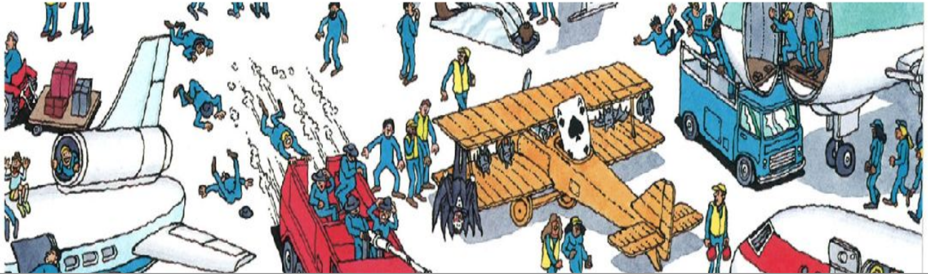
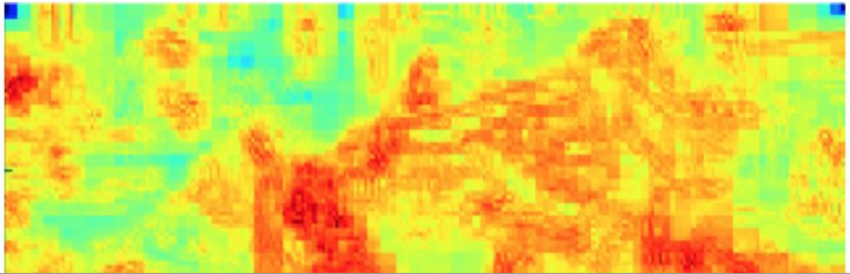
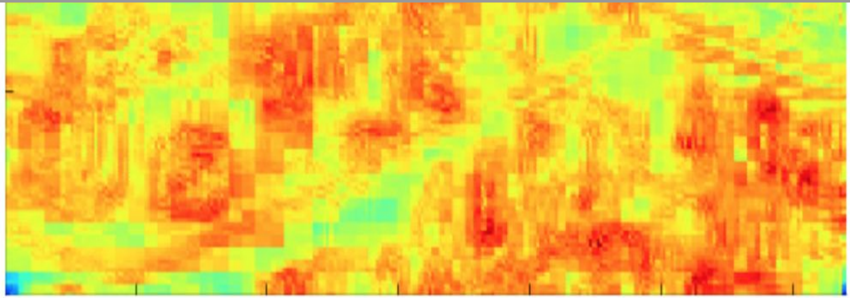



Language and Visual Processing

Micha Elsner
The Ohio State University

Thanks to:

Alasdair Clarke:
Psychology, University of Essex

Hannah Rohde:
Linguistics and the English Language, University of Edinburgh

Manjuan Duan:
Amazon

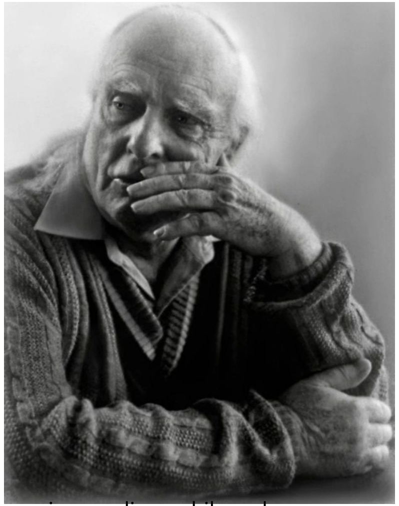
Marie-Catherine de Marneffe:
Linguistics, OSU

and Stephanie Antetomaso, Marten van Schijndel, Emma Ward, Amelia Hunt

Reference in visual worlds

Speaker's task: identify a **target** object

Gricean principles


img: ordinaryphilosophy.com

Quantity: Give as much information as necessary and not more.

In visual setting, implies two design goals:

Uniquely identify the target

But don't **overspecify**

Candidate description

Speaker's task: identify a **target** object

Doesn’t uniquely identify

Speaker’s task: identify a **target** object

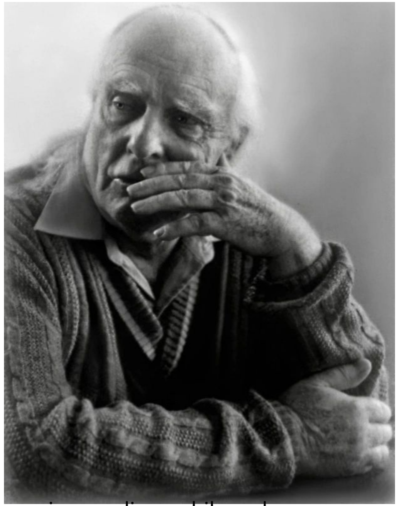
Candidate description

Speaker's task: identify a **target** object

Overspecified

Speaker's task: identify a **target** object

Gricean reasoning is expensive



img: ordinaryphilosophy.com

Gricean principles require us to think about **counterfactuals**…

What description strategies could I use?
(“black”, “right of”, “has little spikes”)

What other objects might they apply to?

influential neo-Gricean research by Frank and Goodman 2012; Jaeger 2010; Degen, Franke and Jaeger 2013, and others

Is "spire" adequate?

Speaker's task: identify a **target** object

Reducing the burden on listeners

Taking vision into account helps listeners to find the target quickly:

"black spire" not only eliminates some competing spires
but does so **efficiently**---
white buildings can be screened out pre-attentively

Overspecification, particularly of color, is probably helpful

see Arts et al 2011; Koolen et al 2011 and others

Reducing the burden on speakers

Speakers take shortcuts, leading to descriptions which are not always optimal for their listeners…

Especially under pressure!

Horton and Keysar, 1996; Beun and Cremers 1998; Bard et al 2003 and others

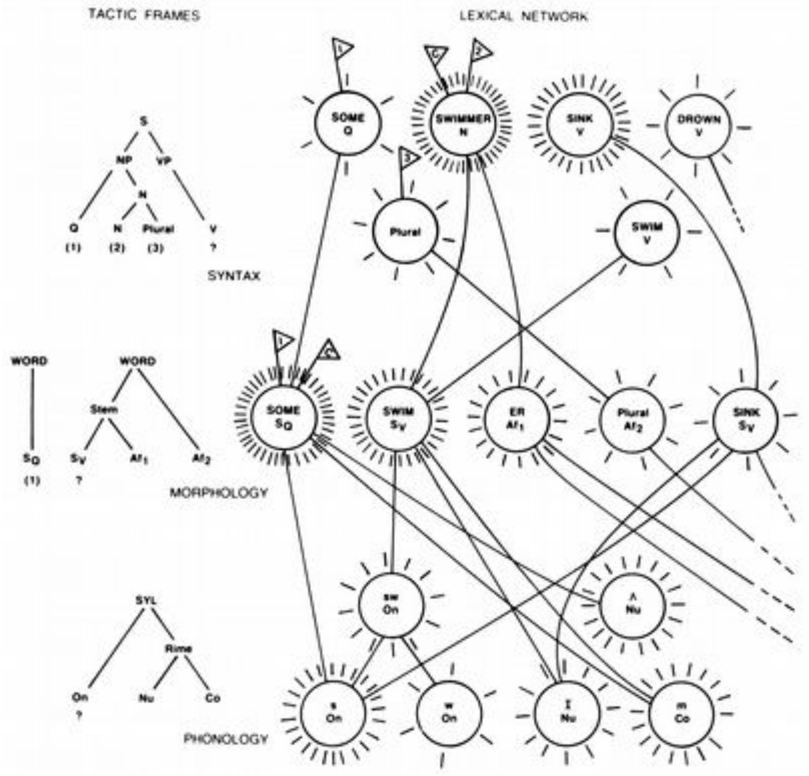
Standard models of speech production

Dell 1986, Levelt 1989 and others

Speech planning is:

- Incremental
- Hierarchical
- Subject to revision

Real-time planning can't always keep up with Gricean ideals



Dell 1986

How vision makes a difference

What is said? Content and discourse structure

When is it said? Eye-tracking and timing data

Why is it said? Cognitive modeling with neural nets

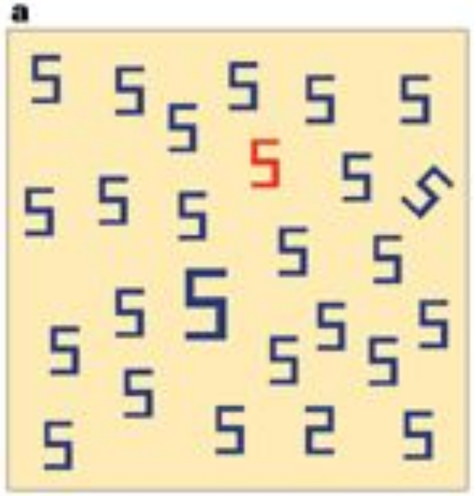
Content and discourse: Where's Wally?

Material from Clarke, Elsner and Rohde 2013,
Duan, Elsner and de Marneffe 2013,
Elsner, Rohde and Clarke 2014
Clarke, Elsner and Rohde 2015

“Visual salience”

The visual system is good at finding unique colors…

Not so good at finding uniquely sized objects quickly



It is easy to find the red, tilted or big '5'. It is not easy to find the '2' among the '5's.

Wolfe and Horowitz 2004

“Where’s Wally” corpus

“Where’s Wally” (Handford)...

- A game based on visual search

Corpus collected on Mechanical Turk

- Selected human targets in each image
- Subject instructed to describe target so another person could find them

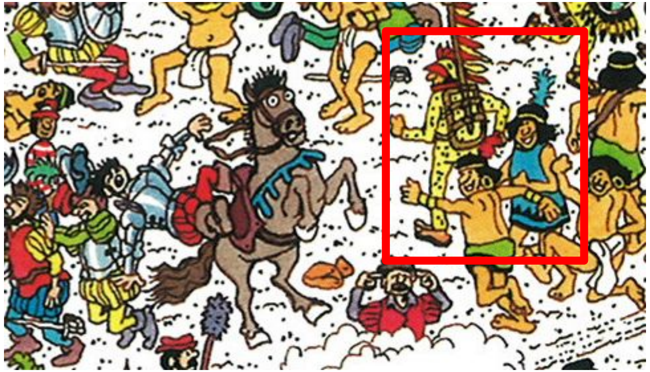
Download: http://datashare.is.ed.ac.uk/handle/10283/336

Sample descriptions...

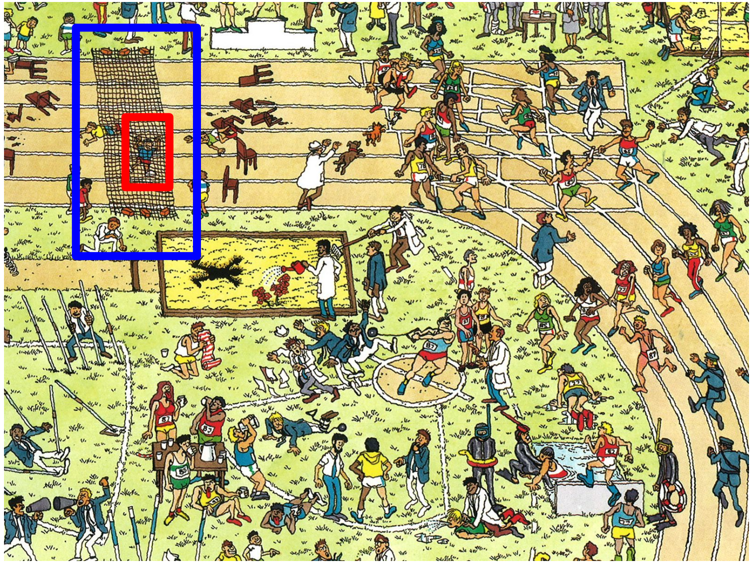
"Man running in green skirt at the bottom right side of picture across from horse on his hind legs."

"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."



