Visual complexity and referring expression generation

THE OHIO STATE UNIVERSITY

Micha Elsner
with
Alasdair Clarke, Hannah Rohde

Where’s the Scott Monument?

See the clock tower? It’s the pointy black spire just to the right.

Terminology: Relational descriptions

More than one object:

- The **target (Scott Monument)**
- One or more **landmarks (clock tower)**

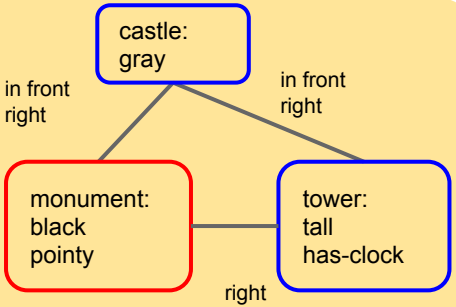
Common for complex scenes

(Viethen and Dale 2011)

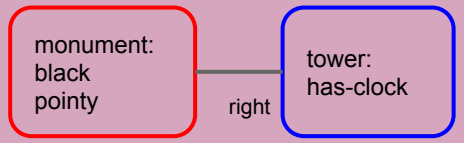
Language generation (pipeline?)

Goal: Identify the Scott Monument

Scene perception



Feature extraction



Content plan

Discourse / sentence plan

See the… uh... clock tower? It's the pointy black---black spire just to the right.

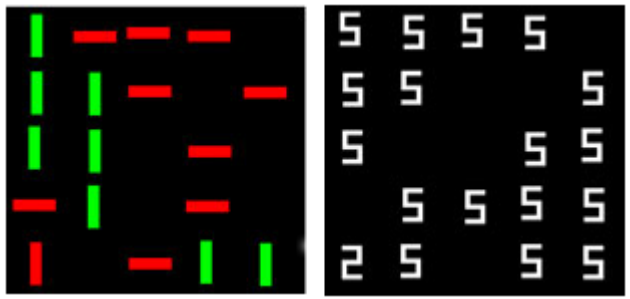
Realization

Incrementality

- Generation works piece-by-piece, and different levels interact…
  - “Incremental” models since (Pechmann 89), (Dale and Reiter 92)
- How does perception affect higher levels?
  - How pervasive are the effects?
  - How powerful?
  - Which perceptual factors?

Modeling visual perception

- Some visual searches are fast; some are slow (Wolfe ‘94 and subsq)
- Two mechanisms: “pop-out” and scanning
- Guided by bottom-up *salience* and top-down *relevance*
  - Salience: color/texture contrasts; relevance: task features
- Psychological models of perception
  - To predict eyetracking fixations and search difficulty



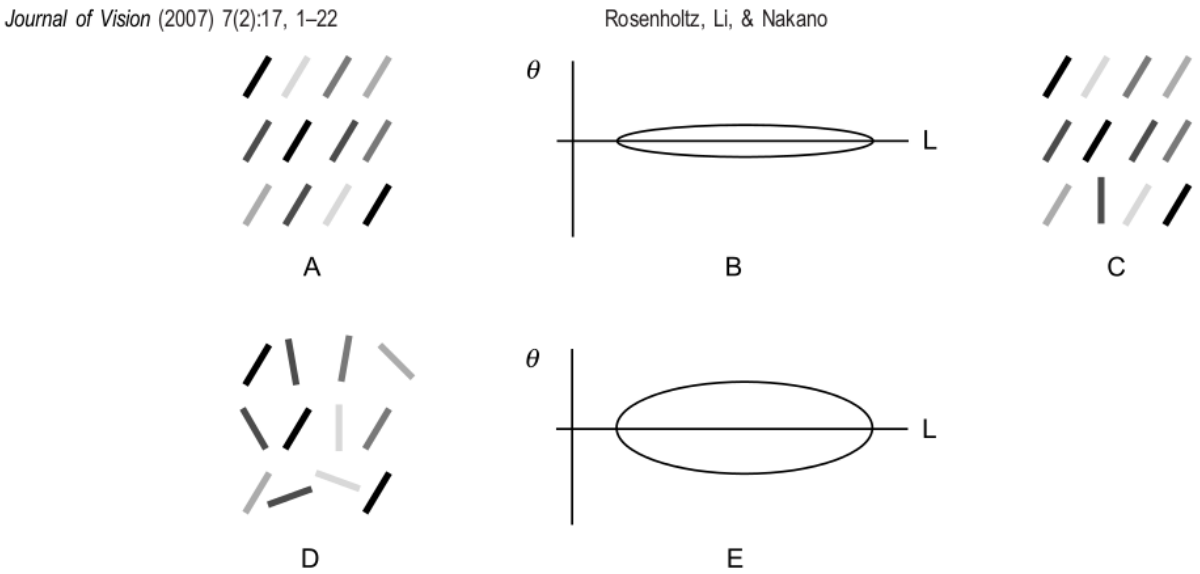
Wolfe and Horowitz 2004

Basic predictors

- Area of object
- Centrality on screen
  - Used extensively in previous work, eg (Kelleher 05)

Visual clutter

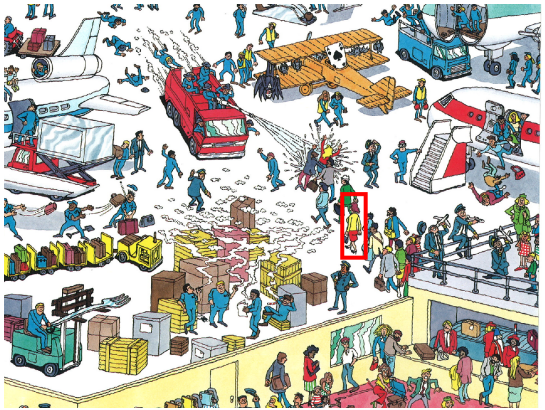
Diversity or variance of global scene statistics

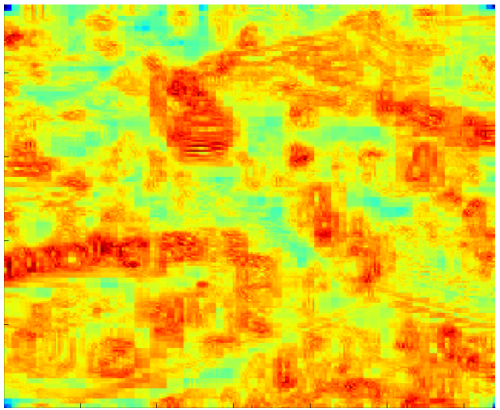


Low-level salience models

Similarity of point to overall scene

Bottom-up part of Torralba et al 2005





Better perceptual modeling?

