NLG as Cognitive Modelling

The case of Referring Expressions Generation

Kees van Deemter
University of Aberdeen
Computing Science dept.

Main message of this lecture

- NLG can be a tool for achieving a better understanding of
  - Language
  - Human language production
- Example: Referring Expressions Generation (REG)
  - Probably the most widely studied area of NLG
  - (RefNet 2013: an entire Summer School devoted to the generation/production of Referring Expressions)

Plan of the lecture

1. Reviewing the goals of NLG
2. Goals of Computational Cognitive Modelling
3. Recap of REG
4. REG as Cognitive Modelling: examples
5. REG as Cognitive Modelling: classification
6. Implications for NLG as a whole

1. Goals of NLG (your turn)

Goals of NLG (my attempt)

a. Automatically producing useful text from non-textual input. (Cf. various lectures from Arria people)

   -- Useful: defined in terms of utility for users

   -- Speeding up understanding/decisions based on "data" (compare Readability course at this summer school)

   -- Improving the quality of enjoyment/understanding/decisions

Implications of this view

- The most useful output may be unlike any human utterance
  - Controlled Natural Language? (e.g., no anaphora)
  - Graphics, multimedia, etc.
- In the end, this enterprise may no longer have much to do with natural language
  - A (highly useful) artificial language?
  - Special lexicons, grammars, etc.

Goals of NLG (my attempt)

b. Automatically producing human-like text from non-textual input. (Simulation!)

-- "Human-like": similar to corpus

-- Possibly the most frequently employed evaluation method in NLG

-- But why? Is generating human-like utterances a goal in its own right?

Goals of NLG (my turn)

Maybe the two aims (**a** and **b** above) co-incide?

Maybe human-like utterances are easy to process by hearers/readers

Evidence that this may sometimes be the case:

Campana et al. (2011) *Natural Language Engineering* ***17*** *(3), p. 311-329*

Goals of NLG (my turn)

~~Maybe the two aims co-incide~~

They do not always co-incide!

“Egocentricity” results in psycholinguistics

W.S.Horton & B.Keysar (1996) When do speakers take into account common ground? *Cognition* **59** p.91-117.

L.W.Lane et al. (2006) Don’t talk about pink elephants!: Speakers’ control over leaking private information during language production. *Psychological Science* **17,** p.273–277.

Egocentricity
Horton & Keysar 1996

Speakers (S) often fail to take the Hearer's (H) knowledge into account

Set-up of their experiment:

What S and H see

S and H observe different halves of a screen

S and H see a target object (which moves)

S also sees a context object c

Conditions:

Shared: H also sees c

Privileged: H does not see c.

S knows which condition S and H are in

What S might say

S describes the target object to H, e.g.

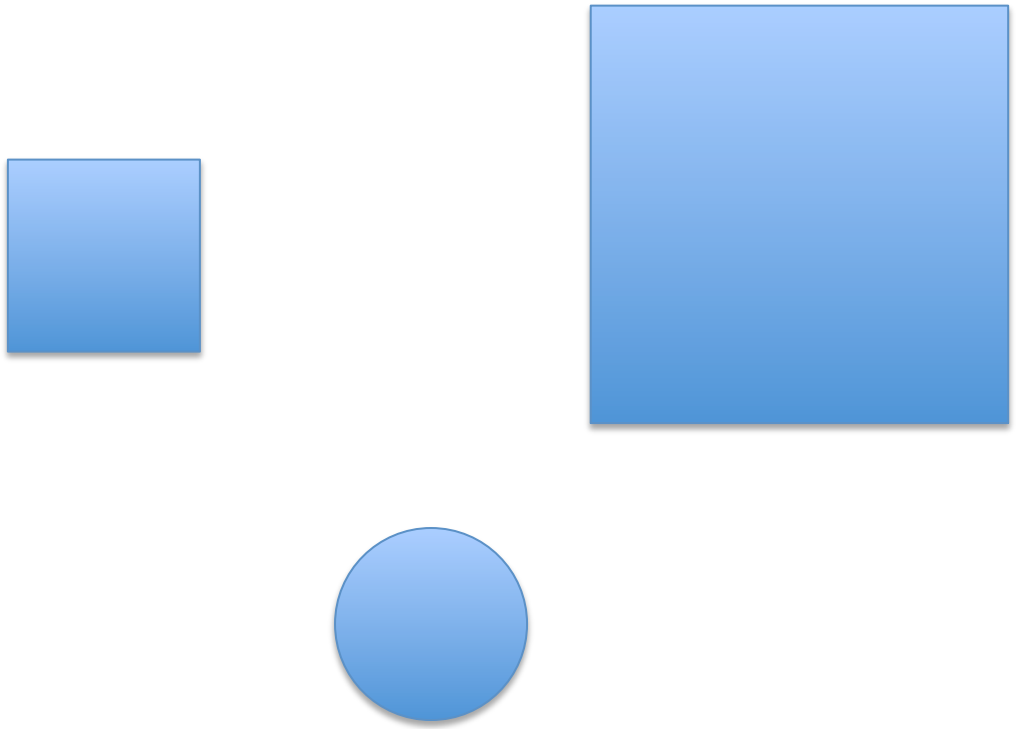
"the small square"

