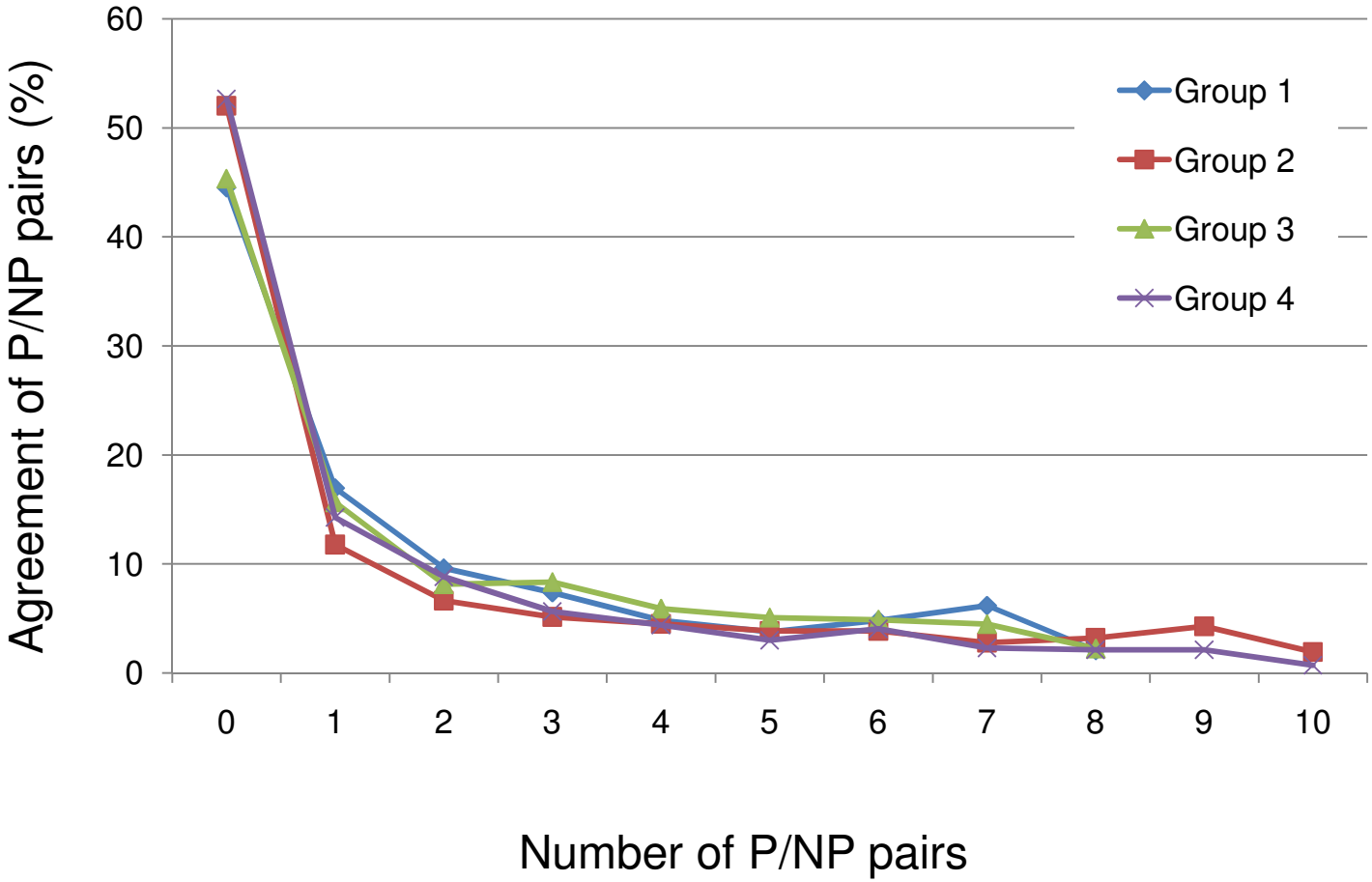
Results by listener: Boundary

Log_2 of Agreement patterns by word



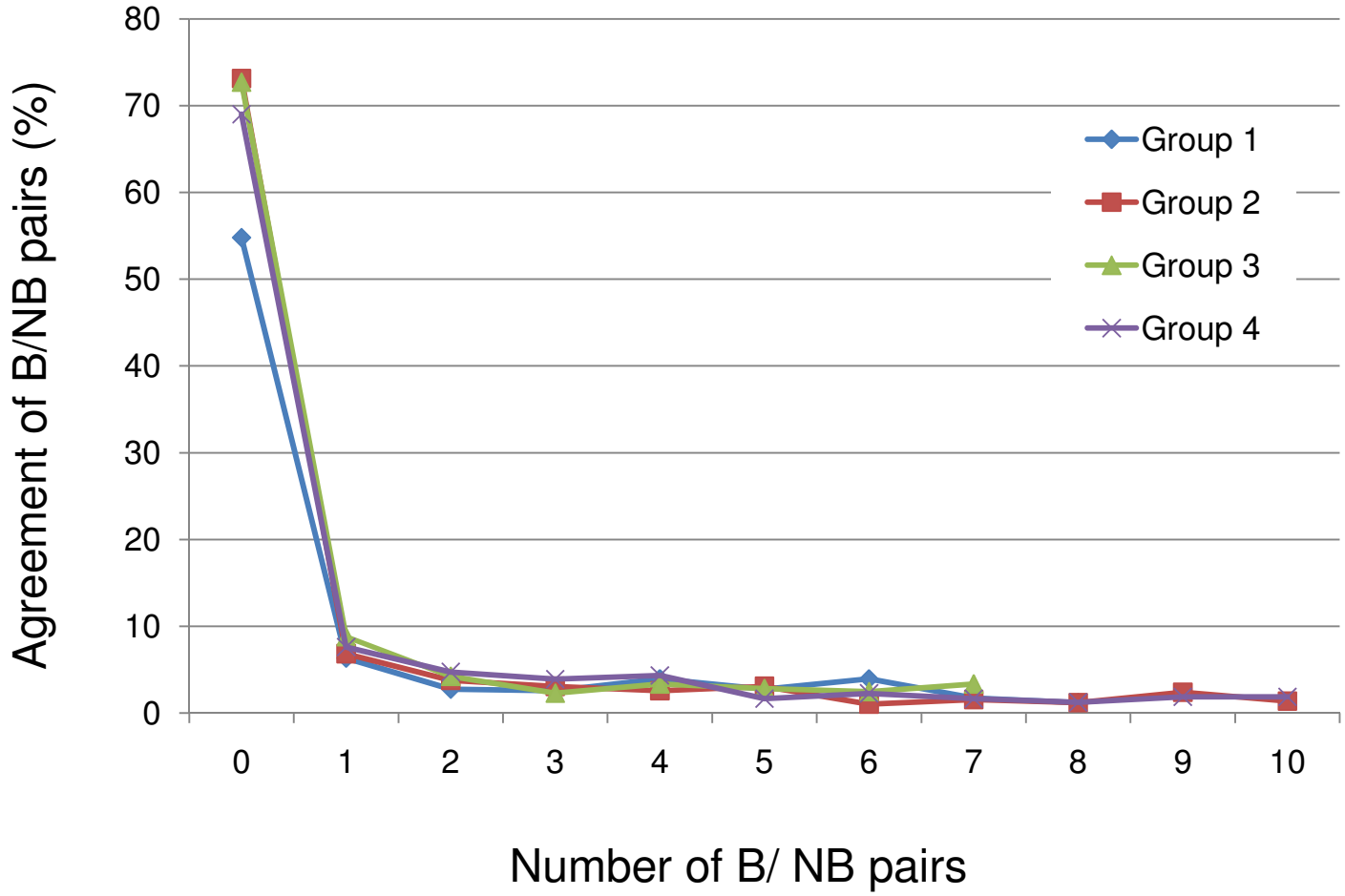
Results by listener: Prominence