Clarke, Dziemianko and Keller, VSS 2014 (poster)
Emma Ward MSc. thesis (in progress; adv Hannah Rohde)

Object-level visual salience

- Perceptual toolkit isn’t perfect…
  - Often weak effects
  - Or only area, not low-level salience etc.

- What’s missing?
- *Objects* vs pixels…
  - Pixel-vs-scene style models poor for objects
  - Large objects are salient but pixels within aren’t

Salience by feature?

Does distribution of feature values affect salience of the feature?

Would you describe these differently?

VS

Content selection

Clarke, Elsner and Rohde Front.Perception 2013

“Where’s Wally” corpus

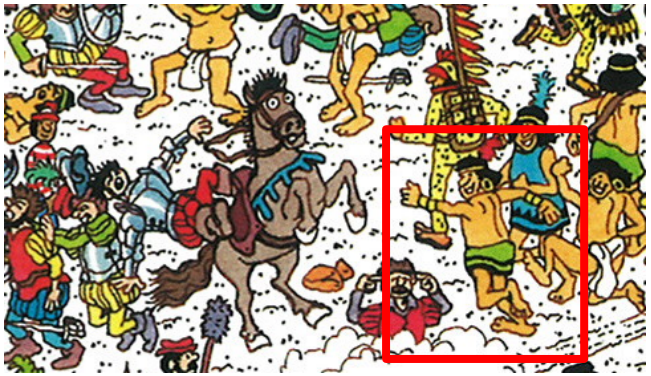
- “Where’s Wally” (Handford)...
  - A game based on visual search
  - Wide range of salient and non-salient objects
- Corpus collected on Mechanical Turk
  - Selected human targets in each image
  - Subject instructed to describe target so another person could find them
- Download: http://datashare.is.ed.ac.uk/handle/10283/336

Sample descriptions...

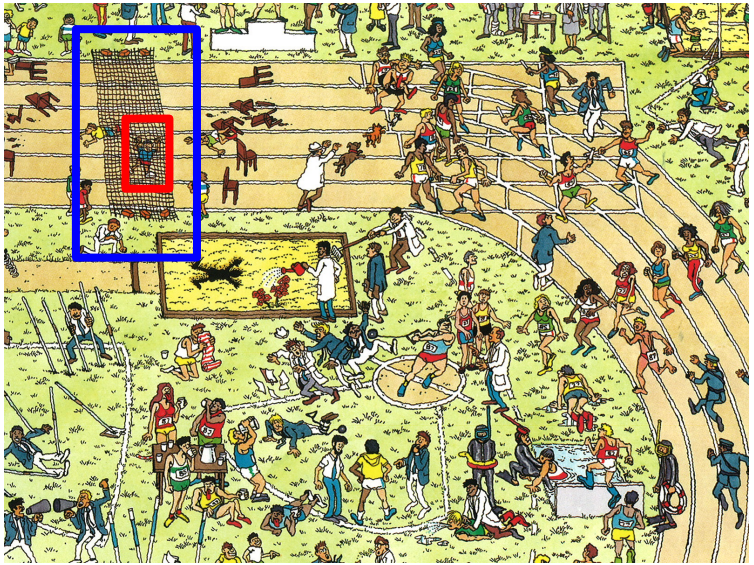
Man running in green skirt at the bottom right side of picture across from horse on his hind legs.

On the bottom right of the picture, there is a man with a green covering running towards the horse that is bucking. His arms are outstretched.

Look for the warrior in green shorts with a black stripe in the lower right corner. He’s facing to the left and has his arms spread.