Note: Degree adjectives (like "small") only have meaning to H if H can see a comparison object.

Only in the shared condition!

The essence of the situation (simplified):

Shared (“the small square”)



Privileged (“the small square”??)

Number of degree adjectives used by S

as a fraction of the number of words in the NP

- **Unspeeded**: 29% (shared), 9% (privileged)

  The difference was significant
- **Speeded**: 19% (shared), 18% (privileged)
  The difference was not significant

  Essentially: <u>speeded speakers did not distinguish between shared and privileged info!</u>

Summing up this part of the talk

NLG can be performed with two different goals in mind

a. Delivering benefits for hearers

b. Simulating speakers

2. Computational Cognitive Modelling

An entirely different research area

See e.g. R. Sun (Ed.) 2008 *The Cambridge Handbook of Computational Psychology*.

(With contributions from J.McClelland, Ph.Johnson-Laird, W.Gray, M.Boden, A.Sloman, etc.)

Models of Cognition

Aim to describe/explain an aspect of human cognition

Can be

- Verbal-conceptual [still most frequent?]
- Mathematical
- Computational

Computational Models of Cognition

Examples from Sun (2008): Models of

- Human memory
- Visual information processing
- Logical reasoning
- Inductive reasoning
- Decision making
- Game playing
- Human (and animal!) learning
- Speaking

Differences between models
Example: logical reasoning

- **Aim** (logically valid reasoning, or with human flaws? E.g. Johnson-Laird; Kahneman & Twersky)
- **Granularity** (Propositional? First-order? Modal?)
- **Physiological basis?** (Some models of human reasoning are inspired by neuro-science, e.g. neural nets)
- **Product or process?** (Only what conclusions are drawn, or also how quickly?)
- **Individual or groups?** (How do group processes affect validity & speed of reasoning?)

3. A brief recap of Referring Expressions Generation (cf., Albert Gatt's lectures)

1. Something about algorithms
2. Something about evaluation (TUNA)

Abbreviations:

RE = Referring Expression

REG = Referring Expressions Generation

The "classic" algorithms

Shared KB is a set of properties, e.g., Desk, Red,..

An RE expresses a conjunction of properties

"Monotonic" Algorithms add properties one by one

- Greedy Algorithm: starting with the most discriminating one
- Incremental Algorithm: following a fixed **Preference Order** of properties (Dale & Reiter 1995)

Monotonic approaches to REG

Let's use informal pseudo-code, where

M : domain of elements

D : description under construction

P : set of available properties

Monotonic REG

D := ∅

While not all distractors have been ruled out

and **P** ≠ ∅ do

Select new P from **P**

If P is false of some distractors then

Add P to **D**

Remove P from **P**

Remove from **M** all distractors ruled out by P

The monotonic approach to REG

Using different methods

D := ∅

While not all distractors have been ruled out and **P** ≠ ∅ do

Select new P from **P**

If P is false of some distractors then

"Update **D**, **P** and **M**"