Pairs of prominence/ non-prominence



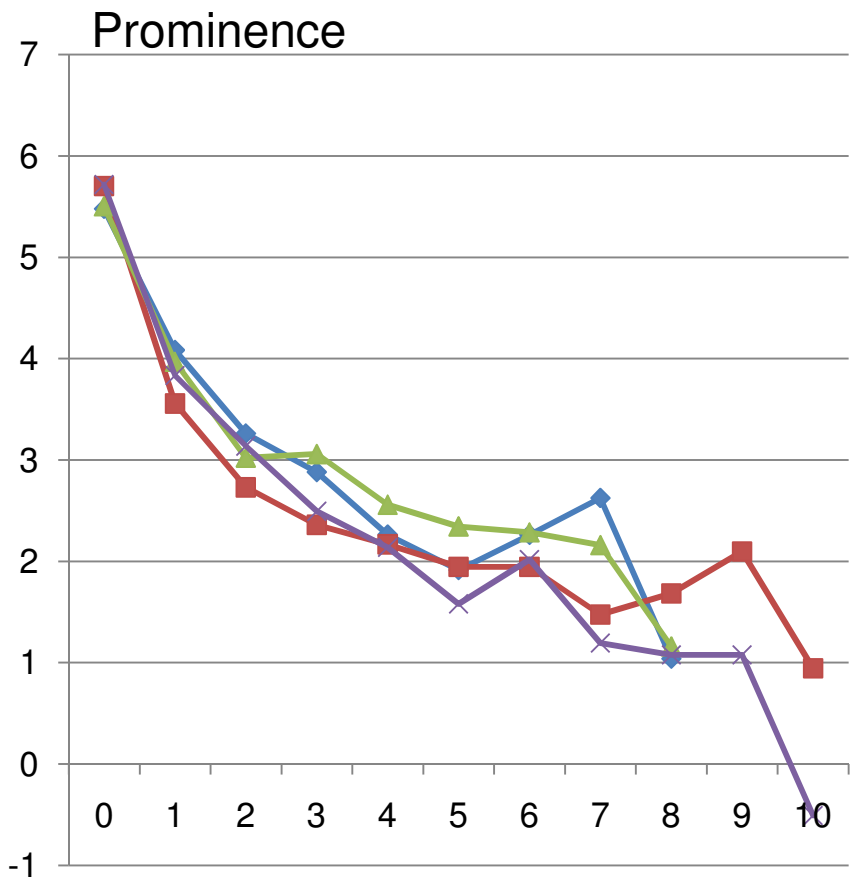
Results by listener: Boundary

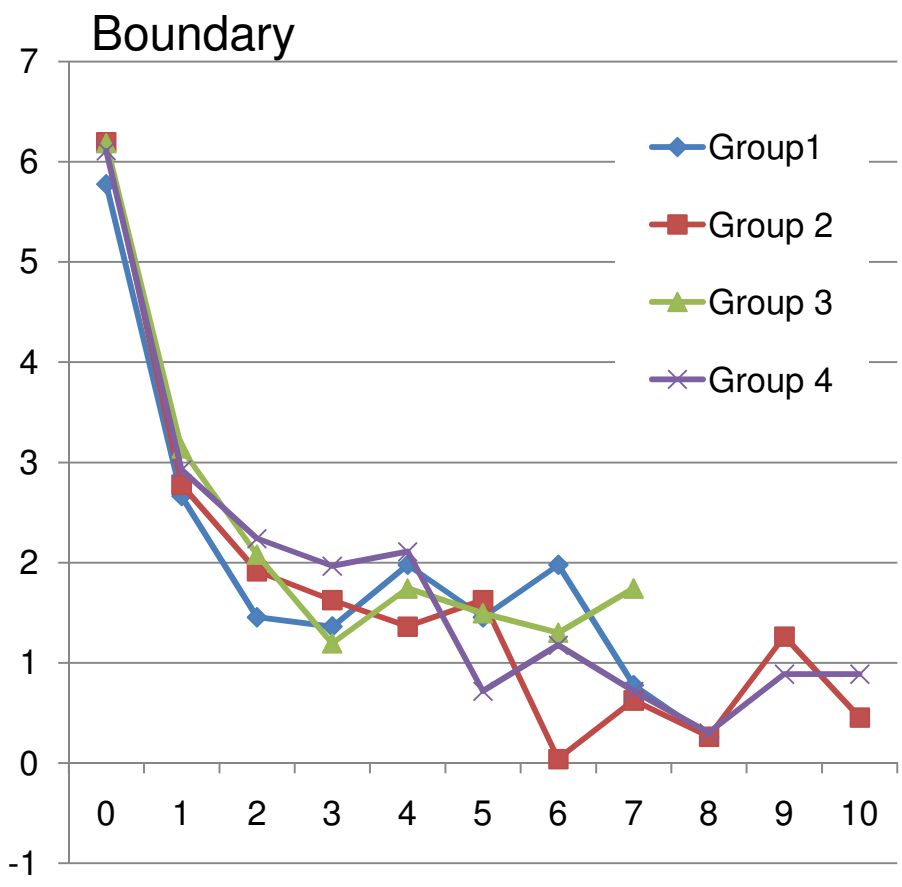
Pairs of boundary/ non-boundary