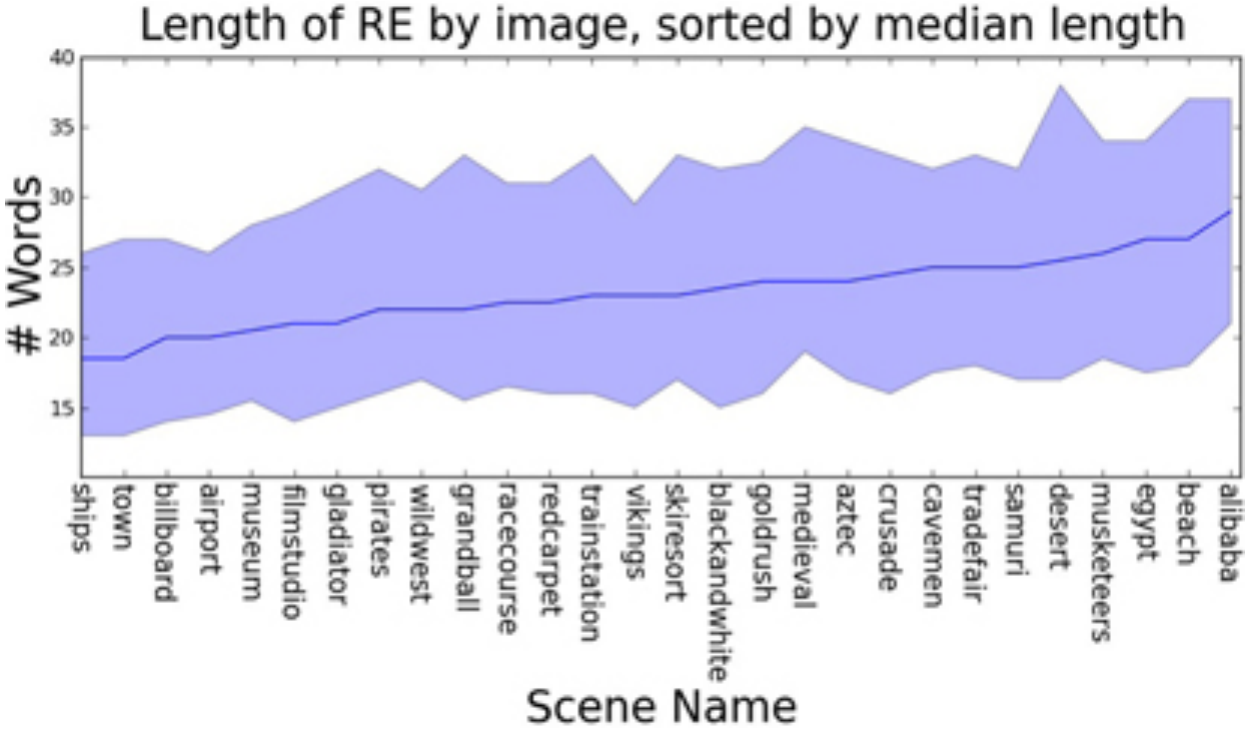
Annotation scheme



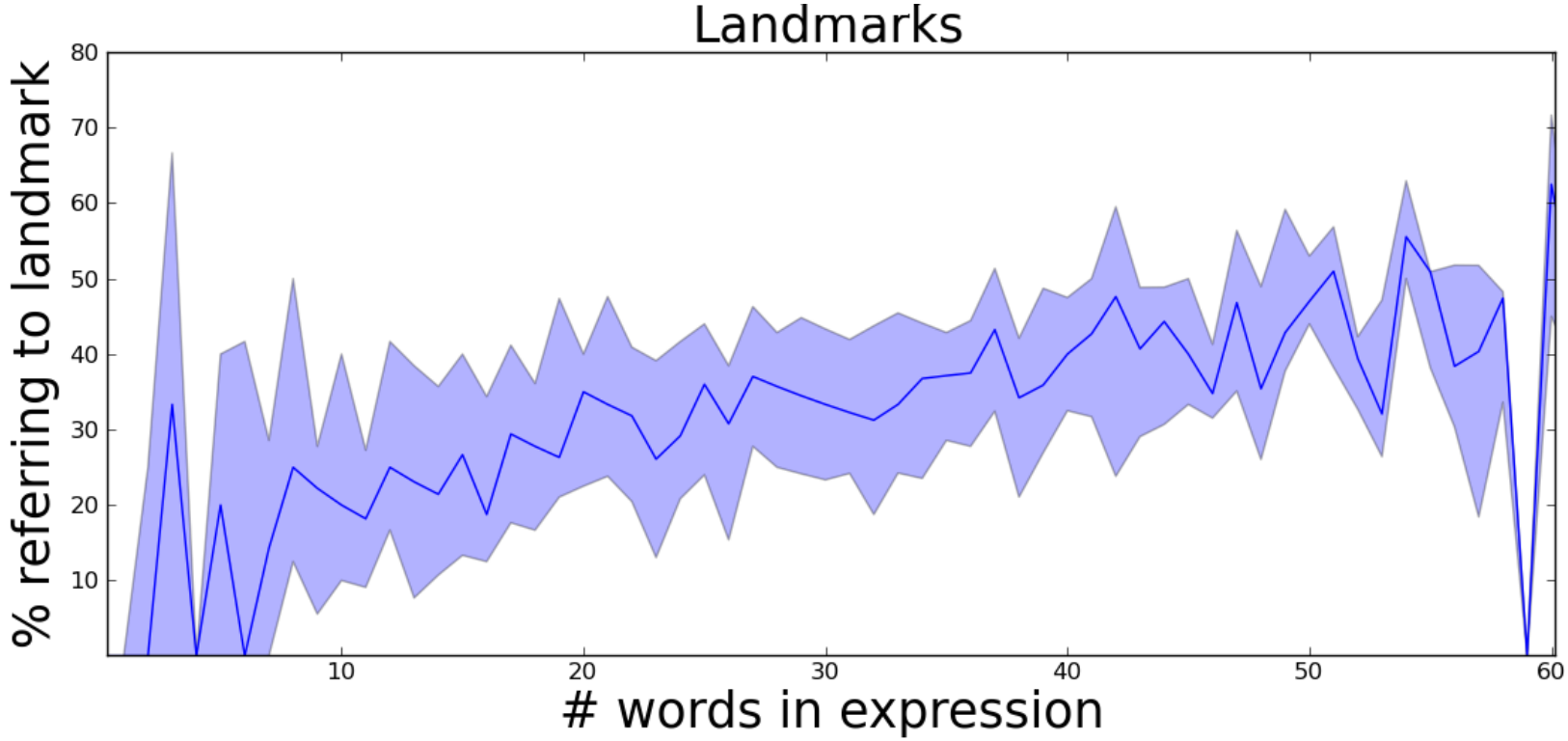
Under <lmark rel="targ" obj="imgID"> **a net** </lmark> is <targ> **a small child wearing a blue shirt and red shorts** </targ>.