Add P to **D**

Remove P from **P**

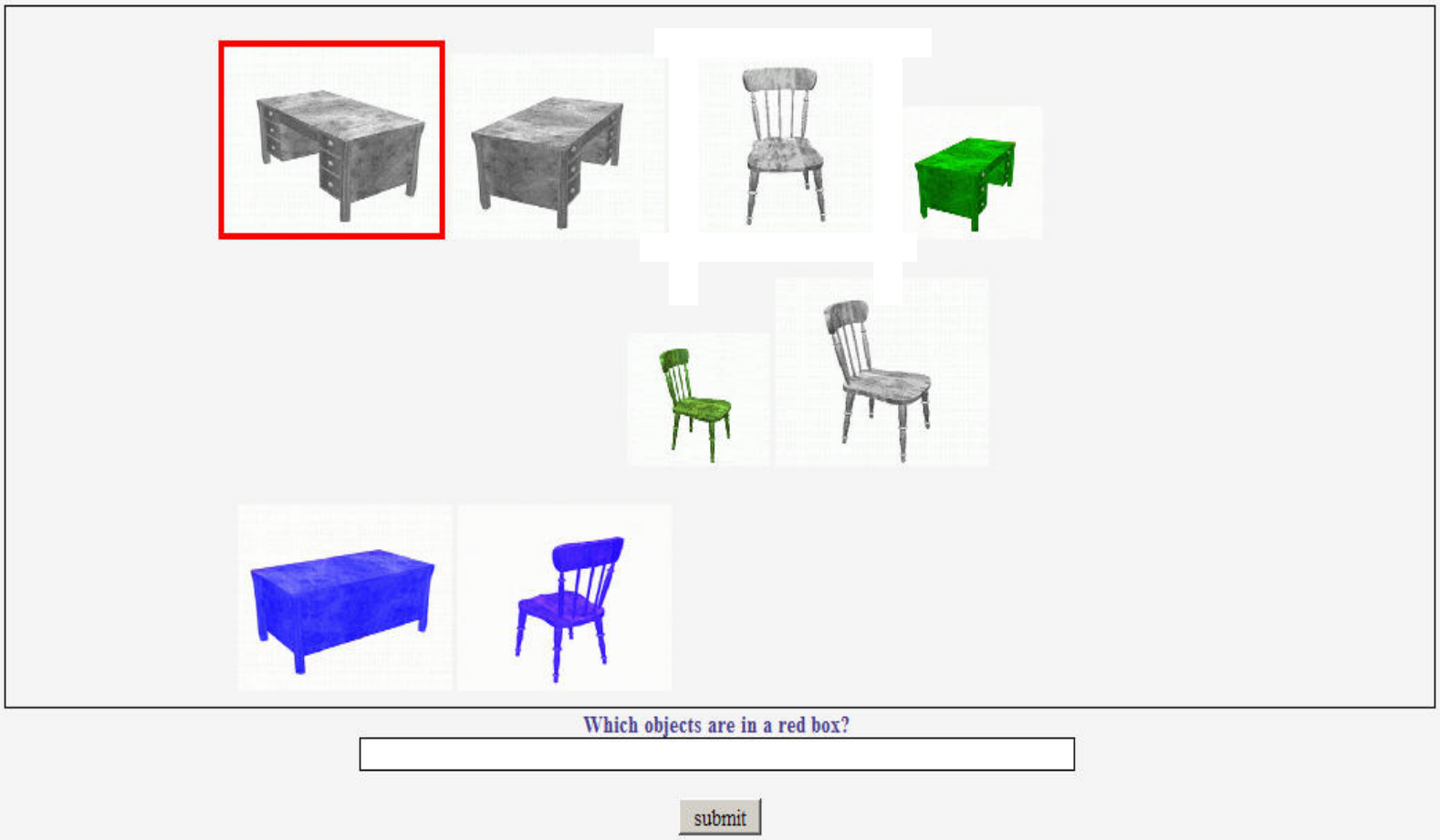
Remove from **M** all distractors ruled out by P

