Influence of Visual Complexity on Referring Expression Generation

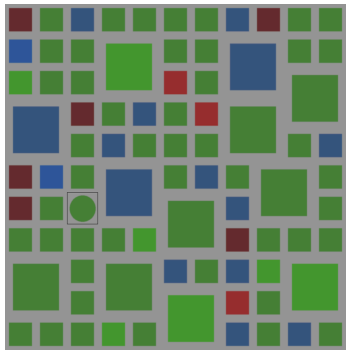
Hannah Rohde, Alasdair Clarke (University of Edinburgh), & Micha Elsner (OSU)
Hannah.Rohde@ed.ac.uk

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., *Psych Science*, 2001) and properties of our visual system influence what we choose to say (Coco & Keller, *CogSci*, 2012), models of RE generation (REG) have largely sidestepped a role for vision (Clarke et al., *Frontiers in Psych,* 2013). Such models include the Incremental Algorithm (Pechmann, *Linguistics,* 1989; Dale & Reiter, *CogSci,* 1995). This algorithm successfully accounts for target overspecification in REG—i.e., speakers' inclusion of features of the target not required for minimal disambiguation. It does so by incrementally assessing target features (e.g., color, size, shape) and adding a feature to the RE if a serial check of distractors confirms that the feature excludes at least one distractor. The process terminates when the RE is unambiguous, though the speaker may start speaking earlier. However, the algorithm makes two large assumptions that are naive in light of findings from visual perception: (i) that the set of candidate features can be identified easily and (ii) that distractor checking is always a serial process.

Here we test how these assumptions hold up in visually complex scenes. A problem for (i) is that the set of potential features of a target is unbounded (beyond color/size/shape, why not consider the target's texture, orientation, lack of horns, etc.?), and more heterogeneous scenes may make the identification of relevant features harder. For (ii), the speaker can sometimes pre-attentively perceive a feature's effectiveness (e.g., shape excludes at least one distractor—in fact all—for a circle among homogenous squares). In visual search studies, the number of distractors slows viewers' response, but only if the distractors are heterogeneous. The Incremental Algorithm makes no prediction for such an interaction between set size and heterogeneity on REG, given that the role for vision is left unspecified. If findings from visual search extend to REG feature extraction/checking, then we predict this interaction in production.

Methods: 18 participants viewed 60 randomly generated stimuli consisting of a grid of colored shapes (see figure). The grid varied in set size (N = 25, 49, 81, 121) and heterogeneity (homogeneous, varying in color only, and varying in color/shape/intensity). Participants were instructed to describe the target so that a listener could quickly and accurately find that shape in the same grid. Because targets could be disambiguated by shape alone ("circle"), we also included 60 fillers with non-unique targets (which required a relative description: e.g., "leftmost red circle"). REs were coded for speech onset, offset, and the inclusion of relative descriptions.

Results: The crucial N x heterogeneity interaction emerged for post-onset speaking time: Little effect of N in homogeneous and color-only heterogeneous conditions (+5ms/object) but a large effect with more heterogeneity (+32ms/object). The lack of an interaction in onset times (and the tiny effect of N on onsets: +2ms/object) suggests that full feature extraction is non-trivial and is completed after speech onset. The post-onset pattern could arise if speakers provide more information (taking longer to communicate) and/or express the same information more slowly (more disfluency). Since a single word disambiguates all critical items, REs are overspecified if information is added. However, our index of overspecification (use of relative descriptions) does not vary across conditions. Full-scale disfluency annotation is under way.



Just as the Incremental Algorithm posits incrementality for feature inclusion in the RE, 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.