Descriptions vary in length

More cluttered images have longer descriptions (ρ = .45)



Longer descriptions, more landmarks



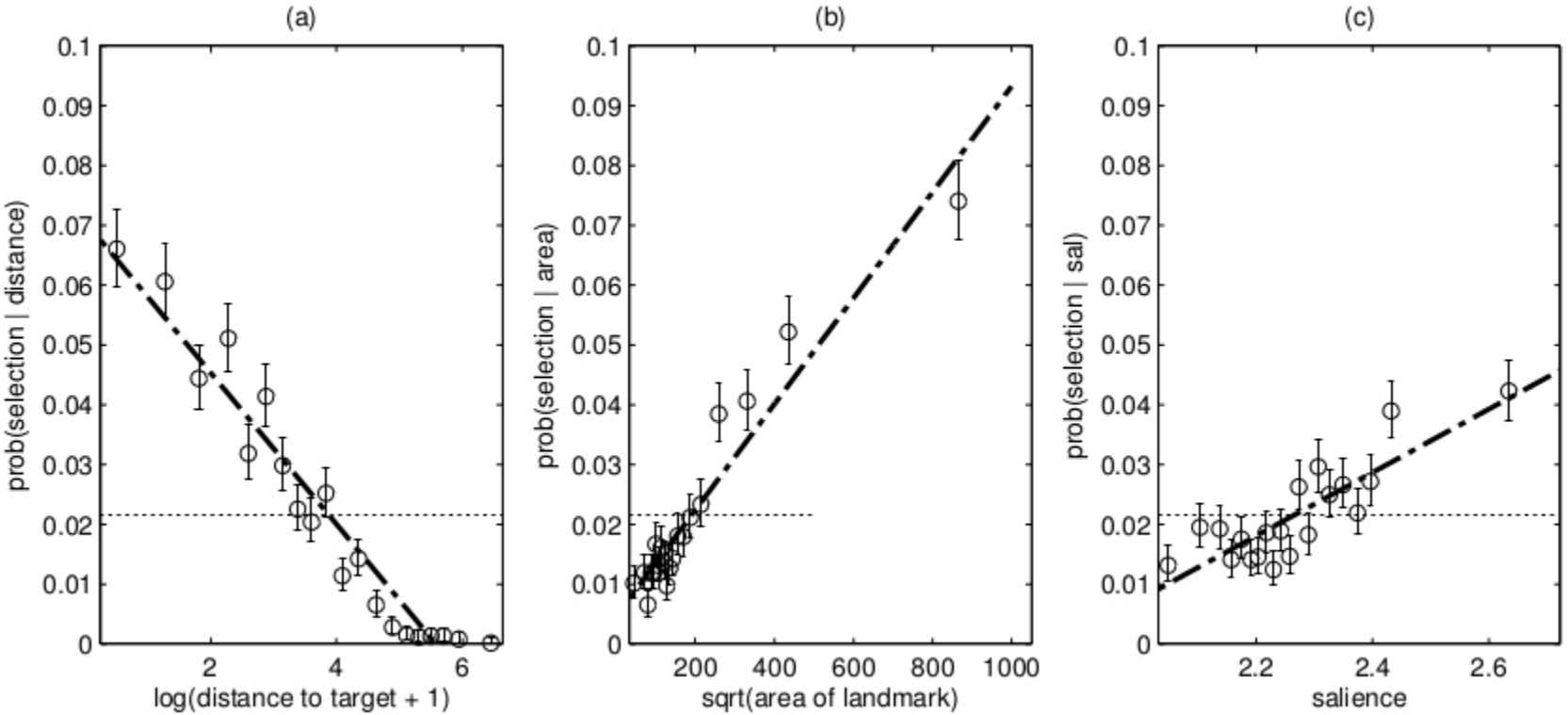
Use a relational description?

Larger, more salient targets take up more of the description

Mixed-effects regression: % of words referencing target
(significant effects only)

	β	std error
Area of target	.25	0.05
Torralba salience model	.20	0.05
Area : salience model	-.11	0.04

Most landmarks: close, large, salient



Discourse structure

Duan, Elsner and de Marneffe, SemDial 2013
Elsner, Rohde and Clarke, EACL 2014

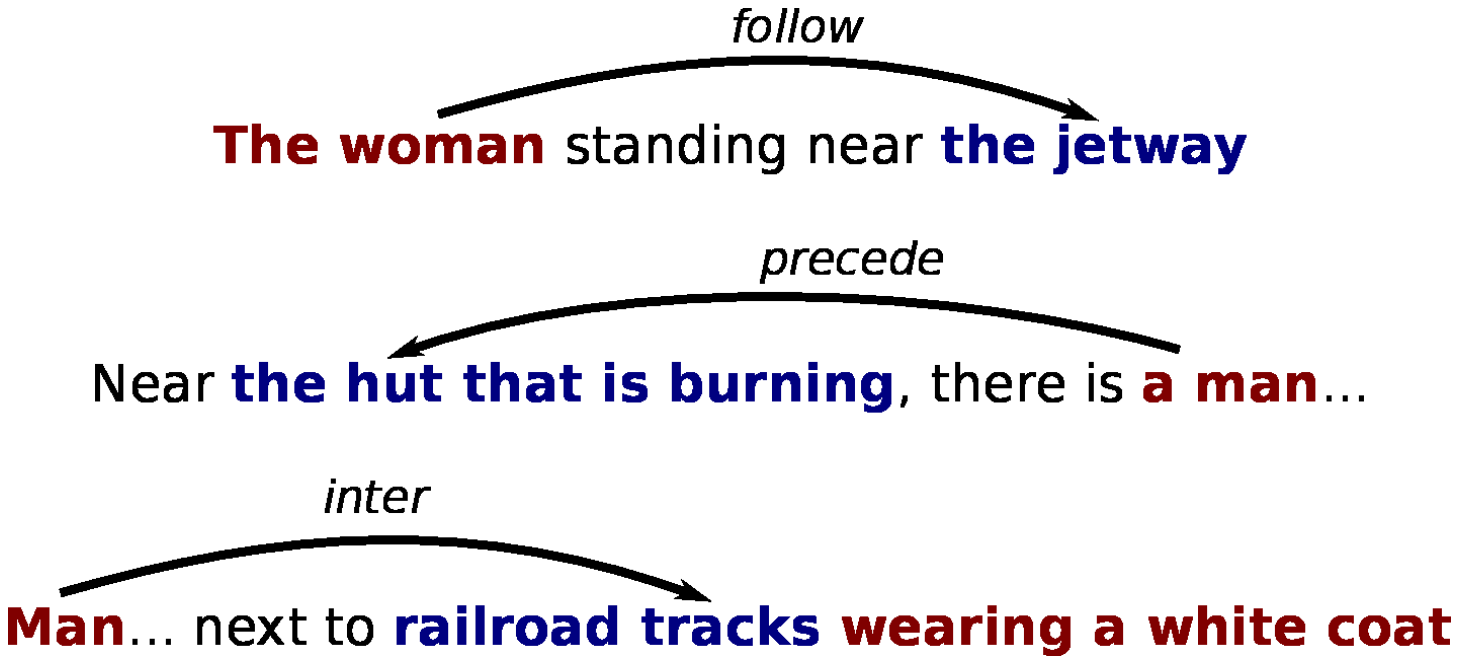
Linguistic form

- So far: what to say
- Also important: **how to say it**
  - Interface between perception and discourse