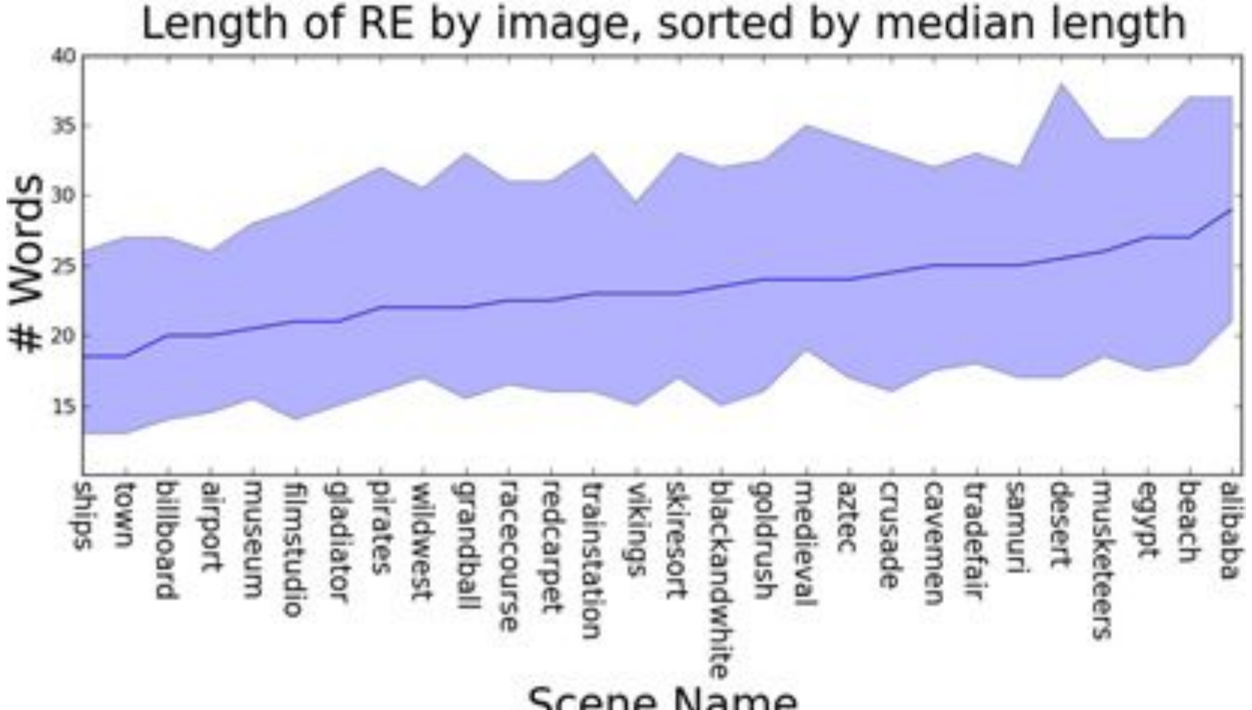
Annotation scheme



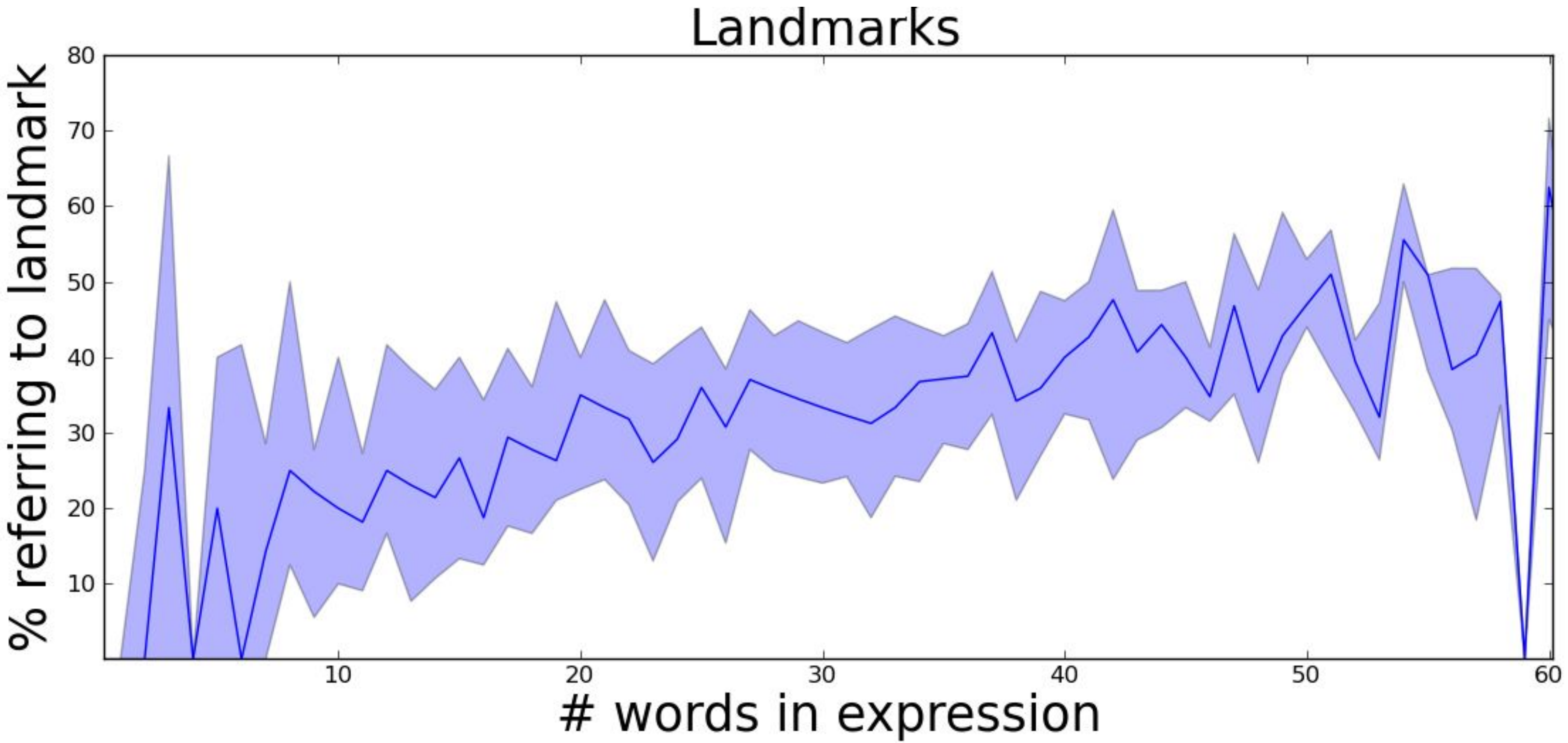
“Under <lmark rel=“targ” obj=“imgID”> **a net** </lmark> is <targ> **a small child wearing a blue shirt and red shorts** </targ>.”

Descriptions vary in length

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



Longer descriptions, more landmarks

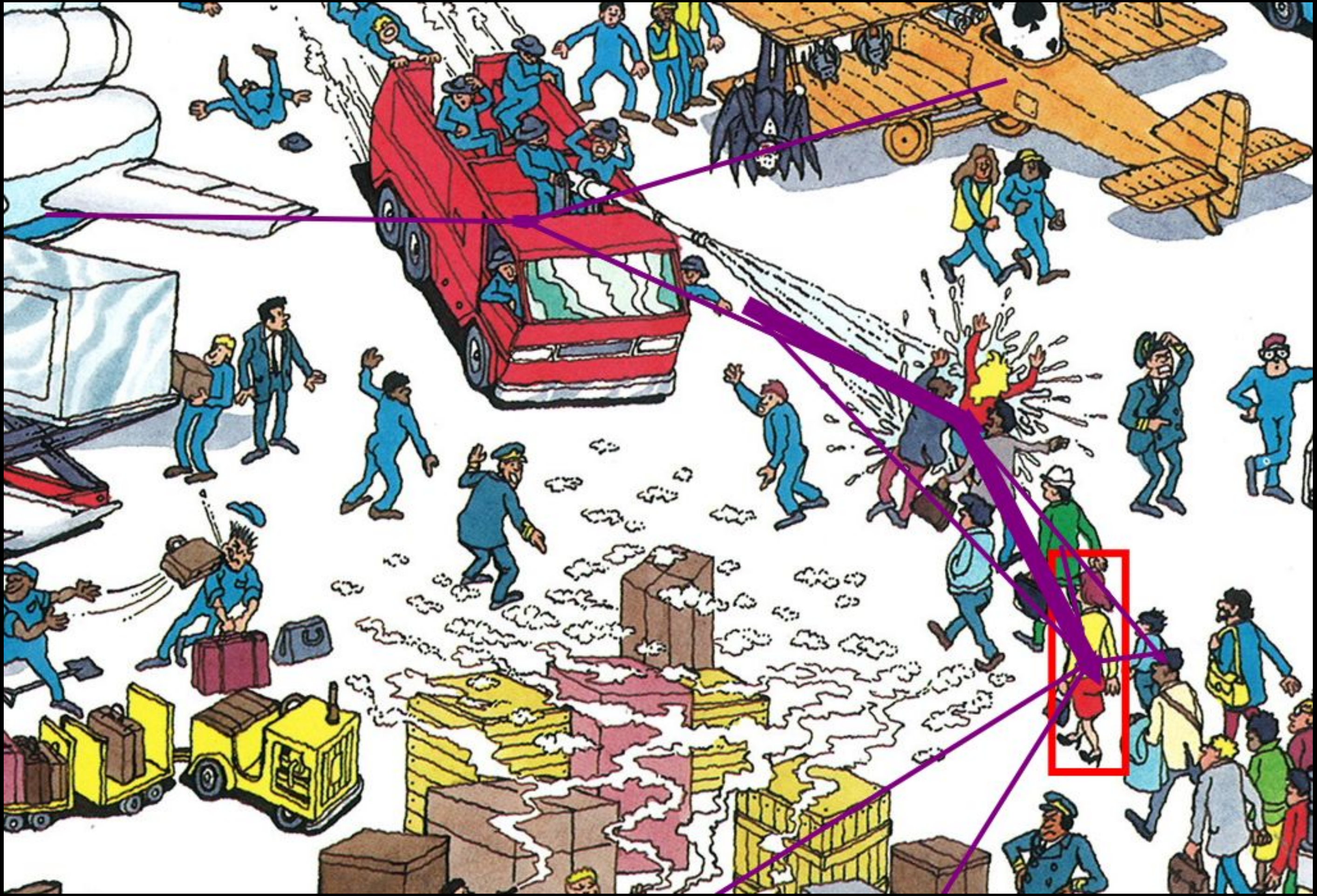


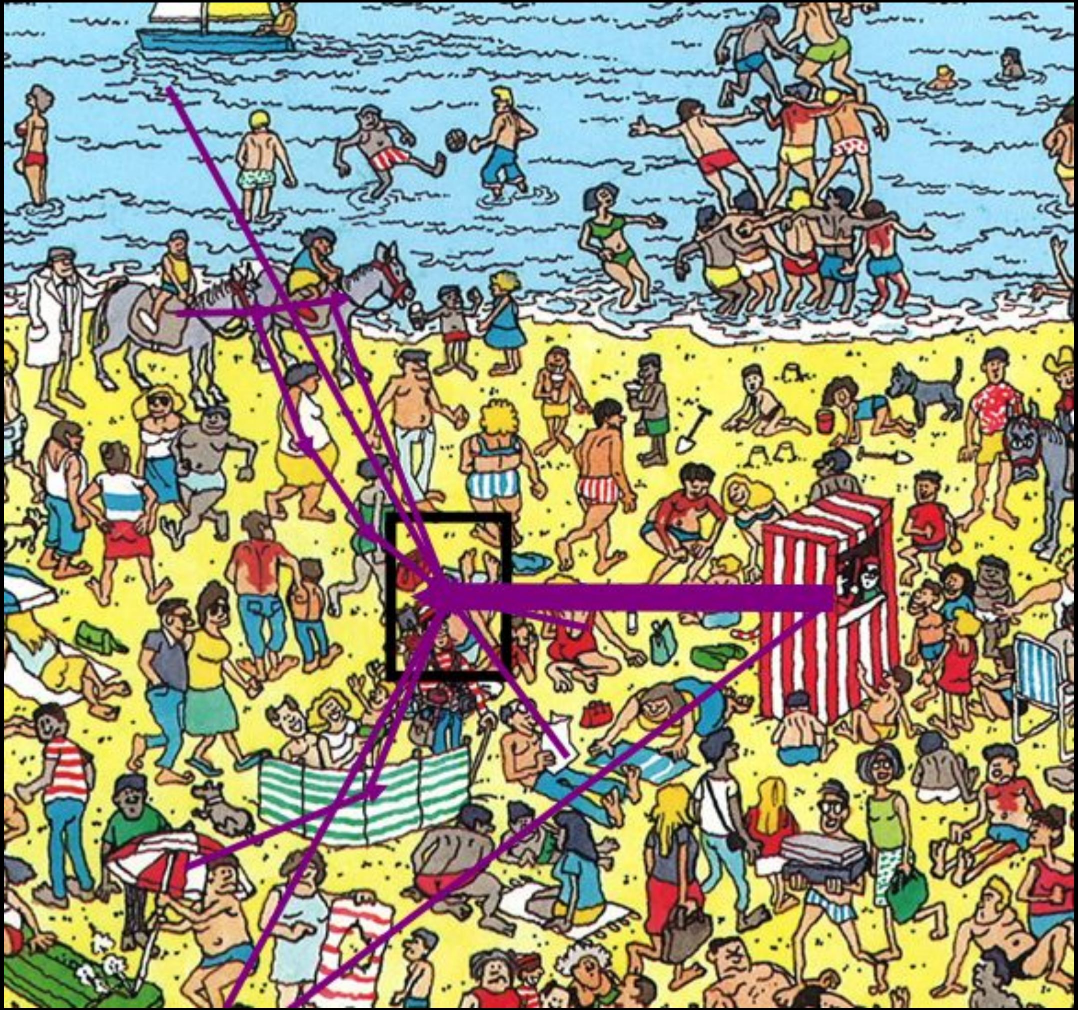
Use a relational description?