Evaluation of these algorithms

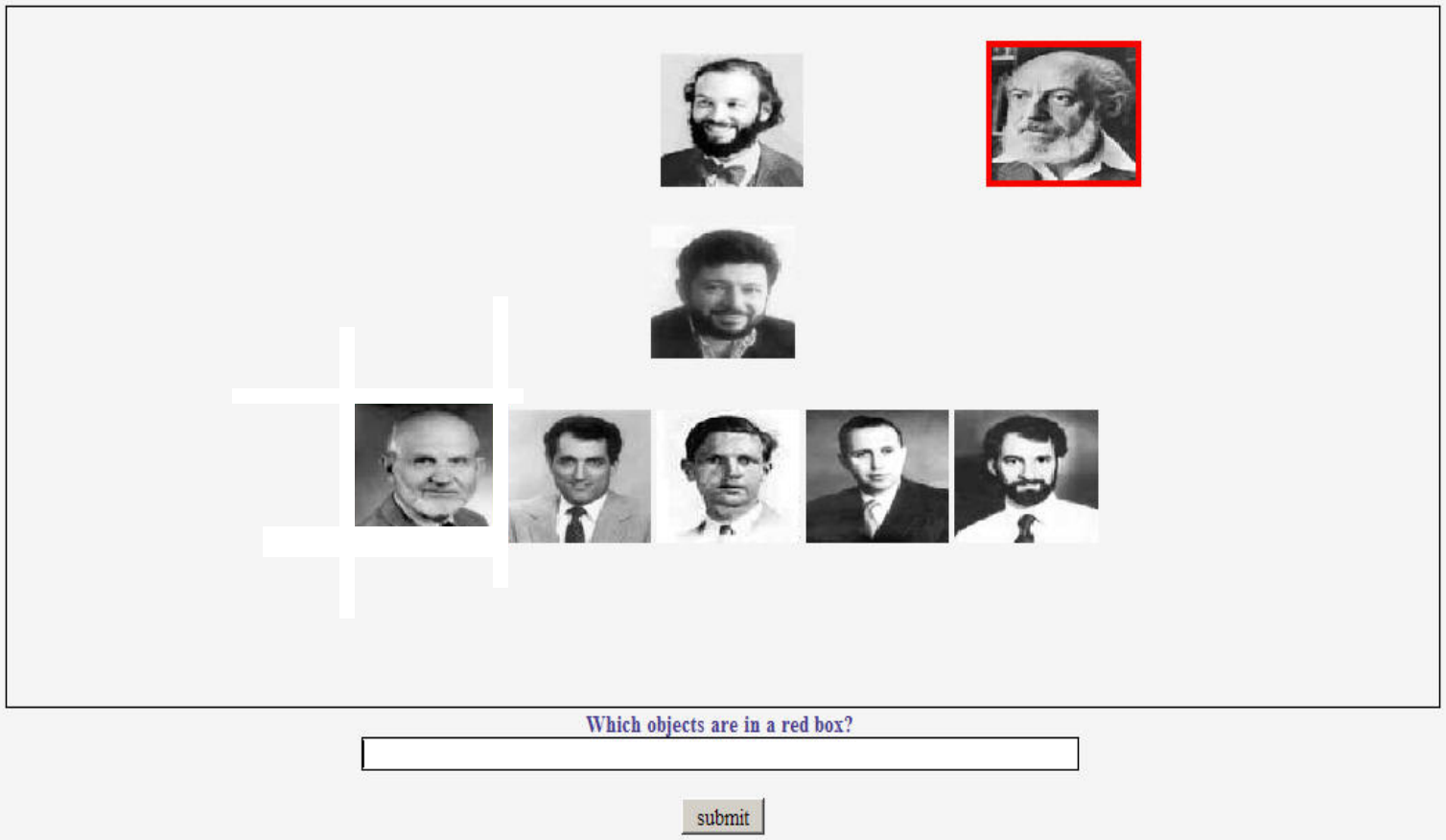
E.g., TUNA (Brighton-Aberdeen, 2006):

- Experiment: REs elicited under controlled circumstances
- These human-produced REs are compared with REs generated by algorithms:
  - Give each algorithm the same input as subjects
  - Compare algorithm's output to subjects' output
  - Count semantic content only