- Two studies:
  - Ordering
  - Definiteness / referring form

Ordering mentions

Surface order of **target** and **landmark**



Establish construction

Look at **the plane**. **This man is holding a box that he is putting** on **the plane**.

- First mention isn't relational
  - *There is*, *look at*, *find the*…
- Almost always with **precede** order

Basic results

- **follow** (38%) and **precede** (37%) equally likely for landmarks
  - Regions usually **precede** (60%): *on the left is a…*
  - **inter** about 25%
- Again, massive individual differences
  - For target / landmark pairs mentioned by two subjects, 66% agreement on direction

Predicting order

Mixed-effects regression; only significant effects shown; visual features of landmark

Feature	Precede	Precede-Establish	Inter	Follow
Intercept	-4.18	-2.66	-2.51	2.72
Img region?	11.46		3.01	-12.62
Lmark area	3.27		1.28	-3.76
Lmark centrality				0.81
Lmark #lmarks		2.38	-1.07	-1.37

- Regions prefer to precede
- Larger landmarks prefer to precede
- Landmarks with landmarks prefer own clause

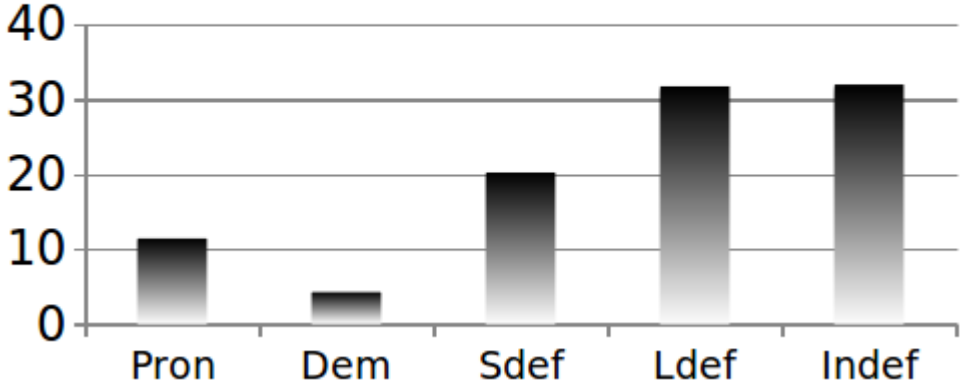
Visual and discourse salience

- Usual ordering principle: given before new
  - **Obama (given)** has **a dog named Bo (new)**

- Similarly, large landmarks prefer to **precede**

Referring form of NPs

- Pronoun: *it*, *she*
- Demonstrative: *that man*
- Short definite: *the car*
- Long definite: *the man in blue jeans*
- Indefinite: *a tree, some people*
- Bare singular: *brown dog* (grouped with definites)



Distribution of referring forms (%)
N=9479

Hierarchy of referring forms

(Ariel 88), (Prince 99), (Gundel 93), (Roberts 03) etc

familiar entities	it	that N	the N	a N	*new entities*

- *Familiarity* usually discourse-based
- Perception also creates familiarity
  - But earlier theories unclear about how
- Again, visual salience like discourse salience

Predicting forms: visual features

Mixed-effects one-vs-all regressions; only significant effects shown

Features	Pron	Dem	SDef	LDef	(Def)	Indef
Area	-1.99	-0.94	0.71	-0.40	1.51	-1.78
Pix.Sal.	-0.25					
Overlap		-0.91		-0.43	-0.45	0.53
Distance	0.38		0.15	0.13	0.43	-0.87
Clutter				-0.43		
Area:Clutter			0.28	-0.09	0.27	-0.22
Sal.:Clutter				-0.09	-0.10	0.15

- More definites for objects far from the target
- Fewer definites in crowded images