Larger, more salient targets take up more of the description

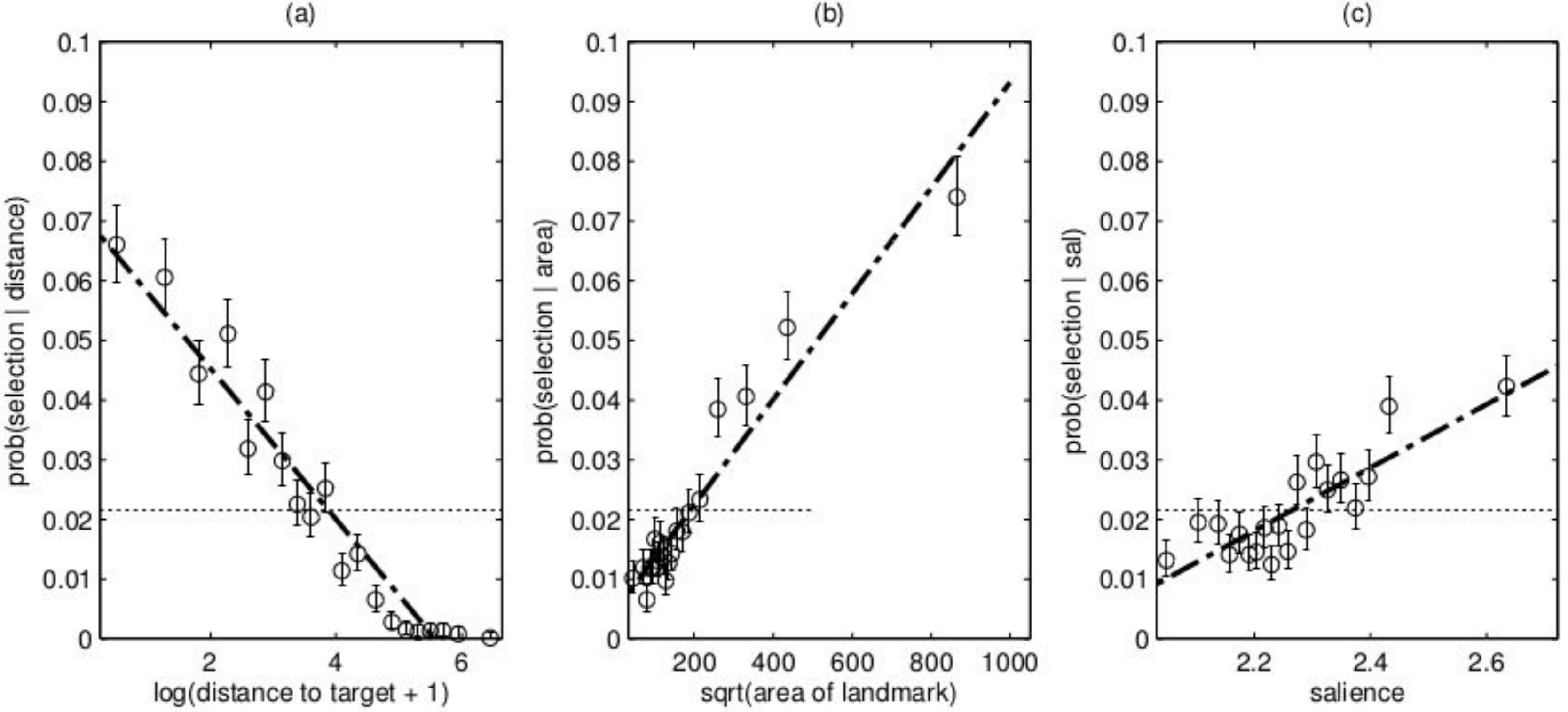
Mixed-effects regression: % of words referencing target
(significant effects only)

	β
Area of target	.25
Visual salience model	.20
Area : salience model	-.11





Most landmarks: close, large, salient



Hierarchy of referring forms

Ariel 1988; Prince 1999; Gundel 1993; Roberts 2003 and others

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

Prediction: Easy-to-see objects more *definite*

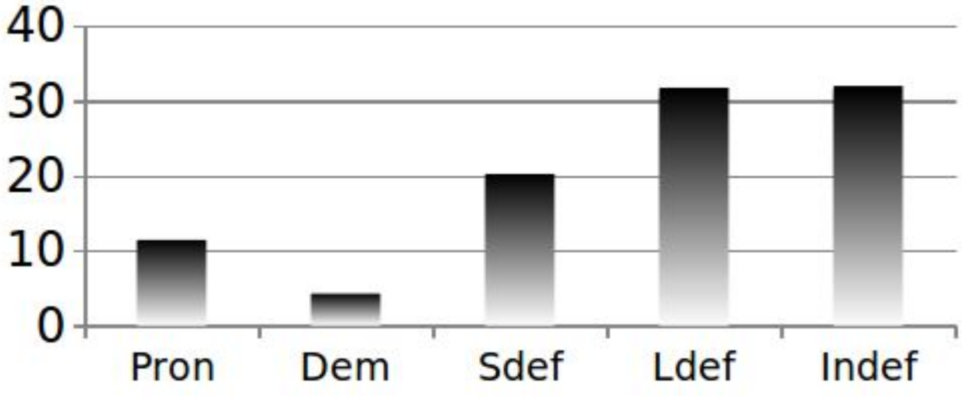
Hard-to-see objects more *indefinite*

Definites require uniqueness (in a set)

Fewer *definites* in cluttered image

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

Predicting forms: visual features

Features	Pron	Dem	SDef	LDef	(Def)	Indef
Area	-1.99	-0.94	0.71	-0.40	1.51	-1.78
Distance	0.38		0.15	0.13	0.43	-0.87
Clutter				-0.43		

- Large objects prefer short definites over indefinites
- More definites for objects far from the target
- Fewer definites in crowded images

Visual and discourse salience

Similar behavior from both kinds of salience

Linguistic effects usually stronger (as in Viethen et al 2011)

But visual effects are important

These experiments focused on speakers

- In a subsequent study, we found that listeners find the target faster when landmark mentions are visually appropriate

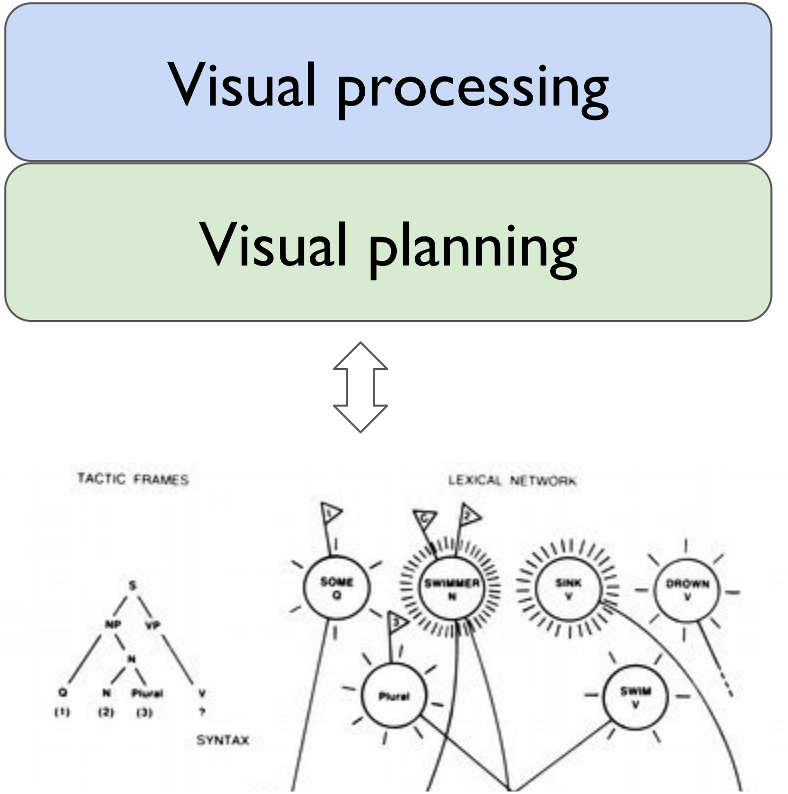
Descriptions in real time

Elsner, Clarke and Rohde 2018

Language in real time

Vision matters for what speakers say…
A window into the planning process.

How far in advance are people planning?
What evidence do they use to make decisions?

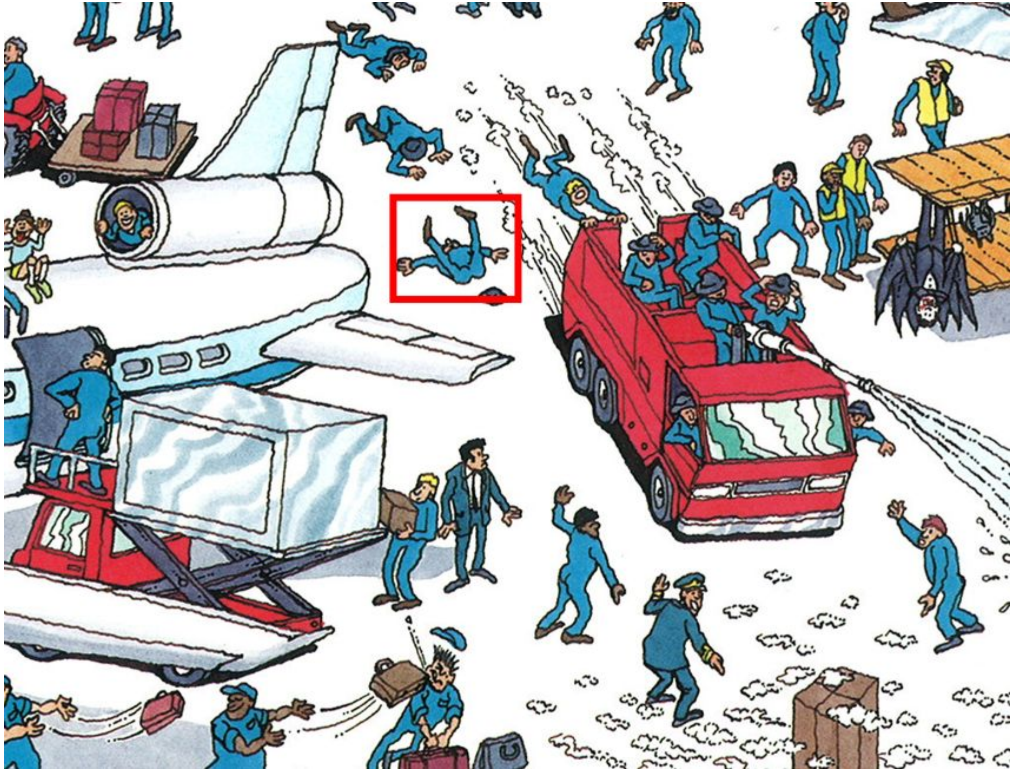


Why do different people make such different plans?

"Man on the ground to the left behind the fire truck laying on the ground with his legs in the air."