Results by listener

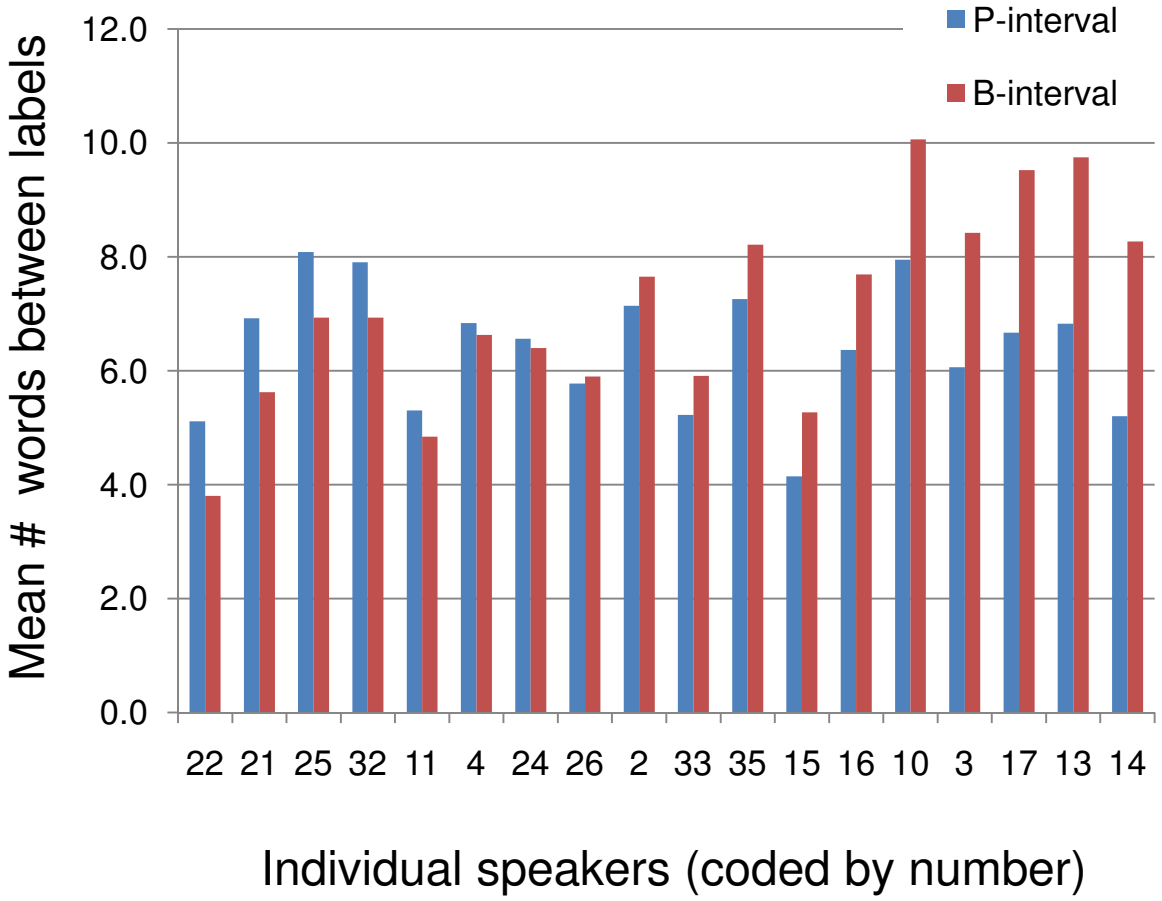
Log_2 (agreement of P/NP and B/NP pairs)





Results by speaker (average over all transcribers)

