From eye to mouth:
Connecting non-linguistic visual grouping and linguistic prosody

Elsi Kaiser, Edward Holsinger, David Cheng-Huan Li and Dani Byrd

Department of Linguistics, University of Southern California, Los Angeles, USA {emkaiser, holsinge, lidc, dbyrd} @usc.edu

Introduction

- ***Grouping*** matters in both language and vision.
  - *Vision*: Grouping parts of a visual stimulus together is crucial for perception (e.g.[3]).
  - *Language*: Words are organized into phrasal units, separated by prosodic boundaries/breaks.
    - Boundary strength is indexed by many acoustic correlates—e.g., segmental lengthening and/or pausing—and influenced by factors like constituent structure [(4]).
- **Different domains:** Encoding of prosodic grouping is inherently temporal (speech unfolds in time), whereas visual grouping is based on distance/proximity, color, etc.
  - Do these domains—in particular, the spoken/temporal and the visual/spatial—connect?
- We explore two possibilities:
  - **Distance Hypothesis:** The greater the distance between objects, the stronger the prosodic boundary between phrases denoting those objects.
  - **Grouping Hypothesis:** Boundary strength is sensitive to a more abstract level: whether objects belong to a spatially-defined group.
- **Does visuo-spatial grouping influences prosodic grouping in the linguistic domain? If a speaker describes a multi-object display, does the spatial configuration/layout influence the strength of prosodic breaks between nouns?**

Perception data

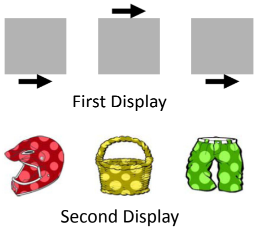
- Analysis = Used listeners' perception of 'connectedness' to estimate ***boundary strength***
  - There are *multiple* cues to prosodic boundary strength
    - Using humans as our measurement tool allows us to tap into multiple potential cues of boundary strength
  - Existing work has shown that listeners can provide 'connectedness' ratings that relate meaningfully to boundary strength (Krivokapic, 2007)

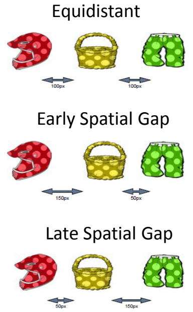
Eye-movement data

- Close connection between **eye movements and speech.**
  - When naming objects or describing scenes, people tend to start to look at the object about 800-1000ms before naming it / before word onset (e.g., Meyer et al., 1998; Griffin & Bock, 2000).
- Analysis = Used speakers' eye-movements to investigate how the sensitivity to visual cues expresses itself in the attentional shifts that take place during production.

Production Study: Design and methods

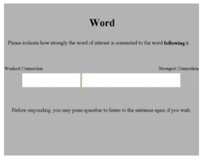
- Participants (n=7) produced scripted utterances based on images on computer screen, eye-movements recorded while speaking.
- Task: Describe the path of an imaginary little brown mouse as he navigates over or under each object before going into a mouse hole
  - E.g. The little brown mouse runs *under the red helmet {break 1} over the yellow basket {break 2} under the green shorts* and into the mouse hole.
- We manipulated the visual scene layout by changing the distance between the three objects: (i) Equidistant/ungrouped (O O O), (ii) Early gap (O _ O O), (iii) Late gap (O O _ O)





Perception Data

- Listeners' perception of prosodic boundary strength
  - Listeners *did not have access* to information about the visual scene
- Participants (n=28) provided ratings of prosodic boundary strength based on what they heard



Word

Please indicate how strongly the word of interest is connected to the word **following** it.

Weakest Connection

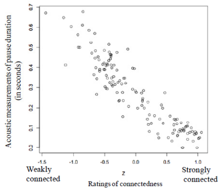
Strongest Connection

Before responding, you may press spacebar to listen to the sentence again if you wish.

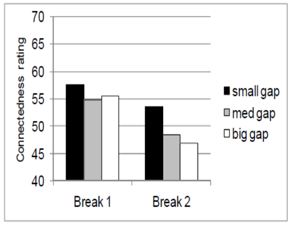
- Task: To rate *how strongly connected* the word of interest is to the word following it, using slider

Results

- Checking task validity: Connectedness ratings are negatively correlated with pause duration
  - Strongly connected = short pause
  - weakly connected = long pause
  - Ratings provide meaningful information about prosodic boundaries



- Significant main **effect of grouping on connectedness rating** (p<.001)
  - Nouns that are *grouped together* are perceived as more connected (separated by weaker boundaries) than nouns that are *excluded from group or ungrouped.*

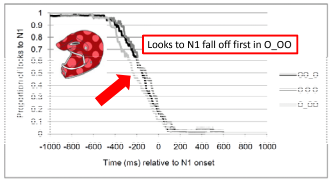


Eye-movement Data

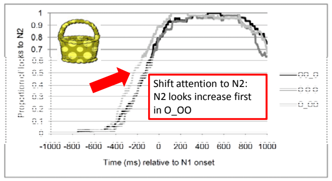
Overall, the data support the Grouping Hypothesis:

Before saying Noun1 (e.g. helmet), where look?

Looks to **Noun1** before uttering Noun1



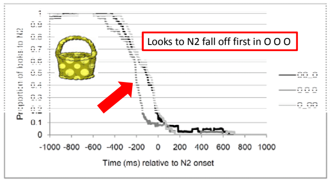
Looks to **Noun2** before uttering Noun1



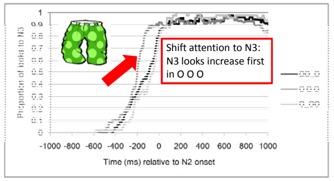
Effect of grouping: Shift from 1st object (N1) to 2nd object (N2) is earlier when 1st object is 'alone' (excluded from a group, O_OO) than in other configurations.
=> Move on rapidly from ungrouped objects

Before saying Noun2 (e.g. basket), where look?

Looks to **Noun2** before uttering Noun2



Looks to **Noun3** before uttering Noun2



Effect of grouping: Shift away from 2nd object (N2) to 3rd object (N3) is earlier when 2nd object is ungrouped (O O O) than when it is in a group (O_OO, OO_O).
=> Linger on grouped objects

Conclusions

- Visual grouping influences temporal aspects of production, namely prosodic boundaries and eye-movement patters.
  - Eye-movements exhibit sensitivity to visual grouping information in ways that relate to the prosodic groupings that speakers produce:.
  - In both cases, it is the higher-level property of grouping that matters, rather than straightforward physical distance.
- Our results suggests that the level at which linguistic and visual representations interface with each other is *abstract*
  - reflects cognitive structuring, not the detailed physical dimensions of either speech or visual information.
- Prosodic grouping effects are temporal, image manipulation was visuo-spatial: Domain-general consequences of the abstract notion of grouping.

This research was supported by NIH grant DC03172 awarded to Dani Byrd. We gratefully acknowledge useful feedback and comments from Louis Goldstein.