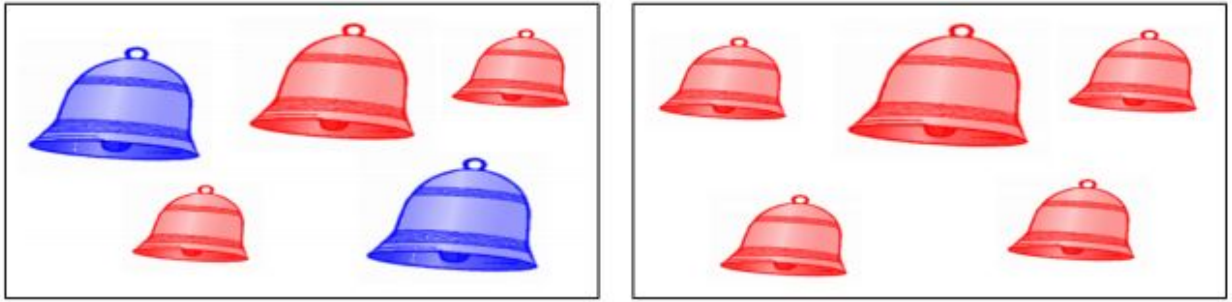
"Top left of the picture. A man falling from the sky with his legs up in the air. Next to the side wing. Looks like he is sleeping with his face up."

"In the top left of the picture, between the plane and the fire engine, a man is falling backwards, dressed in blue with his legs up."



Gatt's experiments

"Reference production as search: The impact of domain size on the production of distinguishing descriptions" Gatt et al 2017

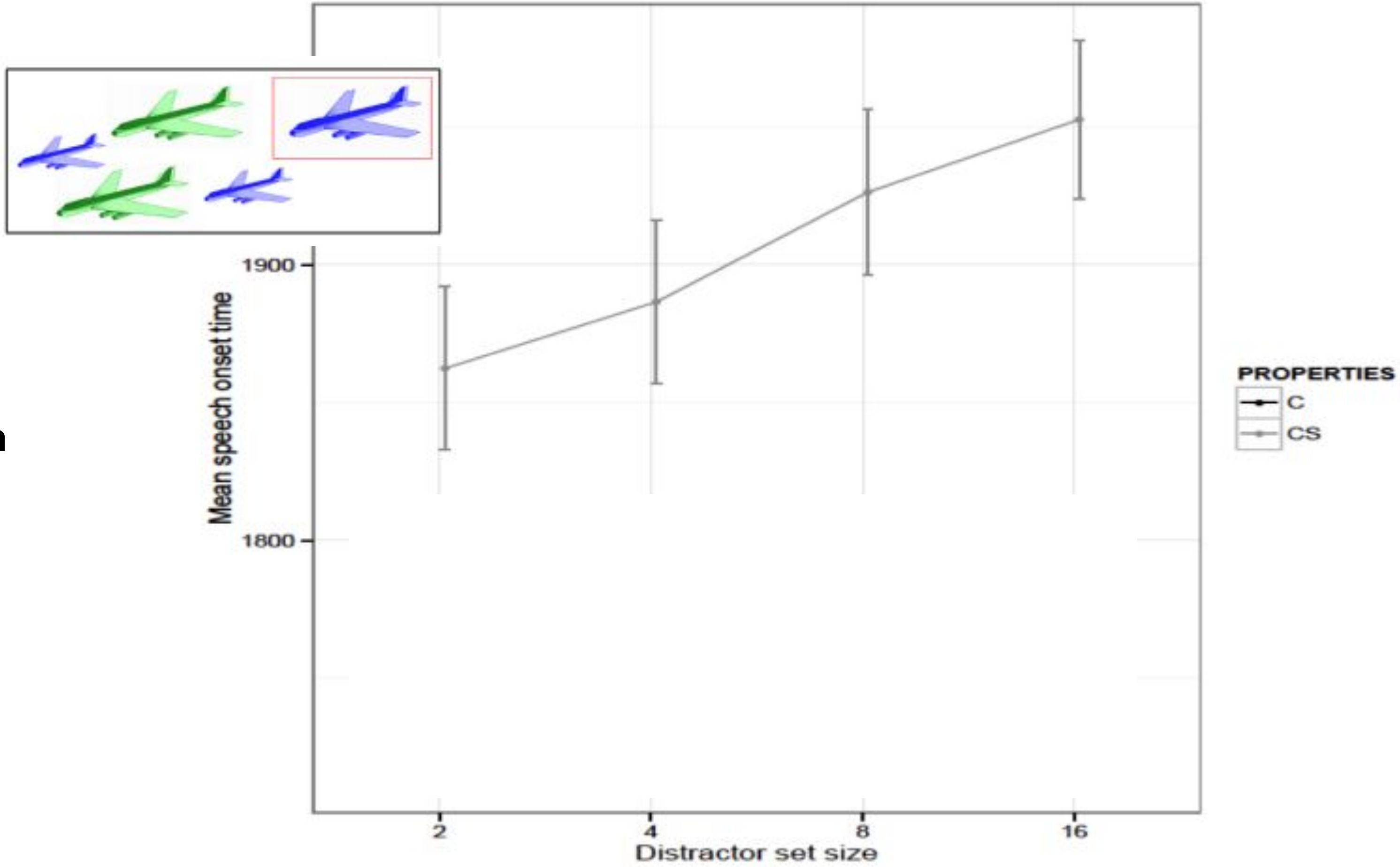


(b) A *large red bell* among large blue and small red distractors

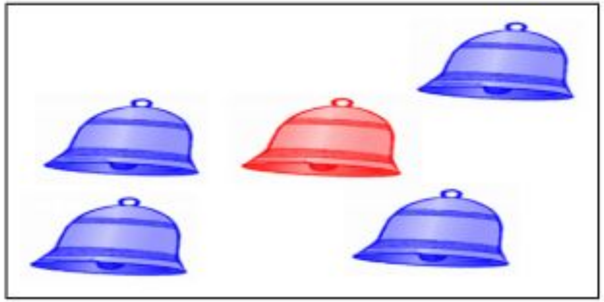
(c) A *large bell* among smaller distractors

Gatt varied the number of bells in the scene...

Time
before
speech
onset



But some cases are easy

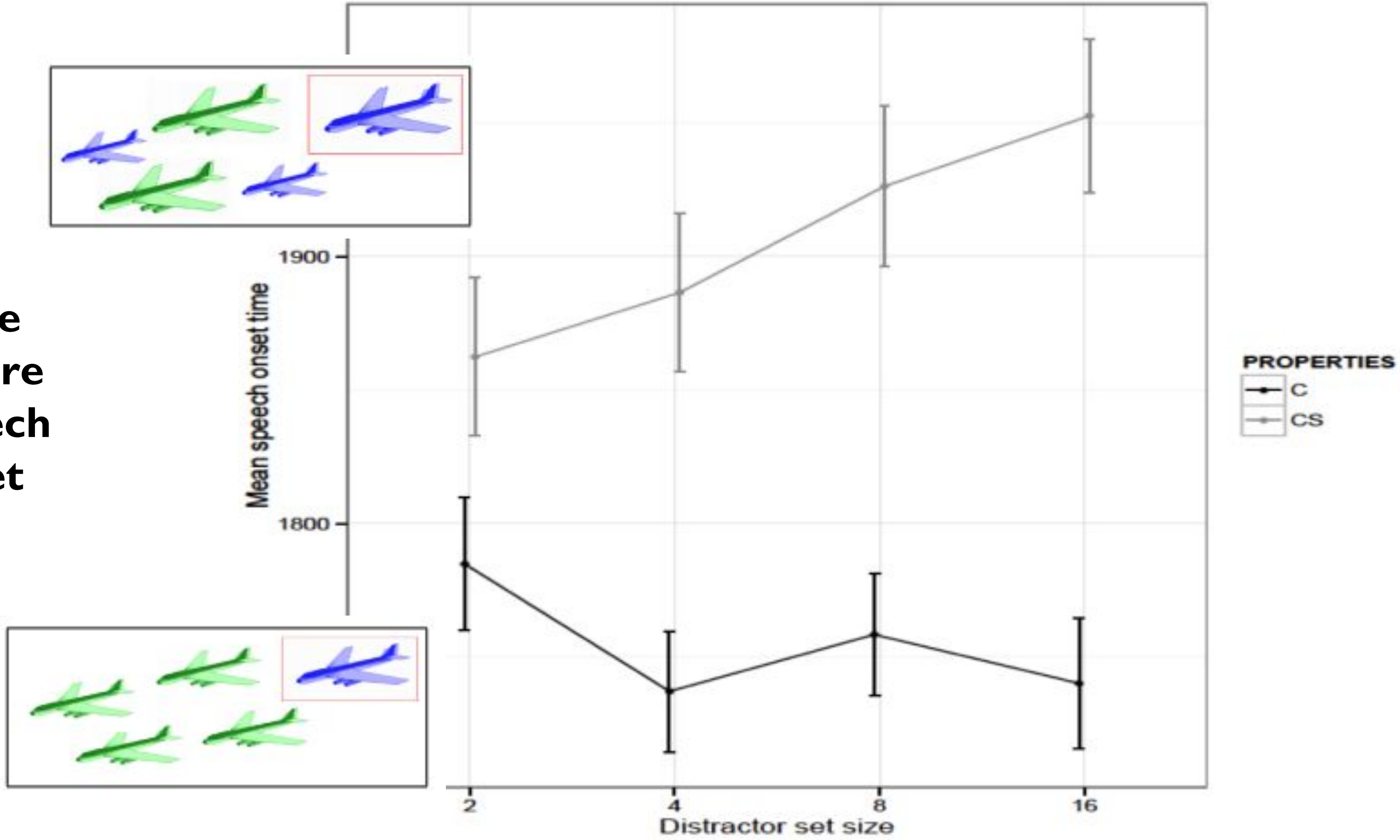


(a) A *red bell* among blue distrac-

tors

Gatt varied the number of bells in the scene...

Time before speech onset



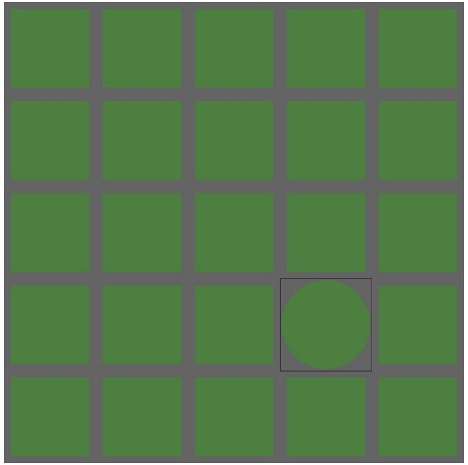
Gatt's results suggest a simple model

For these stimuli, the referring expression is **precomputed** by an **optimal Gricean process**

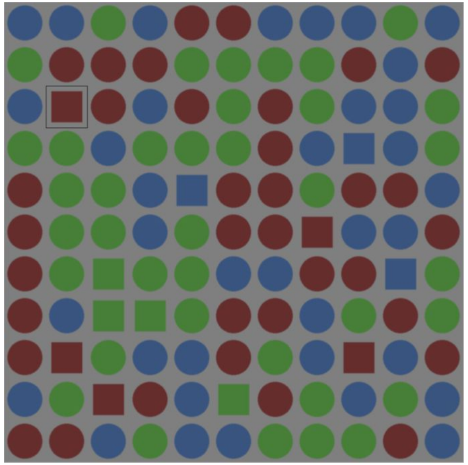
Potentially involving **exhaustive** search of the other objects to check uniqueness

But this is only true if search is **easy**!