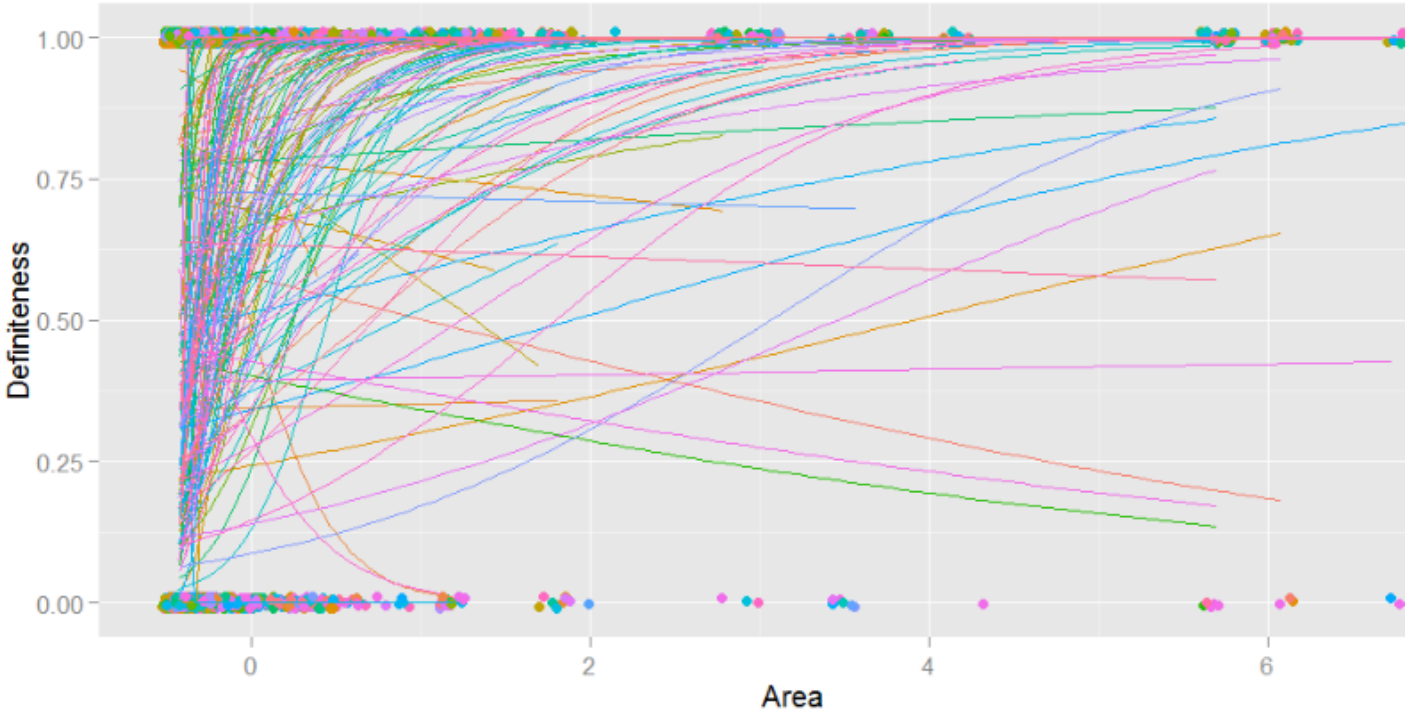
Linguistic features

Mixed-effects one-vs-all regressions; only significant effects shown

Features	Pron	Dem	SDef	LDef	(Def)	Indef
Coref	4.68	0.73		-1.63	-1.37	-2.60
Existential			-3.64	-3.89	-4.70	5.77
After *be*	-3.31	-3.21	-2.12	-2.78	-3.07	4.24
Sent. Initial	0.91		-0.52	-0.28	-0.56	0.46
After prep					0.26	-0.40
Establish: “find the”		2.20		-0.54	-0.71	0.45

- Linguistic effects larger than visual
- Essentially as expected

Effects vary across individuals



Classification

On held-out test sets:

- 57% order (**precede**, **follow**, or **inter**)
  - 42% baseline (lmarks **follow**, regions **precede**)
  - 66-76% subject agreement
- 62% referring form (**pron**, **dem**… etc)
  - 56% without visual features

Descriptions in real time

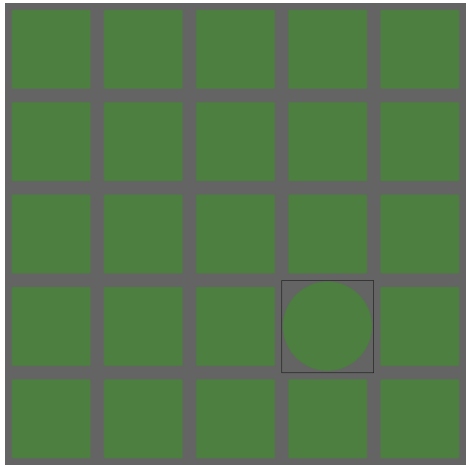
Rohde, Elsner, Clarke CUNY 2014
(poster)
and work in progress

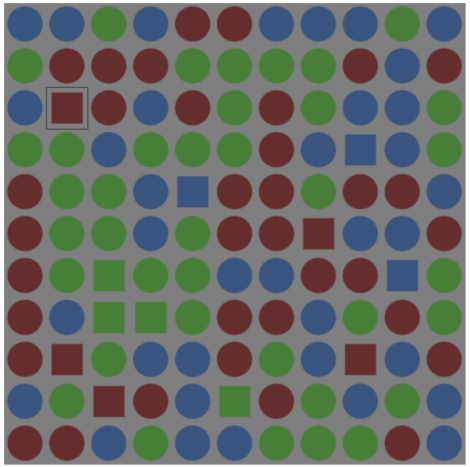
How does incrementality work?

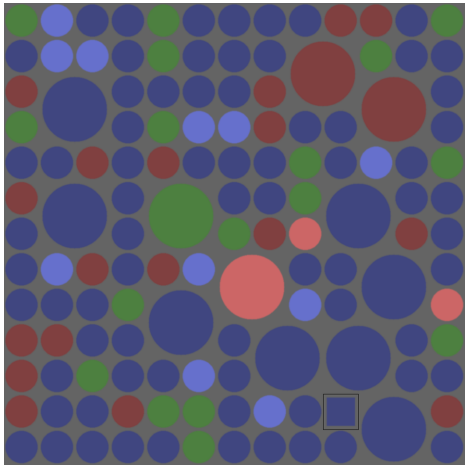
When do speakers do the visual ‘work’ for descriptive elements?

What do they know, and when do they know it?