TUNA: a Furniture trial



TUNA: a People trial



Main evaluation metric

The Dice metric:

$$\frac{2 \times |\text{Common properties}|}{|\text{total properties}|}$$

Corpus: {A,B,C,D}

Algorithm: {B,C,D,E} ➔ Dice = (2*3)/8 = ¾

Dice score of 0 is awful, 1 is perfect

Alg_1 beats Alg_2 iff $Dice(Alg_1) > Dice(Alg_2)$

Details of the TUNA experiment:

Van Deemter, Gatt, van der Sluis, and Power (2012) "Generation of referring expressions: assessing the incremental algorithm." *Cognitive Science* **36** (6)

REG evaluation challenges (open competitions):

Belz & Gatt (2010) Introducing shared talk evaluation to NLG. In Krahmer & Theune (Eds), *Empirical Methods in NLG*

REG algorithms in general:

Krahmer & van Deemter (2012) Computational Generation of Referring Expressions: a Survey. Comp. Ling. **38** (1).

4. REG as Cognitive Modelling

- Observe: TUNA/Dice treated algorithms as simulations, not in terms of their utility
- TUNA does not count the production *process* , only the *product*

Note: whether an algorithm is a Cognitive Model depends on what its aim is / how it is evaluated

Caveat

- This is not the only kind of REG evaluation
- E.g., the Direction-Giving (GIVE) challenge looked at task success (time to find referent)

Koller et al. (2010) The first challenge on generating instructions in virtual environments. In Krahmer and Theune (Eds), Empirical Methods in Natural Language Generation

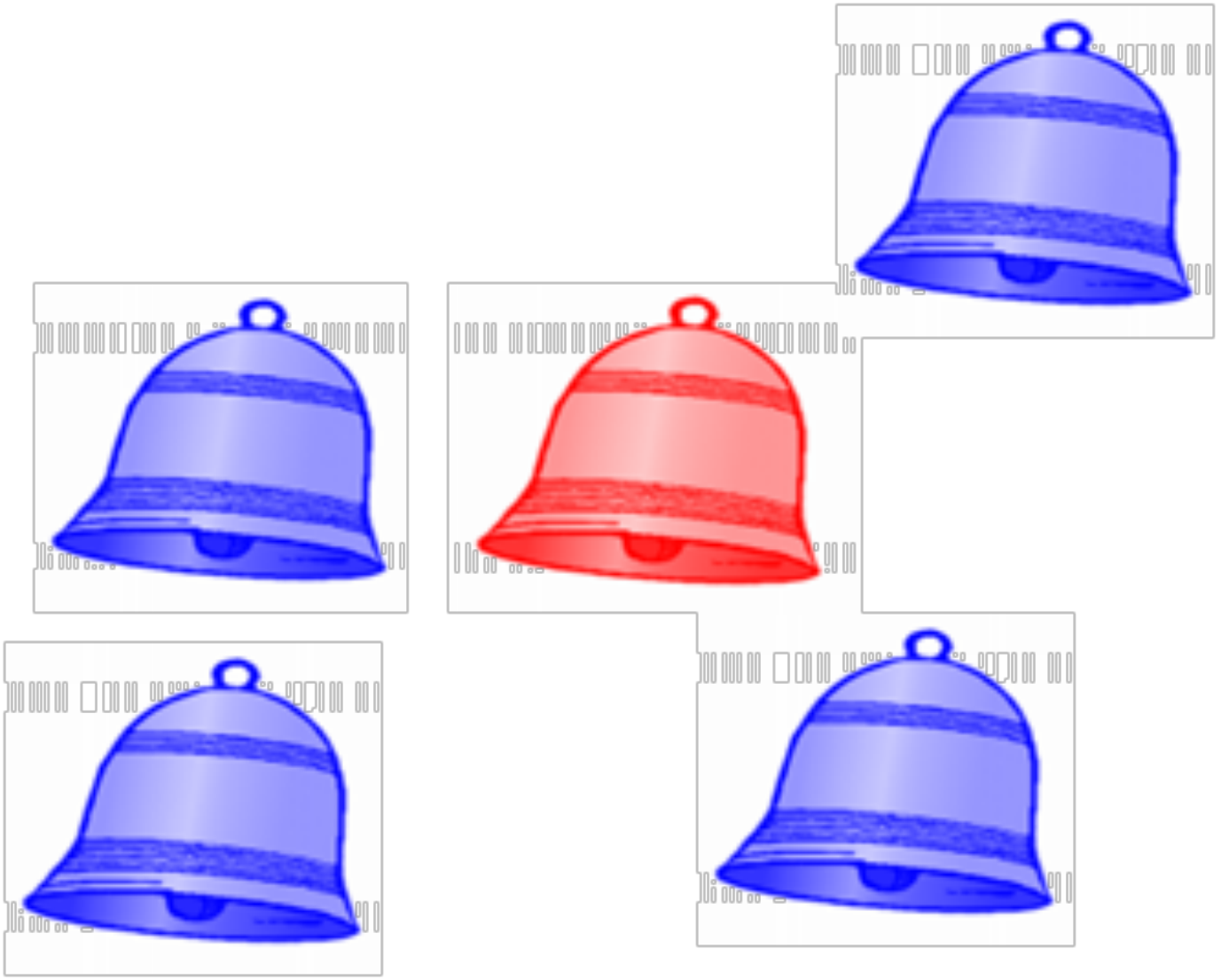
Now: a study in which the **process** is evaluated

