The 27th CUNY Conference on Human Sentence Processing, 13-15 March 2014, OSU

Influence of Visual Complexity on Referring Expression Generation

Hannah Rohde, Alasdair Clarke, & Micha Elsner

The Ohio State University

1. Abstract

In describing a target among distractors, speakers must extract relevant properties of a visual scene and formulate a coherent referring expression (RE) to pick out that target. Although linguistic cues influence how we see the world (Spivey et al., 2001) and properties of our visual system influence what we choose to say (Coco & Keller, 2012), models of referring expression generation (REG) have largely sidestepped a role for vision (Clarke et al., 2013). We ask how the assumptions of a well-known model, the Incremental Algorithm, hold up in visually complex scenes. Our findings suggest that, just as the Incremental Algorithm posits incrementality for feature inclusion in the RE, feature extraction/checking is likewise an active and ongoing process. To scale well, REG models should incorporate cues like scene complexity and be informed by findings from visual perception.

2. Modeling Referring Expression Generation

Incremental Algorithm (Pechmann, 1989; Dale & Reiter, 1995)
Goal: account for target overspecification in REG

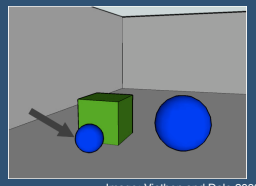


Image: Viethen and Dale 2008

Speaker says:
"the small blue ball"

even though disambiguated by:
"the small ball"

IA procedure:
- given a set of target features (e.g., color, size, shape)
- for each feature
  - for each distractor
    - add feature to RE if distractor is excluded
  - terminate when RE is unambiguous
- speech may start before termination

3. A Role for Visual Perception?

Naïve assumptions:

(i) "the set of candidate features can be identified easily"
➔ Do certain visual contexts impede feature idenfitication?

(ii) "distractor checking is always a serial process"
➔ Can speakers pre-attentively perceive feature effectiveness?

Possible impact of visual context:

Heterogeneous distractors ➔ Feature identification is hard?
Many similar distractors ➔ Single feature (e.g. shape) is enough?

4. Hypothesis from visual search

Visual search research shows: Finding a target is slower with more distractors, but only with heterogeneous distractors.

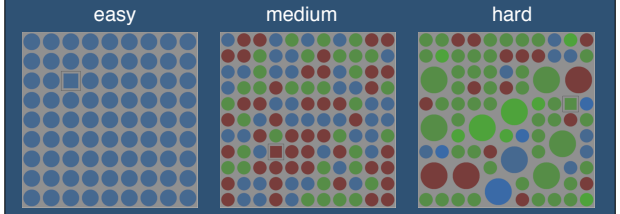
Question: Does this extend from perception to production?

Perception-Informed Hypothesis: If visual search findings extend to REG feature extraction/checking…
Then speaker production will be **slowed** by large distractor N, via an **interaction** with Heterogeneity.

Incremental Algorithm Hypothesis: IA has no prediction for an N x Distractor Heterogeneity interaction because the role for vision is left unspecified.

5. Stimuli

- 60 grids of colored shapes that varied…
  - **set size N** (25, 49, 81, 121)
  - **heterogeneity (=difficulty of visual search)**



➔ Target always disambiguated by shape alone ("square")

Fillers: 60 grids with non-unique targets which required relative descriptions ("leftmost red circle")

6. Production Experiment

Instructions: "describe the target so that a listener could quickly and accurately find that shape in the same grid"

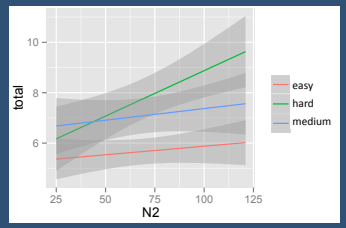
Participants: 18 English speakers from University of Edinburgh

Coding:
- speech onset
- speech offset
- inclusion of relative descriptions

Also have eyetracking; transcription and analysis ongoing!

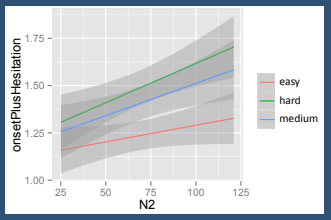
7. Results

Total speaking time:
Effects of N ($p<.001$), heterogeneity ($p<.05$), driven by predicted interaction ($p<.001$).



➔ Confirmed hypothesis from visual search:
- Little effect of N in easy, medium (+5ms/obj)
- Large effect of N with most heterogeneity (+32ms/obj)

Onset to speech time:
Only observed effect is small effect of N on onsets ($p<.001$; +2ms/object).



➔ Interaction only in post-onset speech suggests full feature extraction is non-trivial and is completed after speech onset

Source of post-onset delays:
- Speakers express same information more slowly (disfluency)
- Provide more information (taking longer to communicate)

…However, overspecification does not vary by condition

➔ No difference in use of relative descriptions for N or Heterogeneity
➔ Variation likely reflects patterns of disfluency

8. Conclusion

- Just as the Incremental Algorithm posits incrementality for feature inclusion in the referring expression, our results suggest that feature extraction/checking is likewise an active and ongoing process.

- To scale well, REG models should incorporate cues like scene complexity and be informed by findings from visual perception.

References

Clarke, A., Elsner, M. & Rohde, H. (2013). Where's Wally: the influence of visual salience on referring expression generation. *Frontiers in Perception Science: Special Issue on Scene Understanding*, 4(329).
Coco, M. I., & Keller, F. (2012). Scan pattern predicts sentence production in the cross-modal processing of visual scenes. *Cognitive Science*, 36.
Dale, R., & Reiter, E. (1995). Computational interpretations of the Gricean maxims in the generation of referring expressions. *Cognitive Science*, 19.
Pechmann, T. (2009). Incremental speech production and referential overspecification. *Linguistics*, 27(1).
Spivey, M., Tyler, M., Eberhard, K & Tanenhaus, M. Linguistically mediated visual search. *Psychological Science*, 12(4).
Viethen, J., & Dale, R. (2008). The use of spatial relations in referring expressions. *Proceedings of INLG*, 5.

Contact: Hannah.Rohde@ed.ac.uk