Experimental setup

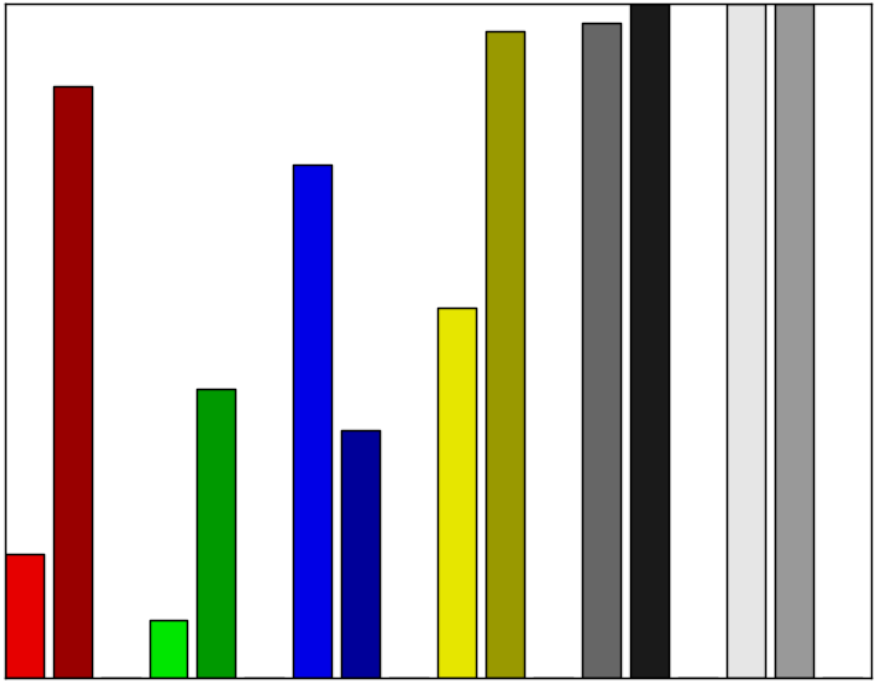






- 20 subjects each saw 120 random object arrays
- Varied heterogeneity, size, presence of distractor
- Speech and eyetracking

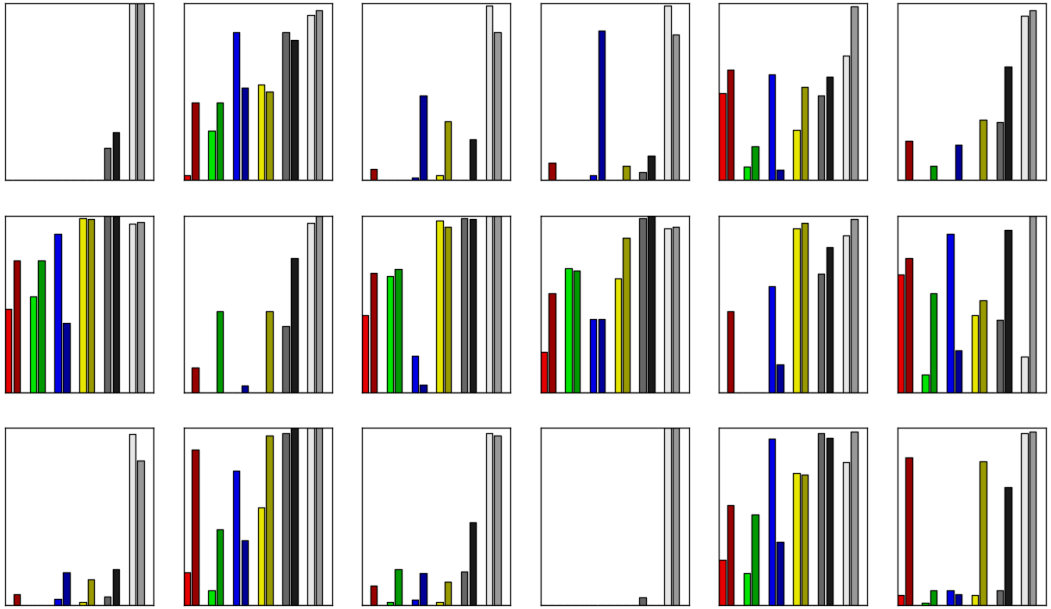
Phrase type effects



Proportions of descriptive elements, single subject:

Coordinates:
two rows down
Landmark:
next to the big square
Scene-relative:
the only circle
Region:
on the left
Other:
you're looking for
Target:
small red circle

Major effects of distractor



Proportions of descriptive elements, 18 subjects:

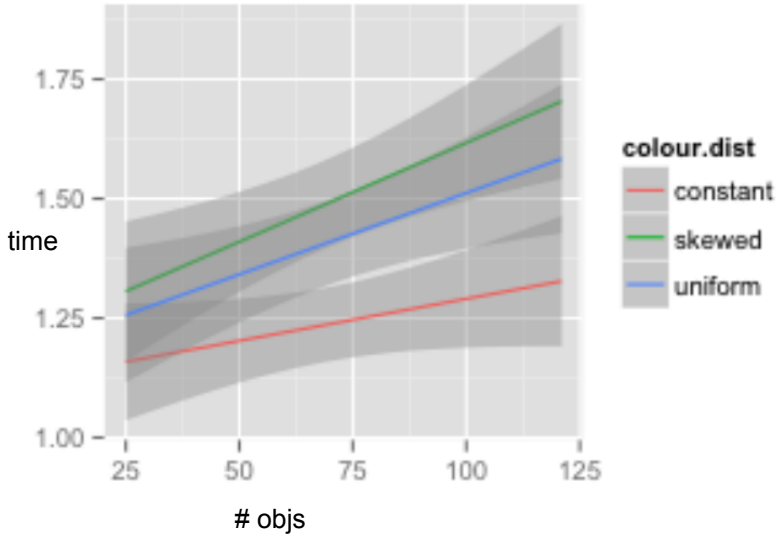
Coordinates:
two rows down
Landmark:
next to the big square
Scene-relative:
the only circle
Region:
on the left
Other:
you're looking for
Target:
small red circle

Speech onset times

- When do subjects notice a distractor?
- Before or after they talk?

