Influence of Visual Complexity on referring Expression Generation

Citation for published version:

Link:
Link to publication record in Edinburgh Research Explorer

Document Version:
Publisher's PDF, also known as Version of record

Publisher Rights Statement:

General rights
Copyright for the publications made accessible via the Edinburgh Research Explorer is retained by the author(s) and / or other copyright owners and it is a condition of accessing these publications that users recognise and abide by the legal requirements associated with these rights.

Take down policy
The University of Edinburgh has made every reasonable effort to ensure that Edinburgh Research Explorer content complies with UK legislation. If you believe that the public display of this file breaches copyright please contact openaccess@ed.ac.uk providing details, and we will remove access to the work immediately and investigate your claim.
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

**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?

**Naive assumptions:**
(i) "the set of candidate features can be identified easily"
- Do certain visual contexts impede feature identification?
(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)
  - easy, medium, hard

**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).

**Onset to speech time:**
- Little effect of N in easy, medium (+5ms/obj)
- Large effect of N with most heterogeneity (+32ms/obj)

**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**


---

**Contact:** Hannah.Rohde@ed.ac.uk