- Intervals between prominences and boundaries by speakers



- Avg. intervals b/w P: 6.4 w

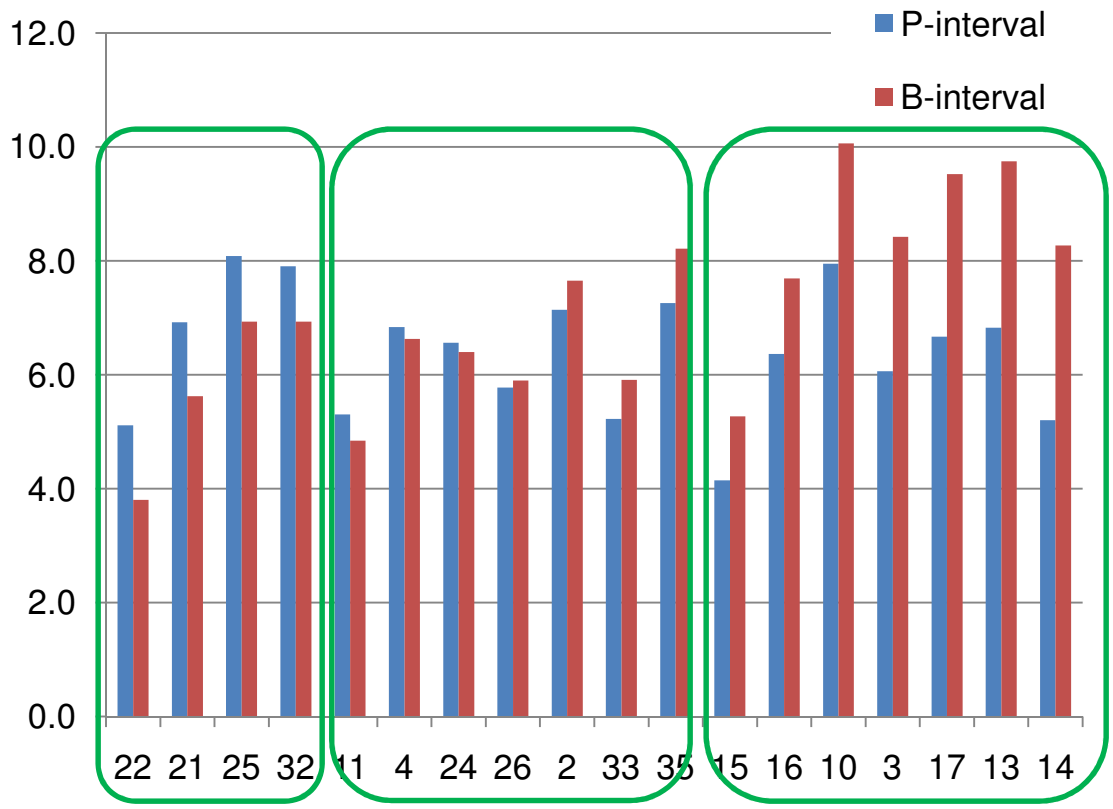
  Range: 4.1 – 8.1 w

- Avg. intervals b/w B: 7.1 w

  Range: 3.8 - 10.1 w

Results by speaker (average over all transcribers)

- Intervals between prominences and boundaries by speakers



- Avg. intervals b/w P: 6.4 w

  Range: 4.1 – 8.1 w

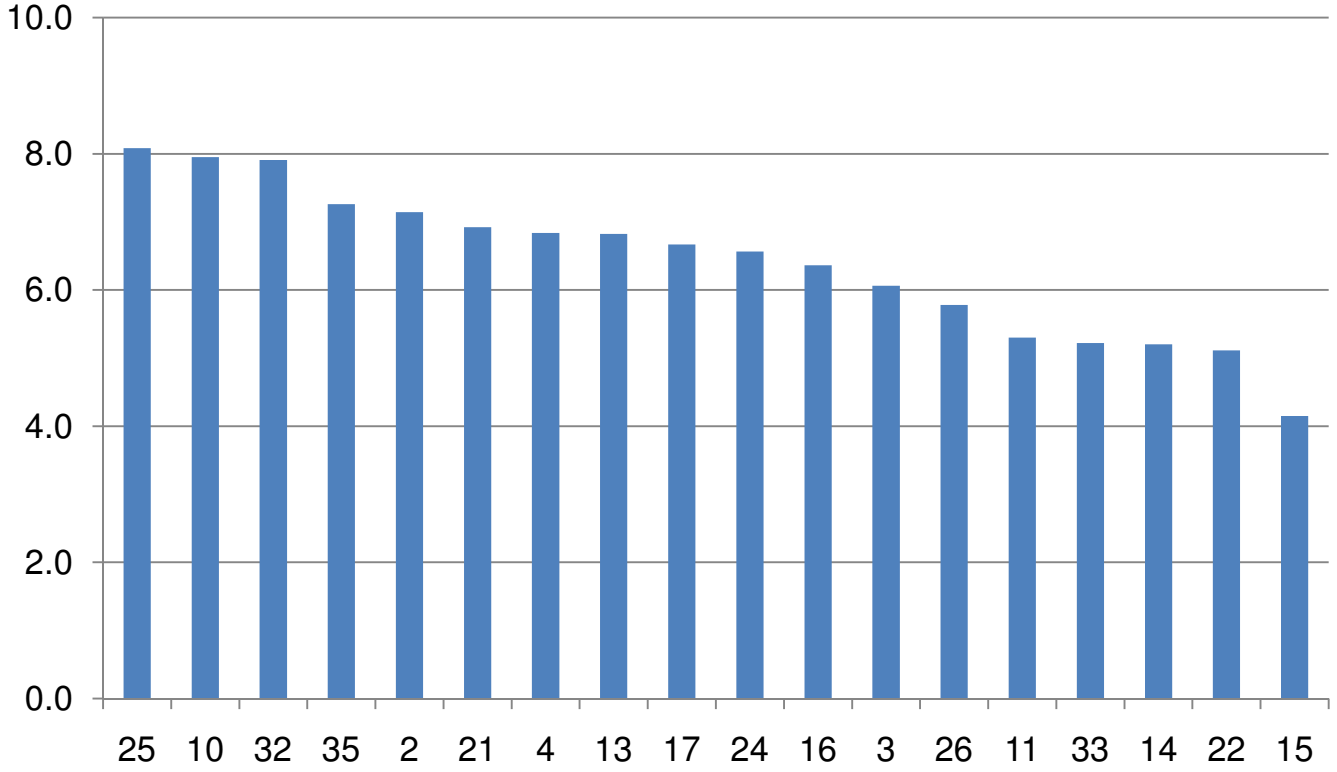
- Avg. intervals b/w B: 7.1 w

  Range: 3.8 - 10.1 w

- Int(P) > Int(B)
- Int (P) ≈ Int(B)
- Int (P) < Int(B)