Models of visual processing (Treisman & Gelade 1980) make predictions about visual search:

- Target can be distinguished from all distractors by using 1 property → search times do not grow with numbers of distractors [Pop-out effect]
- Target can only be distinguished from all distractors by using 2 properties → search times grow linearly with numbers of distractors [No pop-out effect]

two situations where the referent “pops out”

1. “The red bell”



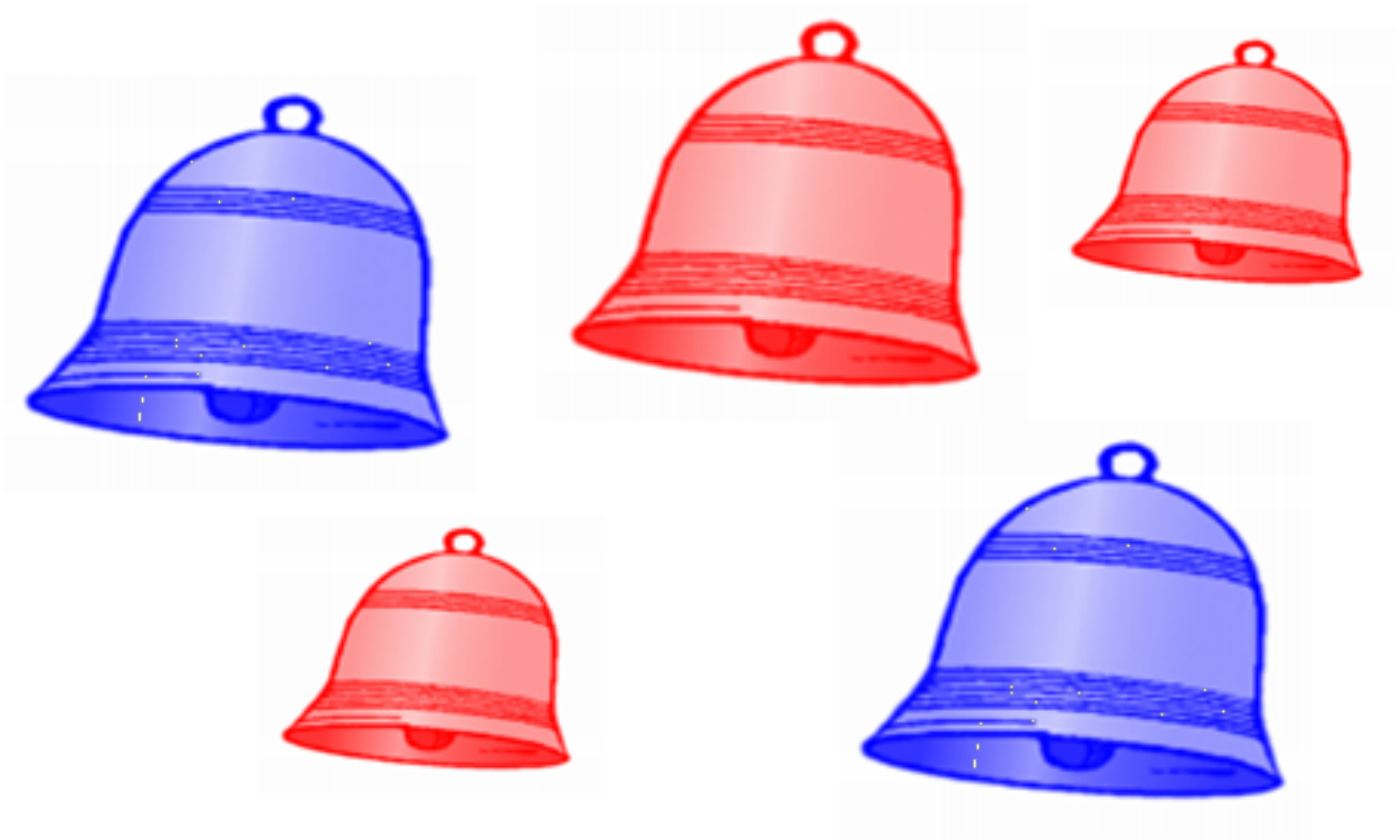
two situations where the referent “pops out”