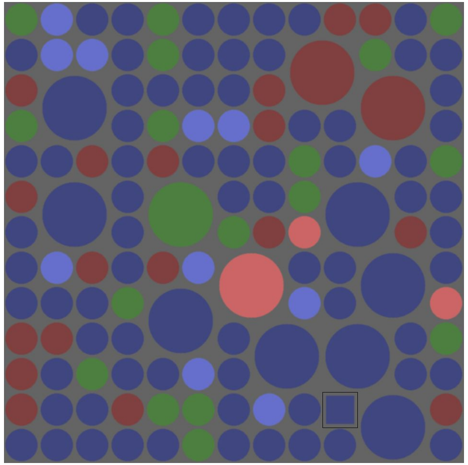
Our follow-up study



5x5 uniform
unique target



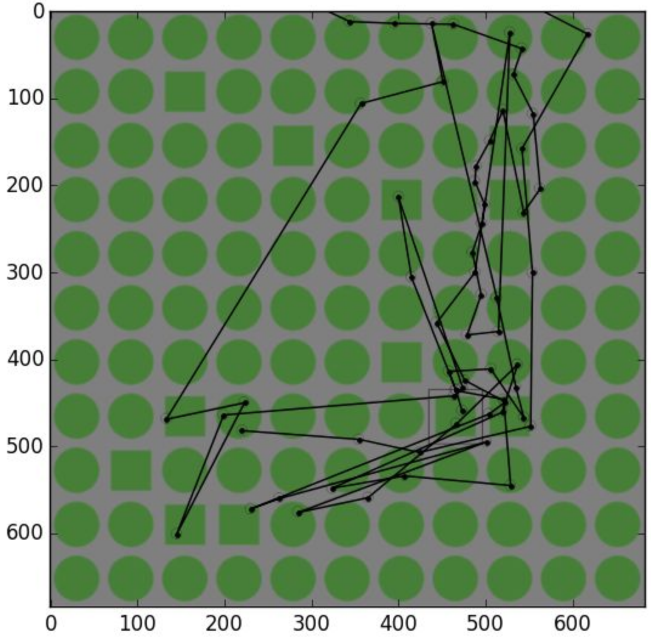
11x11 multicolor
ambiguous target
“red square” not enough



11x11 skewed
distr.
unique target

Gaze tracks

We can see the speaker counting out rows and columns



“ooh green square uh f[rom] on the eighth row and ninth column”

Planning is complex

Speakers choose adaptively between strategies:

Unique, distinctive target ⇒ simple target description

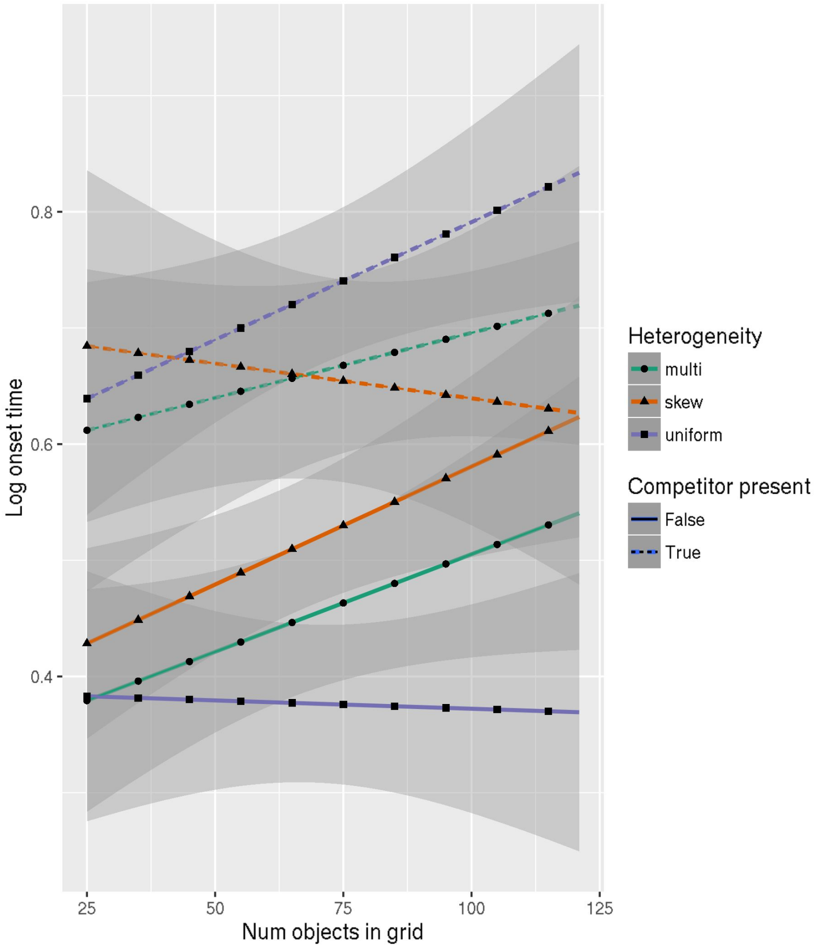
Larger grids ⇒ spatial descriptions like "top left"

"Skewed" scene ⇒ landmarks

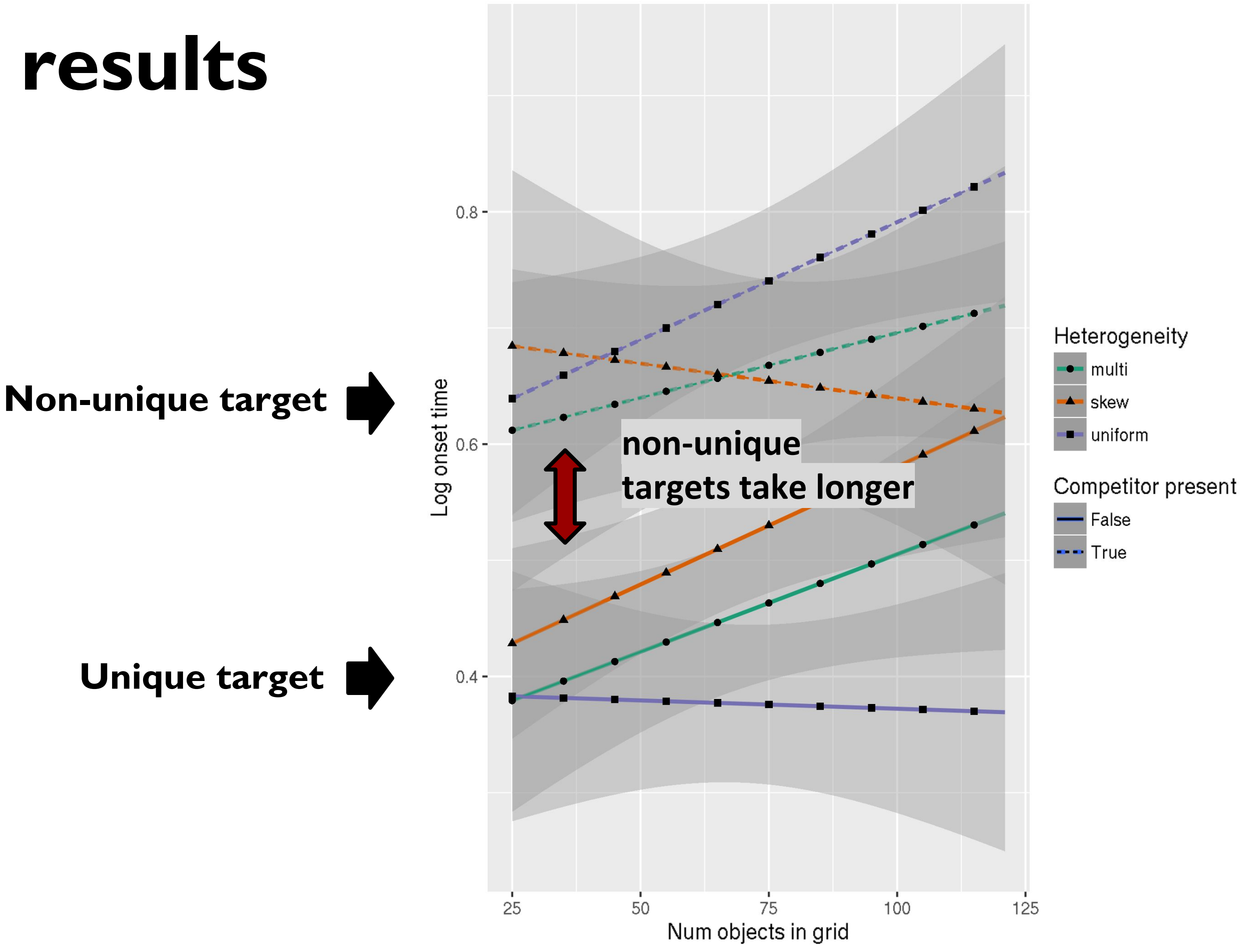
Large uniform scene ⇒ coordinates

Coordinates (visually difficult to compute) as a fallback

Timing results



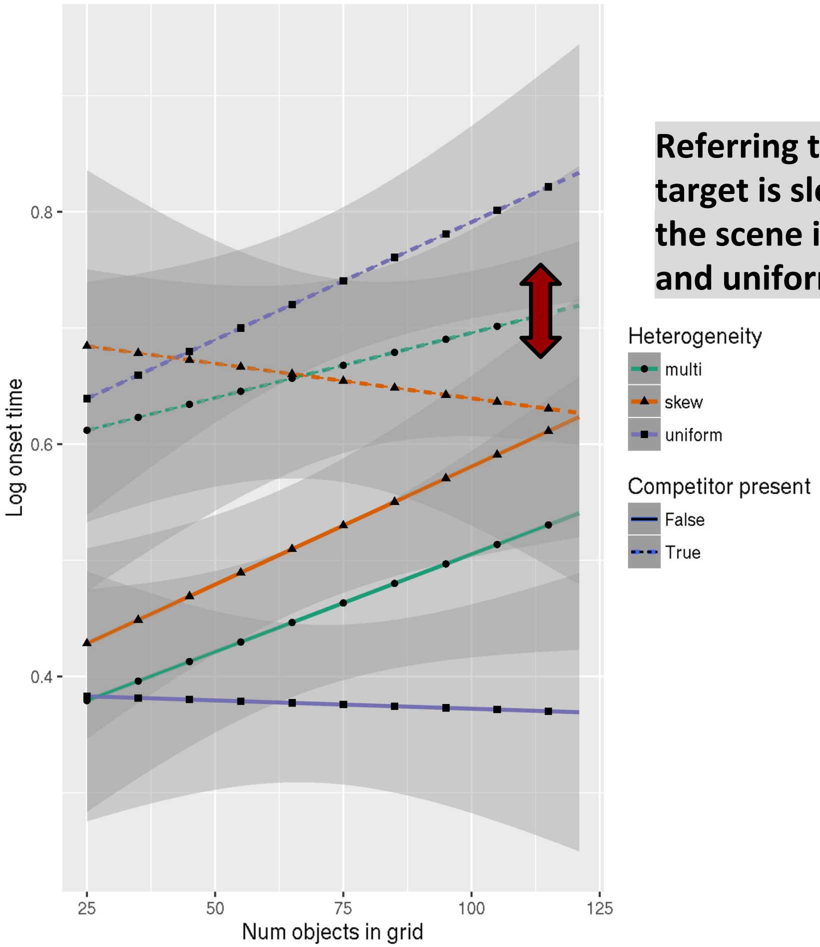
Timing results



Timing results

Non-unique target

Unique target

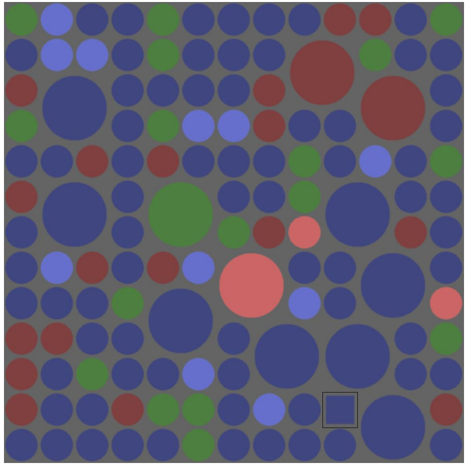


Referring to the target is slow if the scene is large and uniform

Initial perception guides strategy choice

Non-unique targets require more explanation; speakers know this.

Skewed scenes are visually hard to parse
But enable **quick** linguistic strategies to screen out most of the chaos



11x11 skewed distr.
unique target

When things go wrong...