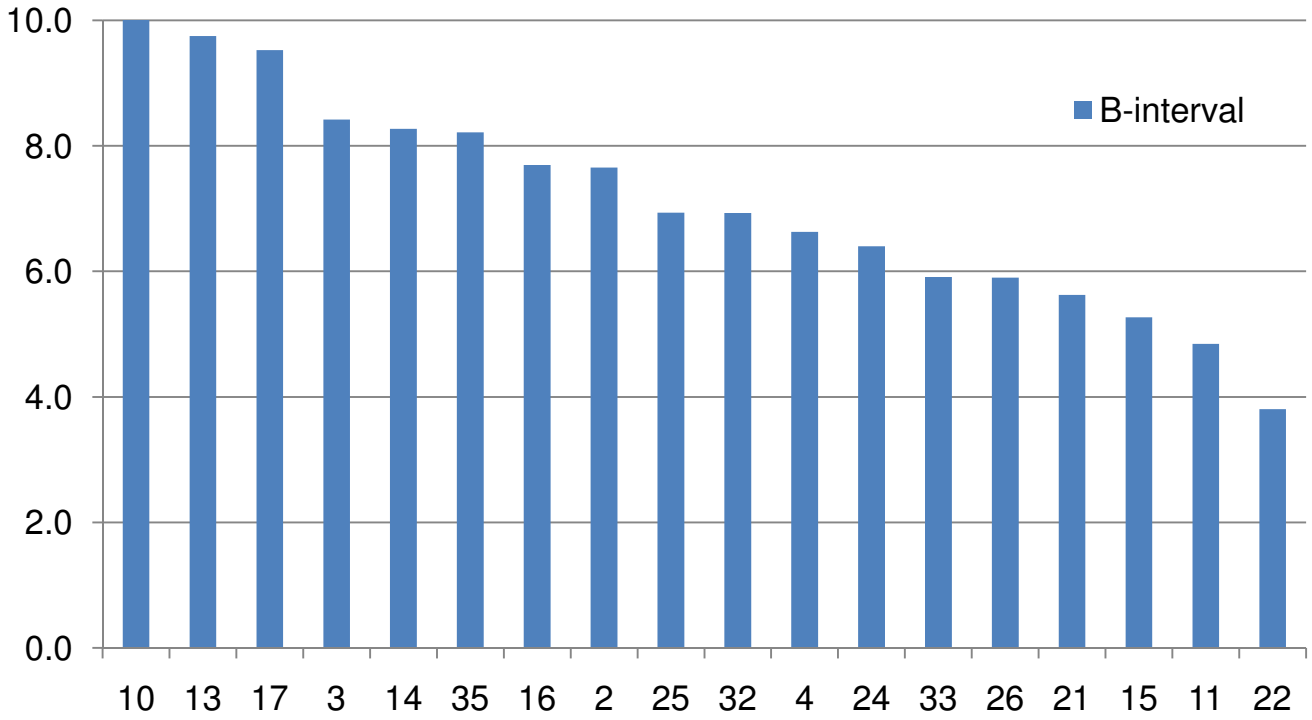
Results by speaker

- Variation by speaker in the intervals between prominences; each bar represents average over 15-22 listeners



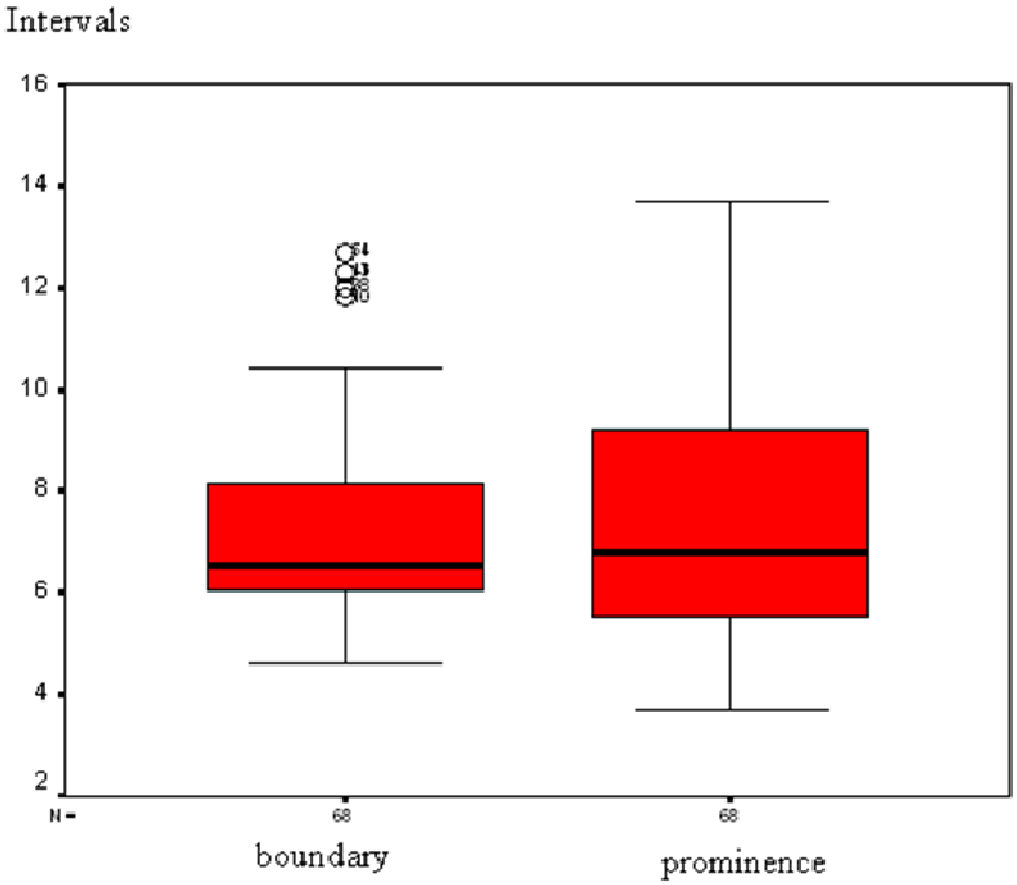
Results by speaker

- Intervals between boundaries by speakers



Results by listeners

- Intervals between prominences and boundaries by listeners (N = 72)