2. “The large bell”

No pop-out effect

"The large red bell"

Search time increases linearly
with the number of distractors



Research question of this study

Is it the same for generation?

- You might expect YES (because the speaker needs to compare the referent along 2 dimensions)
- On the other hand, the SPEAKER doesn't have to SEARCH for the referent

  Suppose REG said "List all properties of the referent"
  This would be independent of the number of distractors!

What do REG algorithms predict?

- Recall the shape of most algorithms

Recall: The monotonic approach to REG

D := ∅

While not all distractors have been ruled out

and **P** ≠ ∅ do

Select new P from **P**

If P is false of some distractors then

Add P to **D**

Remove P from **P**

Remove from **M** all distractors ruled out by P

Predictions of the monotonic algorithmic pattern

- These algorithms were not intended as process models
- Yet they can be viewed in this way

Some predictions:

1. Production latency increases with the number of distractors
2. Production latency increases with the number of properties ending up in **D**

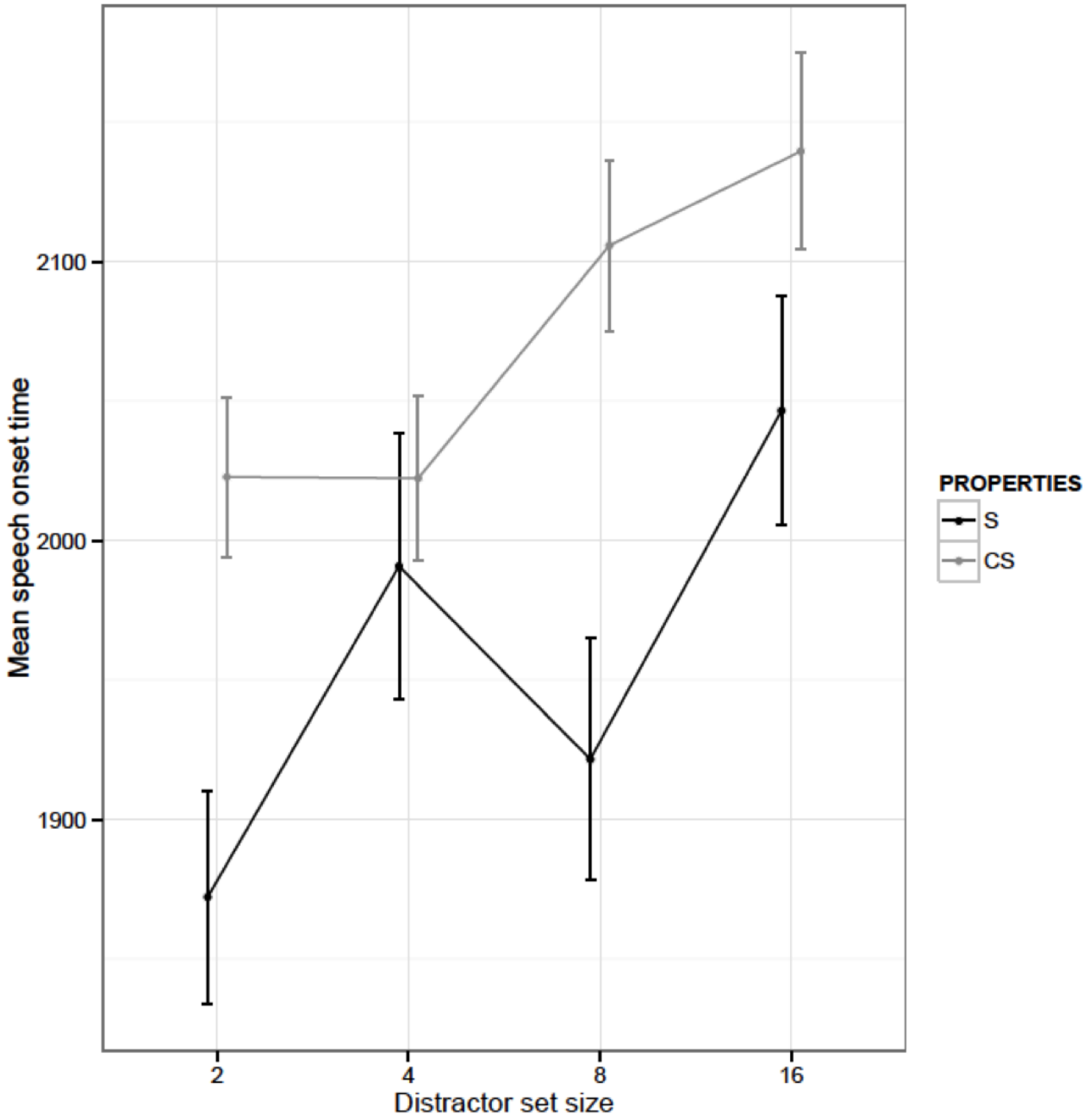
2 experiments
(only 1 experiment reported here)

Domains were varied in terms of

- the number of distractors (2,4,8,16)
- the number of properties required (1,2)

- Standardised pictures (Snodgrass & Vanderwart 1980)
- Domain elements were always of the same type
- 64 experimental items, 108 fillers
- 40 Speakers of Dutch
- Items occurred in the same order for all participants
- Participants were asked to describe items for an imaginary hearer

Results


Mean speech onset time
2100
2000
1900
2
4
8
16
Distractor set size
PROPERTIES
S
CS

Predictions of the monotonic algorithm pattern

1. Production latency increases with the number of distractors **v**