Speakers can miss an important detail and make a bad plan…

Early observation (Pechmann 1989): Speakers sometimes produce mis-ordered adjectives:

"red big square"

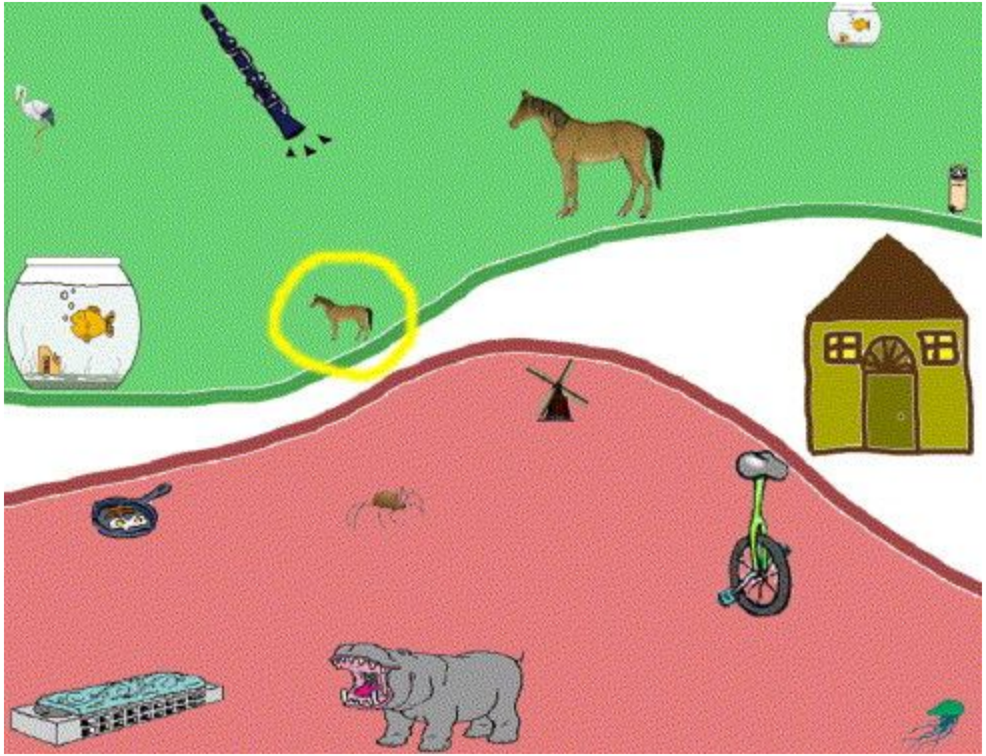
Vision as a source of speech errors

Is this a "small horse" or a "horse"?

When would you expect one vs the other?

"Watching the eyes when talking about size: An investigation of message formulation and utterance planning"

Brown-Schmidt and Tanenhaus 2006

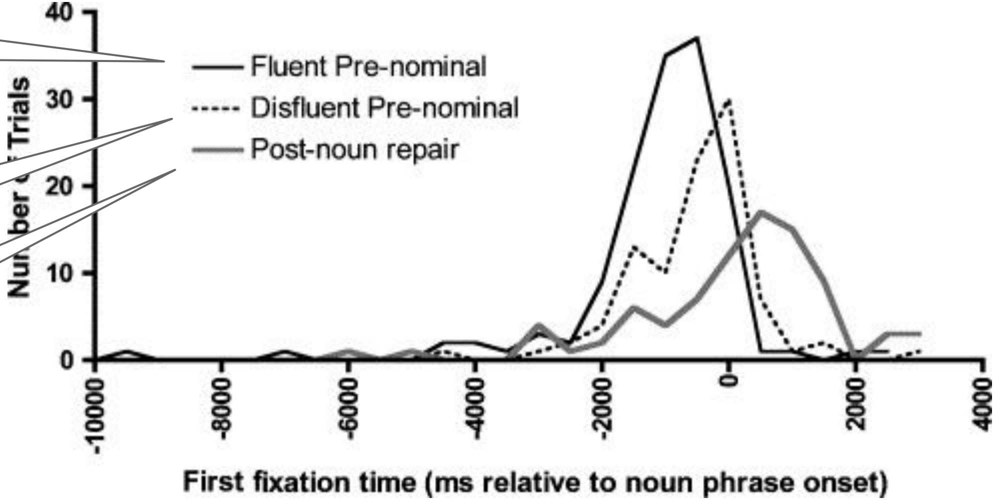


Eye track: first look at large horse

“the small horse”

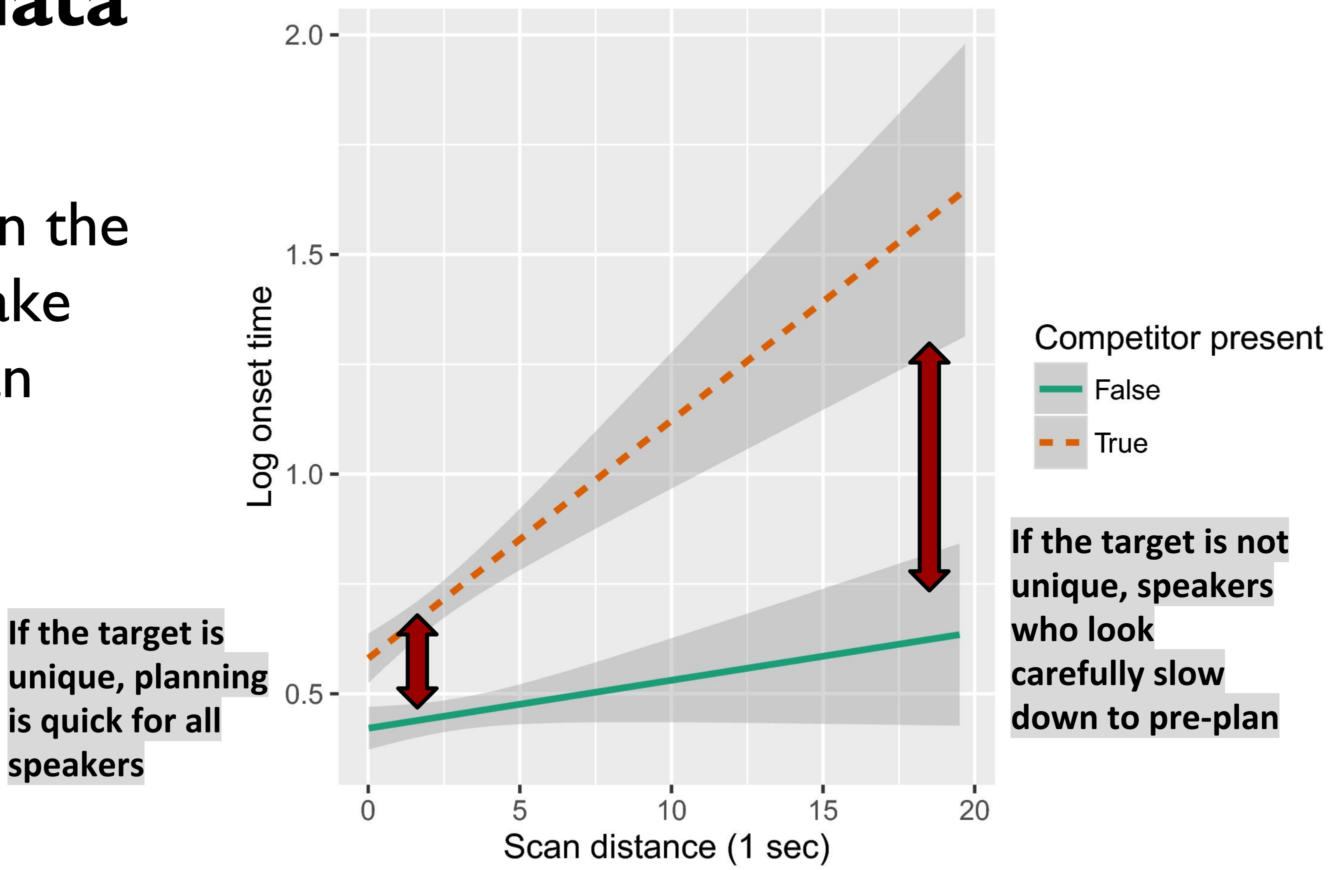
“the… uh... small horse”

“the horse… uh… small one”



In our own data

Speakers who scan the scene carefully make different plans than careless ones



Language and processing

When scenes are sufficiently complex, speakers use a variety of strategies to balance:

How much they're going to need to say
How hard they have to look at the image

Time pressure sometimes causes an error/revision

Open question: Why variability?

Do people vary so much because of:

individual-level factors

How good they are at visual search?

How good they are at speech planning?

circumstantial factors

What they looked at first?

Entrainment to a strategy?

Can we predict / explain speech errors?

Experiments are ongoing

We’re working on individual variation now…

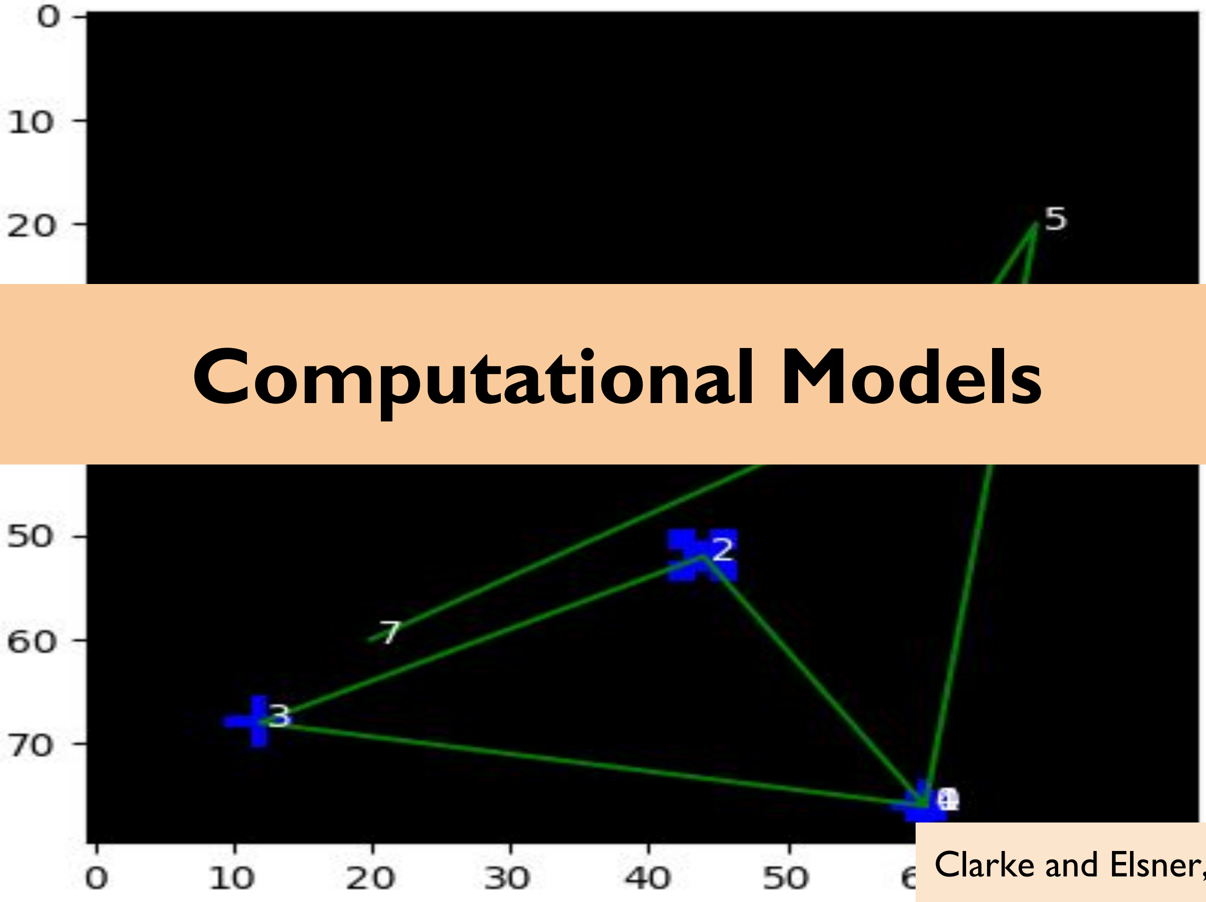
But we’re also trying to model the planning process

With a good model, we could:

Check what utterances are likely given a gaze path

Measure visual costs of different speech patterns

Computational Models

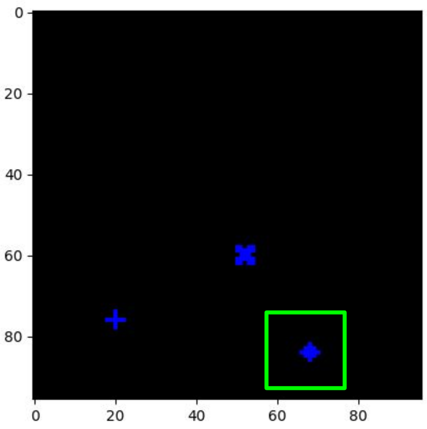


Clarke and Elsner, in submission

We are starting simple

Synthetic data: model learns to imitate a deterministic strategy.