- Avg. intervals b/w P: 7.2 w

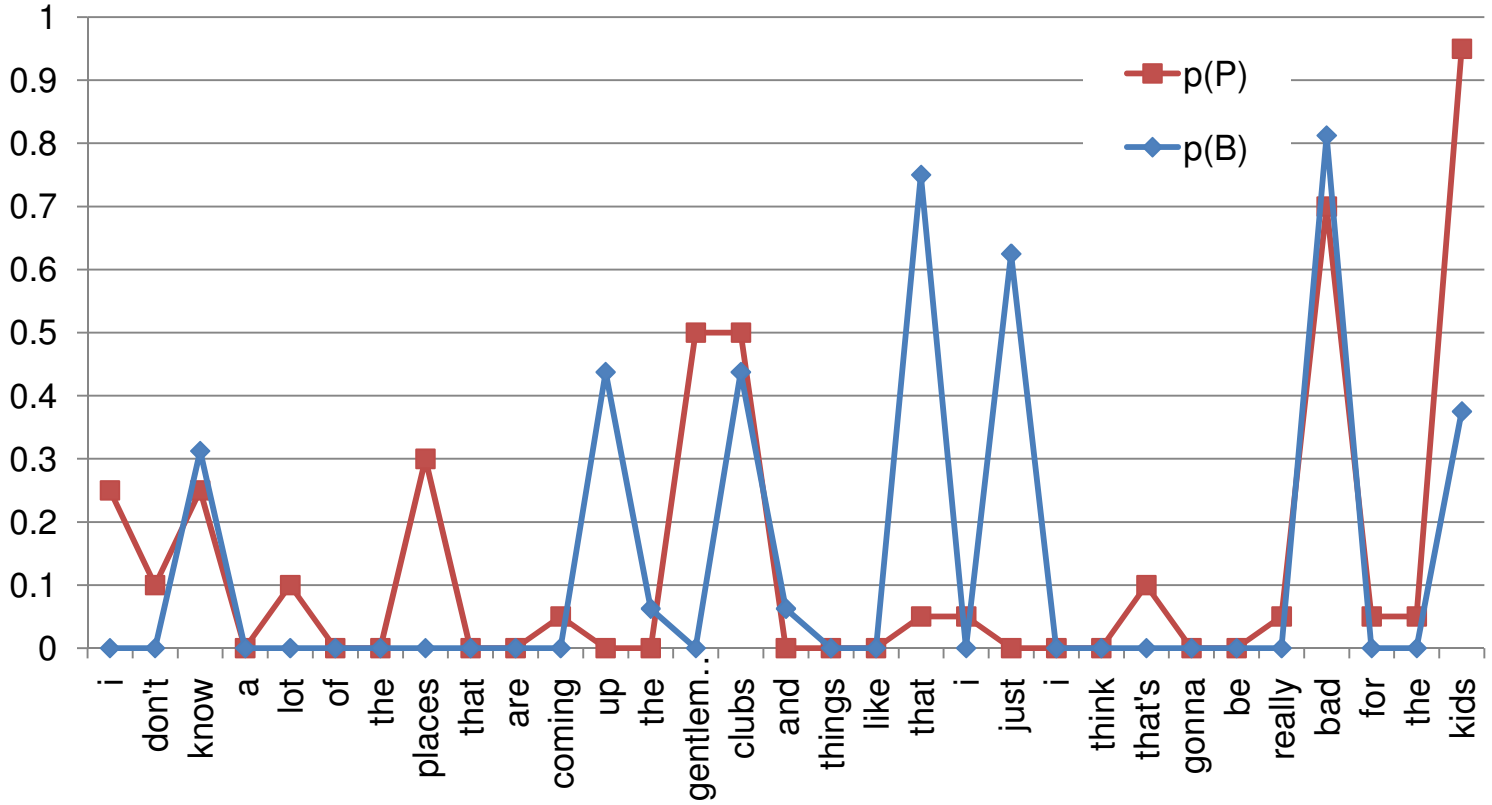
  Range: 3.8 – 18.7 w

- Avg. intervals b/w B: 7.3 w

  Range: 4.6 - 12.7 w

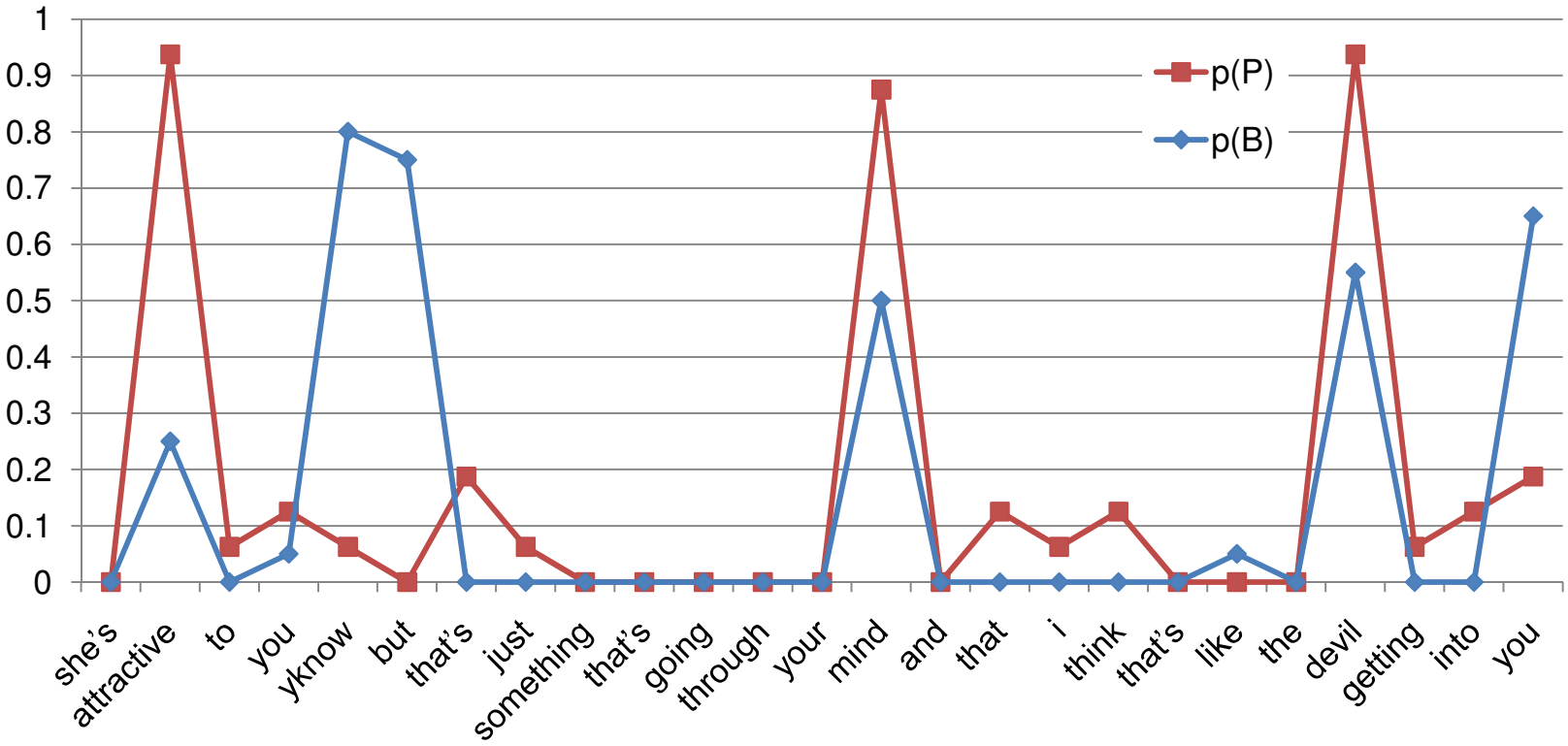
Probabalistic prosody labels

- Distribution of prominence and boundary (s23)



Probabalistic prosody labels

- Distribution of prominence and boundary (s03)



Assessing agreement

- Fleiss' multi-rater kappa coefficients

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

.

  - P(A) = proportions of times that raters actually agree
  - P(E) = proportions of times that raters would agree by chance

Assessing agreement

- Fleiss' multi-rater kappa coefficients and Z- statistics

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

$$P(A) = \frac{\sum_{i=1}^{n}\sum_{j=1}^{2} n_{ij} C_2}{N \times_T C_2} = \frac{1}{N}\sum_{i}^{N} S_i, \quad P(E) = \sum_{j=1}^{2}\left(\frac{\sum_{i}^{N} A_{ij}}{N * T}\right)^2 = \sum_{j=1}^{2} p_j^2$$

	P	No P	S_i
W1	5	5	(5*4+5*4)/10*9
W2	6	4	(6*4+4*3)/10*9