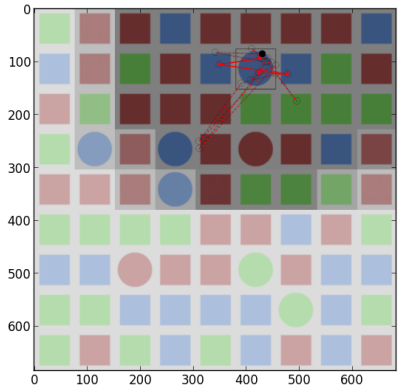
Speech onset:
~1.5 sec

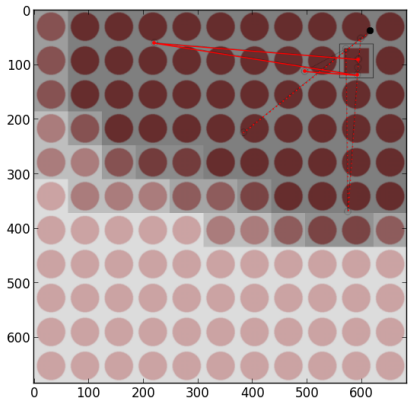
Small effect of scene type



Distractor probably seen early

- Simplistic model of visual search
  - Distance thresholds (per size and type)
  - Estimated heuristically from object - fixation dists





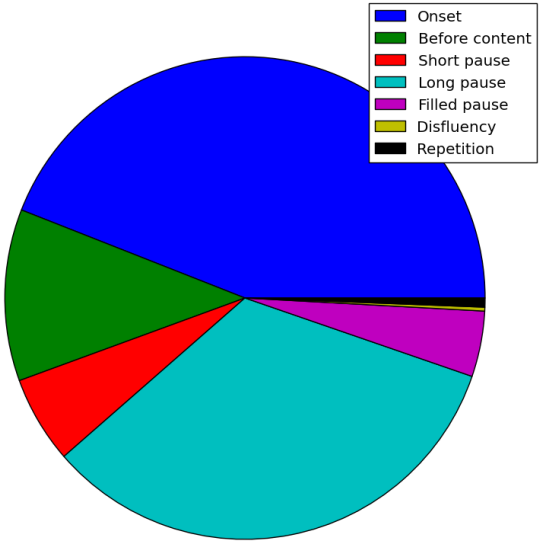
How much incrementality?

- About 1.5 sec to scan before speaking
  - Probably see distractor if present
  - Allows top-level decision about how much content
- Top-level decision not incremental...
- Are finer-grained decisions?

How speakers waste time

- Pre-onset
- Onset to first content
- Pauses (short, < .25s; long otherwise)
- Filled pauses *um, uh, well, okay* etc.
- Disfluencies [*cir---*] *circle*
- Repetitions [*the green*] *the green circle*

Long pauses are common



When speakers waste time

Which descriptive elements are associated with long pauses?

Mixed-effects model of long pause duration, residualized for total words in utterance
Largest fixed effects shown

Intercept	-1.1
Distractor present?	.22
# Shape terms *(next to the square)*	.19
# Scene-relatives *(only)*	-.17
# Coordinates *(second row)*	.13

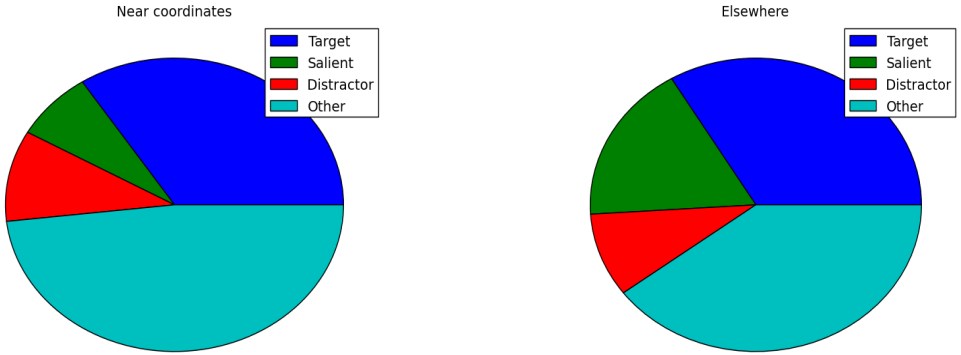
Coordinates and landmarks are slow

- Speakers pause more:
  - When a distractor is present
  - When using landmarks (probably)
  - When using coordinates
- Speakers pause less:
  - When using scene-relative terms
  - Most common is *only*

Visual behavior during wasted time

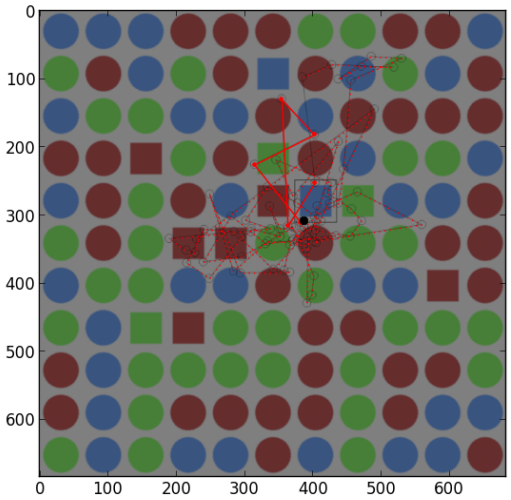
What is this wasted time *for?*

- What do people do while they waste it?
  - Near coordinates, more looks at non-salient shapes



Does this reflect counting?

It's a blue square um **f-- four five columns in from the right and five columns down** from the top. It has a red square on its left hand side and a green square on its right hand side...



What's *really* going on is an open question…

We're still having trouble categorizing visual behaviors.

Interim analysis

- Suggests phrase type is planned first…
- Before specific content is known

- Wasted time: visual confirmation of phrase content?

Conclusions

- Language and visual perception interact at many levels
- Visual effects on form as well as content
  - Including “discourse”-type phenomena
- Tentative support for incremental planning…
  - Contents of phrases underspecified until called for
  - We hypothesize: feature values underspecified too

Open question: Grice vs laziness

Grice's maxim of quantity:

> Make your contribution as informative as is required.
> Do not make your contribution more informative than is required.

- Are perception effects (mainly) Gricean?
  - Intended to make listeners' tasks easier
- Or (mainly) speaker-driven?
  - By perceptual / cognitive limitations or laziness

Planned experiments on listeners may help...