The teacher doesn’t use vision, but the model does.



Teacher says:

UNIQUE BLUE DIAMOND </s>

Model overview

The model has focal and peripheral vision

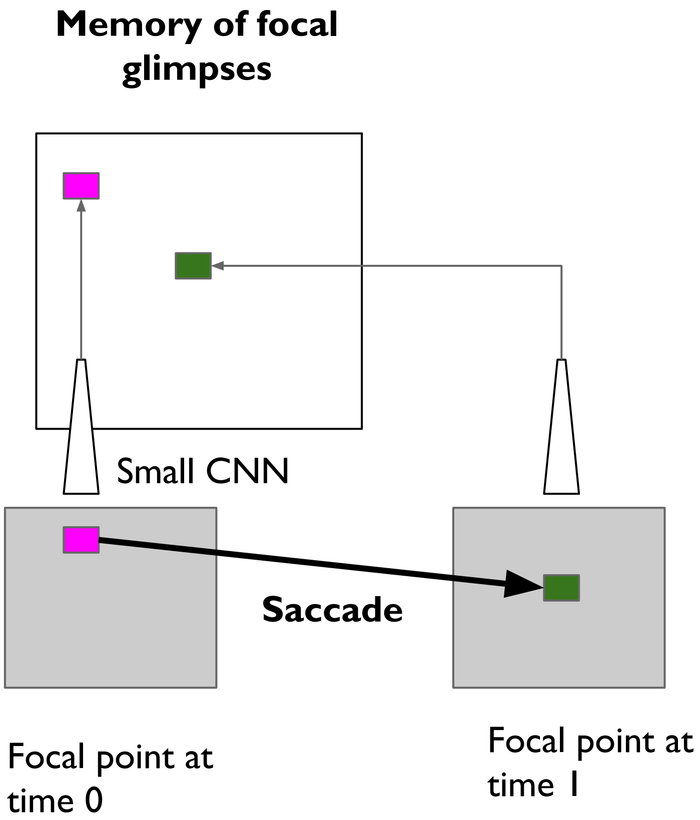
At every step, it moves the focus point…

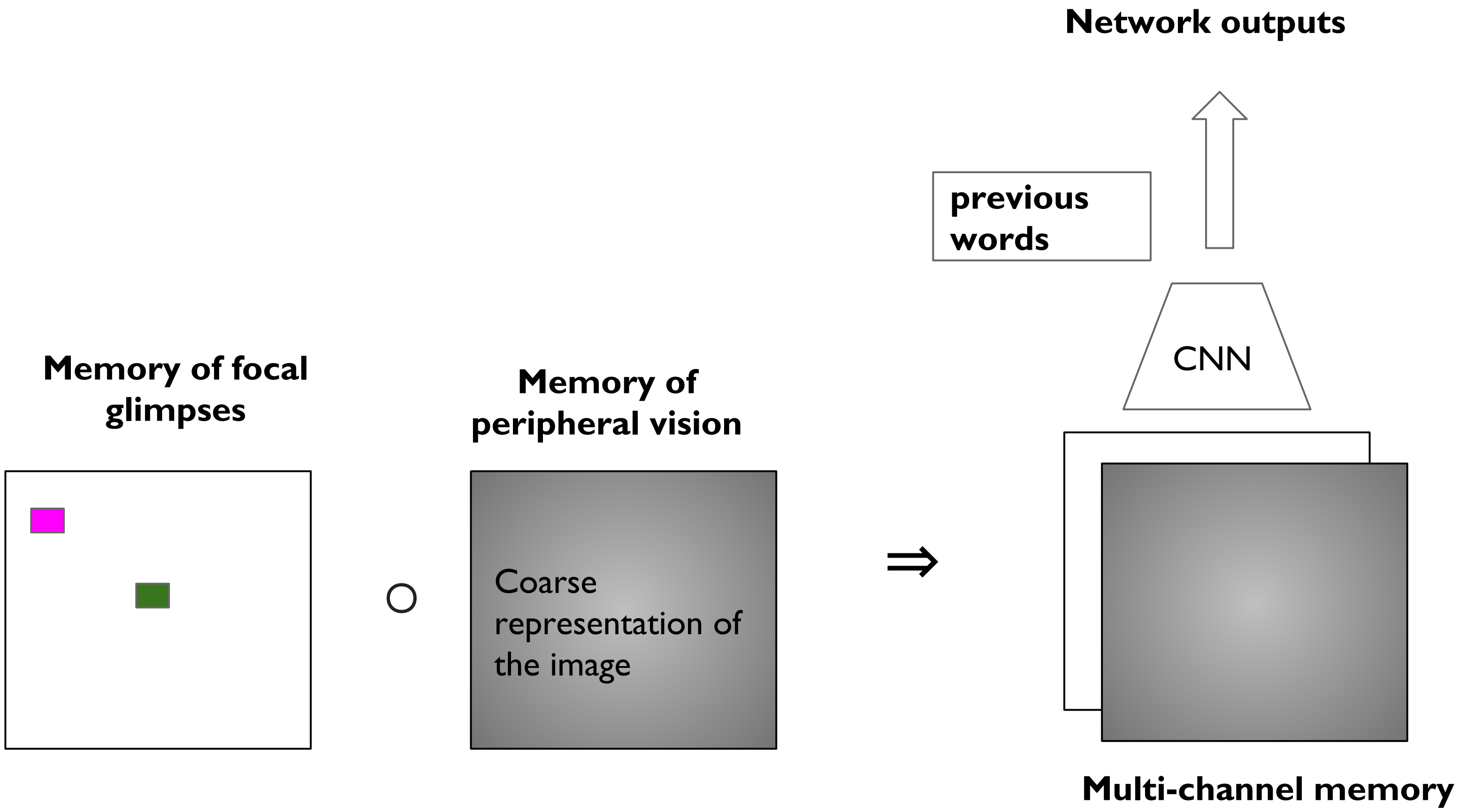
And then decides whether to utter a word…

And then which word to say

Retinotopic memory

When the model makes a fixation, what it sees is stored in a memory array, which is shaped like the image




Network outputs
previous words
CNN
Memory of focal glimpses
Memory of peripheral vision
Coarse representation of the image
○
⇒
Multi-channel memory

Training procedure

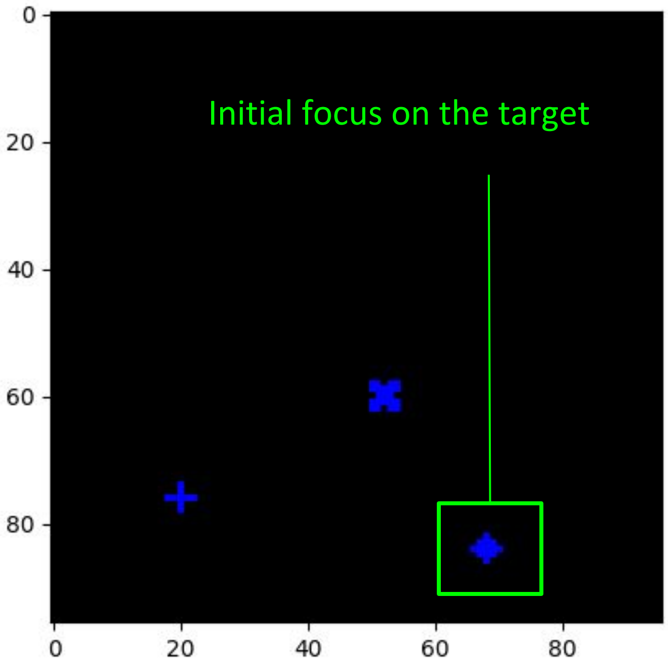
Decisions about **where to look** and **whether to speak or not** trained using deep Q-learning (reinforcement)

Positive reward for the right word, negative reward and trial halt for the wrong word, slight negative for pausing.

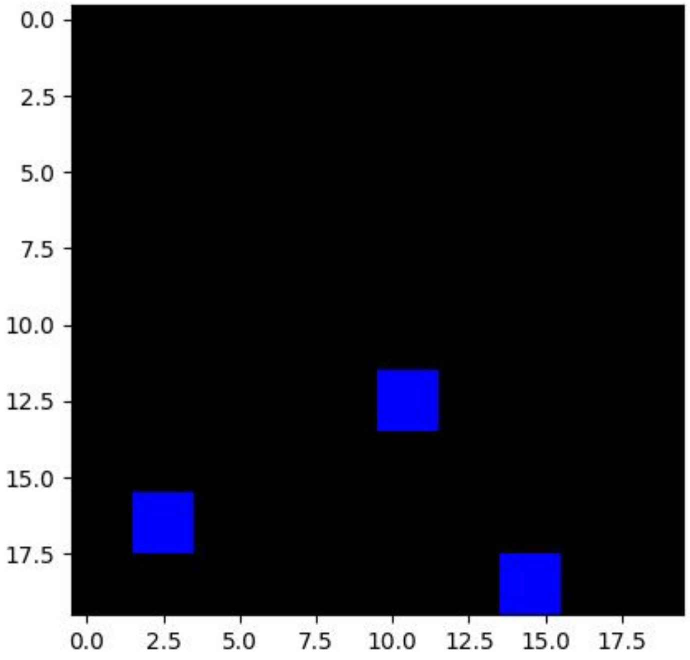
Decisions about **what word to say** (conditioned on whether to speak) trained using conventional max-likelihood