W3	0	10	10*9/10*9
W4	5	5	(5*4+5*4)/10*9
A_j	16	24	
P_j	16/(4*10)	24/(4*10)	

N= 4, T = 10

  - P(A) = proportions of times that raters actually agree
  - P(E) = proportions of times that raters would agree by chance

Assessing agreement

- Fleiss' multi-rater kappa coefficients and Z- statistics

z=2.32, α=0.01		Exp.1		Exp. 2	
		Grp.1	Grp.2	Grp.3	Grp.4
Prominence	Kappa	0.373	0.421	0.394	0.407
	z	19.43	20.48	18.15	18.31
boundary	Kappa	0.612	0.544	0.621	0.575
	z	27.62	21.87	25.05	26.22

  - All agreement scores are statistically significant.
  - Agreement scores for boundary are consistently higher than those for prominence.

Assessing agreement

- Fleiss' multi-rater kappa coefficients and Z- statistics

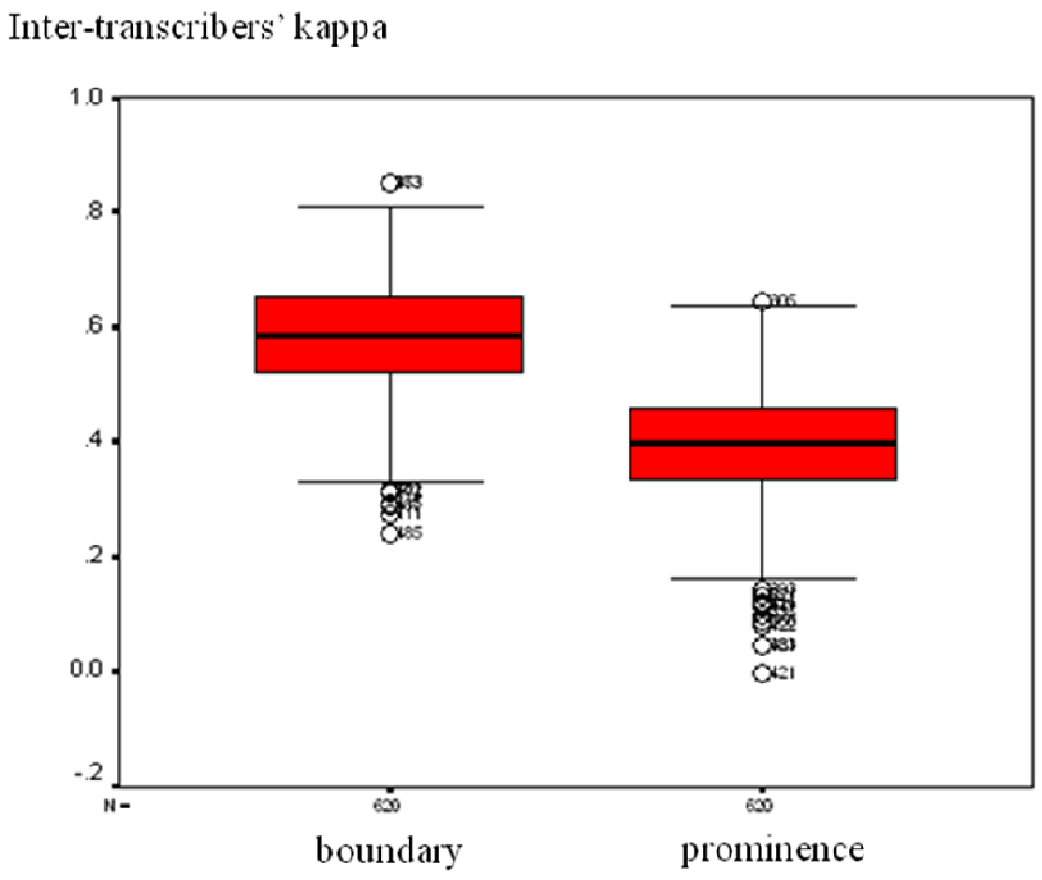
z=2.32, α=0.01		Exp.1		Exp. 2	
		Grp.1	Grp.2	Grp.3	Grp.4
Prominence	Kappa	0.373	0.421	0.394	0.407
	z	**19.43**	**20.48**	**18.15**	**18.31**
boundary	Kappa	0.612	0.544	0.621	0.575
	z	**27.62**	**21.87**	**25.05**	**26.22**

  - All agreement scores are statistically significant.
  - Agreement scores for boundary are consistently higher than those for prominence.

Assessing agreement

- Cohen's inter-transcriber kappa coefficients

  (Both members of pair hear same speakers)



- Avg. kappa for B: 0.582

  Range: 0.240 – 0.850

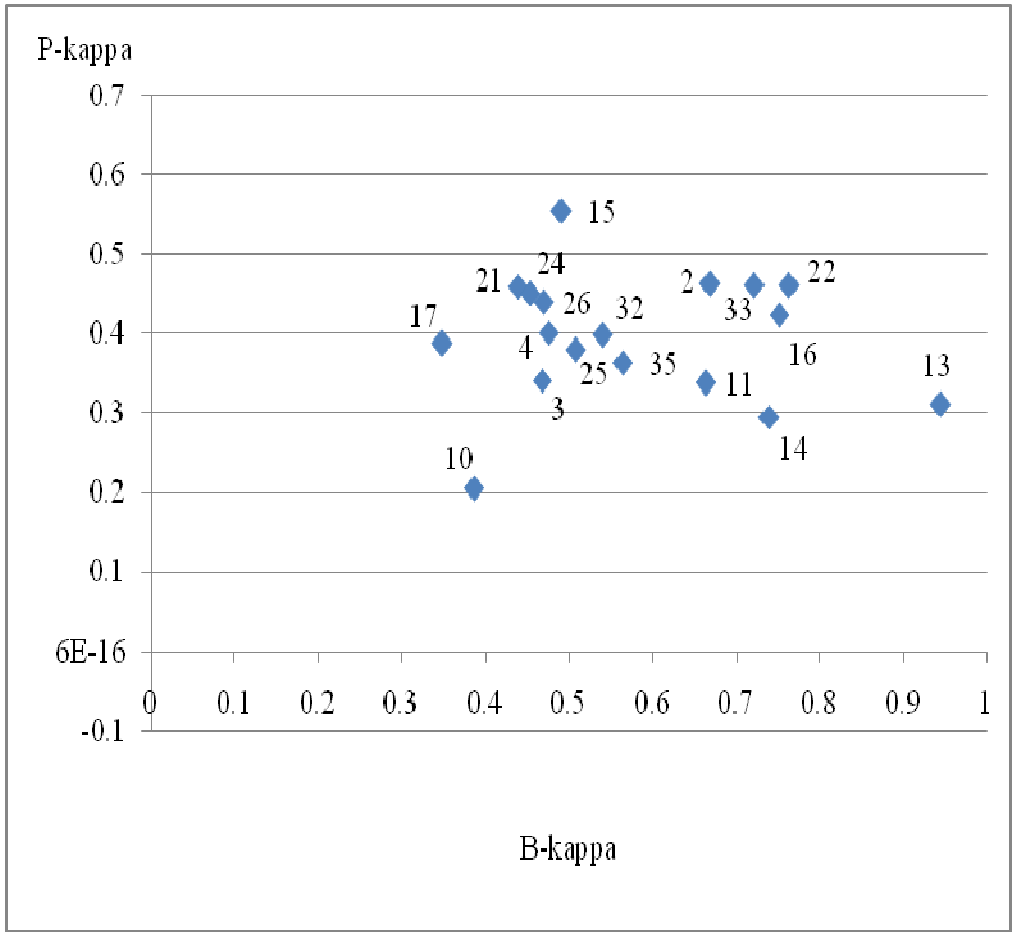
- Avg. kappa for P: 0.392

  Range: -0.003 – 0.644

- Listener induced variability

Plotting agreement scores by speaker:
Prominence x Boundary

- Fleiss' multi-rater kappa coefficients by speaker (set1)



- All agreement scores are statistically significant.
  Range of z: 3.25 – 13.77

- Speaker induced variability

Discussion

- Significantly high agreement scores show some uniformity in prosody perception across listeners.
- Greater uniformity in boundary perception
  - agreement scores: B > P
- Boundary perception is less variable across listener pairs
  - z scores (Fleiss' Kappa): B < P

Discussion

- Observe variability in agreement scores across transcriber pairs
  - Variable listener sensitivity to prosody indicators
- Observe variability in agreement scores across speakers
  - Speakers vary in how clearly they cue prosody
  - Within-speaker variation for prominence vs. boundary agreement
- Observe variability in intervals between prominences and boundaries
  - Speakers vary in frequency of prominence or boundary marking, or maybe in clarity of cues (e.g., in nuclear vs. pre-nuclear prominences or in boundary strength)

THANKS!

This research is funded by NSF IIS-0414117

Thanks to our collaborators in the Prosody & ASR Group:

Mark Hasegawa-Johnson
Chilin Shih
Margaret Fleck
Xiaodan Zhuang
Zak Hulstrom
Tae-Jin Yoon