Experiment 1: replicating Gatt

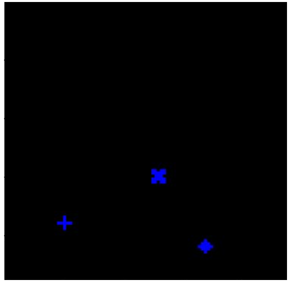
The actual scene



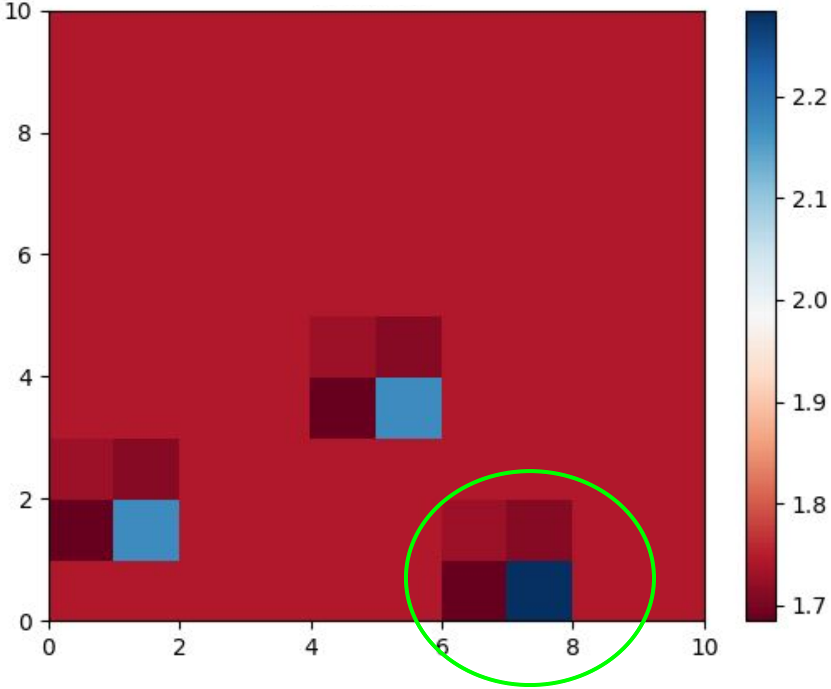
Simulated peripheral vision



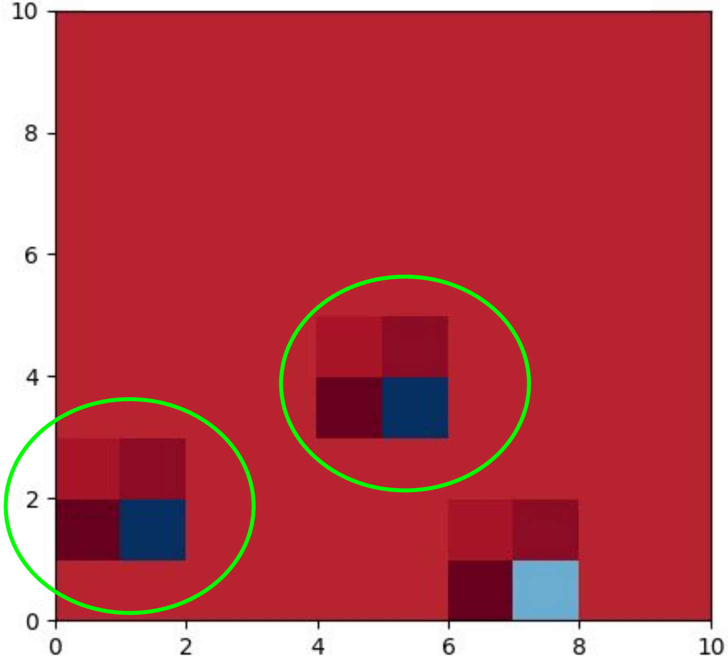
Learning where to look



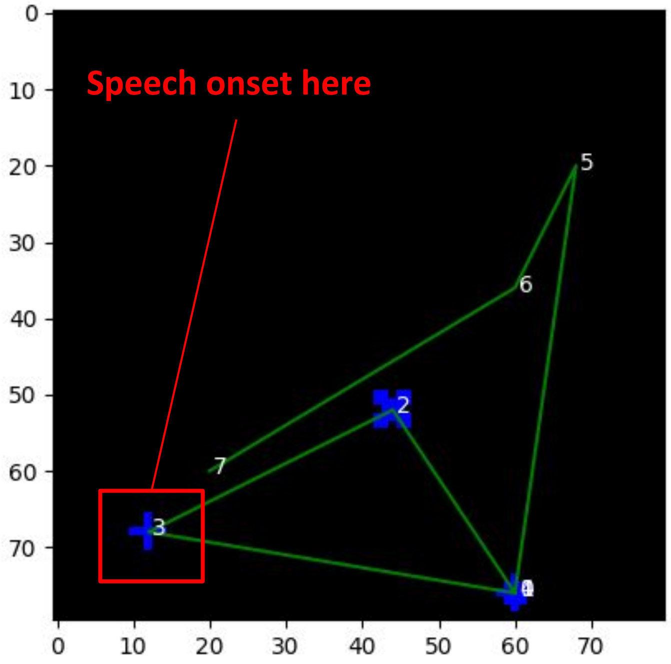
Where does the model want to look first?



What will it do next?



The outcome

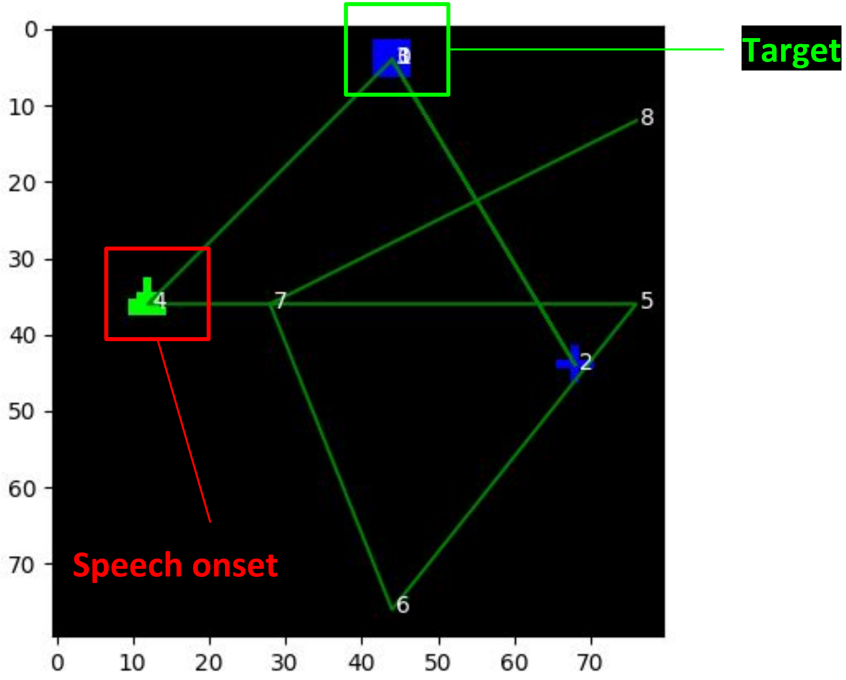


Caption: `unique blue diamond </s>`

What if the image is multicolor?

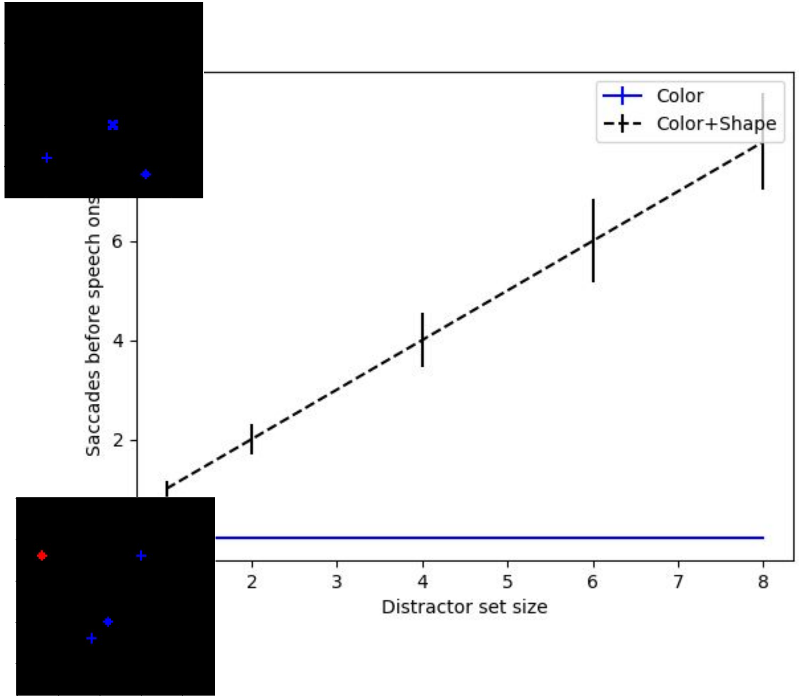
The system has learned to look at the other blue shape first…

(It still saccades to the green one afterwards.)

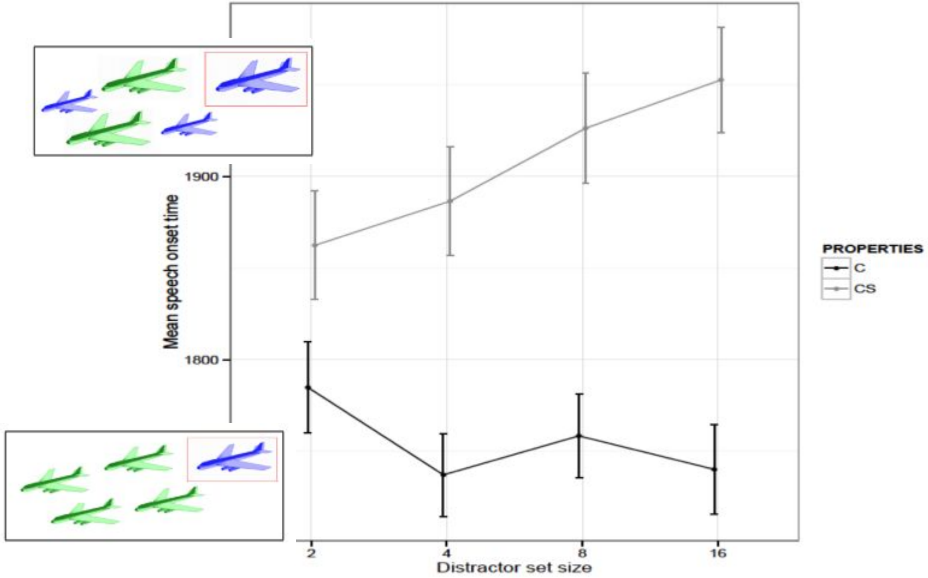


Onset times

Model onset times (saccades)

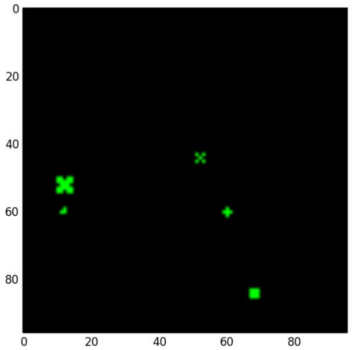


Gatt's onset times (milliseconds)



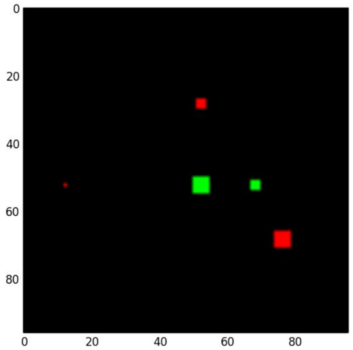
Experiment 2: disfluencies

Add size contrast; hierarchy of color > shape > size



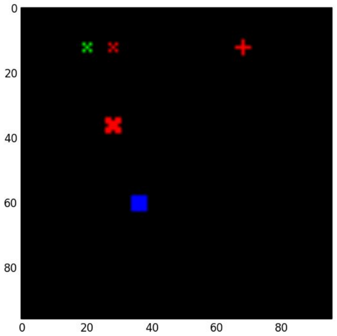
Size/shape:

“small x”



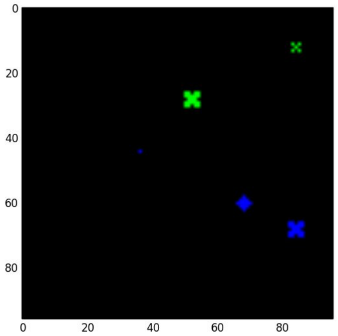
Color/size/shape:

“small green square”



Color:

“green one”



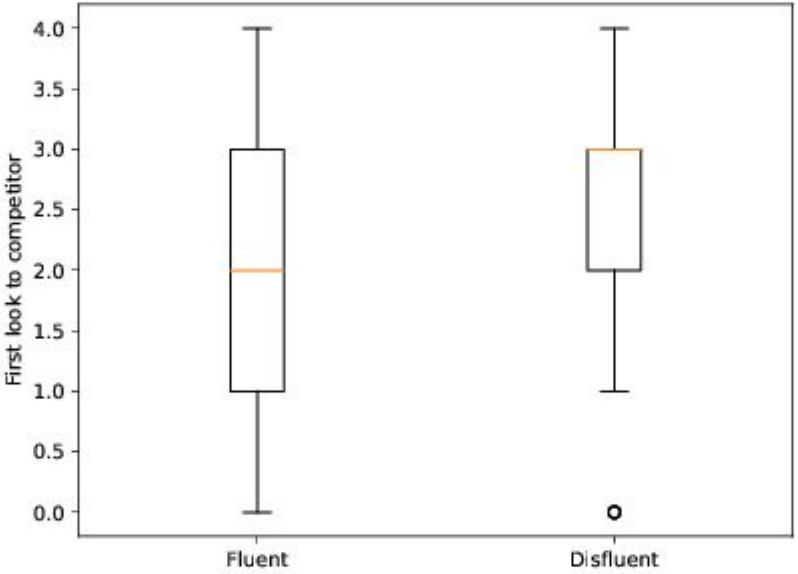
Color/shape:

“blue x”

Simulated disfluencies

Productions like
"red small square"

result when the system sees the other square too late to adjust.



"small red square" "red small square"

Future work

We hope to improve the vision section of this model

In order to make predictions on photorealistic stimuli

And analyze human gaze data

Conclusion

Pragmatics involves compromise between optimal design principles, and costs of various cognitive resources

Speakers reason intelligently about these costs

- Not just vision, but memory, lexical retrieval, …
- This creates complex planning problems and rich linguistic strategies for description

The language-vision interface

Understanding this process can help to improve virtual direction-giving and descriptive programs

Reveal details of human sentence planning

Delimit the boundaries of neo-Gricean theories for reference

Thank you!

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

- Large objects prefer short definites over indefinites
- More definites for objects far from the target
- Fewer definites in crowded images