2. Production latency increases with the number of properties ending up in **D** **v**

If this is correct (and if Treisman & Gelade were also right) then production and comprehension are interestingly different

Predictions of the monotonic algorithmic pattern

If this is correct (and Treisman & Gelade were also right) then production and comprehension are interestingly different

(A nuance: Experiment 2 showed that when the referent had a different colour from all distractors, and this colour stood out sharply, then no effect of distractor set size was found.)

Summing up

This study tested the ability of a REG algorithm to describe the production **process**

Gatt, van Gompel, Krahmer, and van Deemter (2012).
Does domain size impact speech onset time during reference production? In Proc. of the 34st Annual Conference of the Cognitive Science Society (CogSci), pages 1584–1589, Sapporo.

5. REG as Cognitive Modelling: classification

Differences between models
Example: REG

- **Aim**: (most often) to simulate speakers; sometimes to benefit hearers
- **Granularity**: most often a set of properties; sometimes choice of words and syntax too
- **Physiological basis**: not taken into account yet, despite progress in neuro-science (Nieuwland & Van Berkum 2008, the Nref effect)
- **Product or process**? Usually the product (e.g., TUNA); sometimes the process (e.g. Gatt et al. 2014)
- **Individual or groups**? Usually results averaged over a group; sometimes probability distribution over a group

End of this section

- MANY more experiments have been done, testing various aspects of REG algorithms
  - Similarity to the expressions in a corpus
  - Utility for a hearer (e.g., how long does the hearer take to find the referent. Garouffi & Koller 2014; Paraboni 2014; both in *Language, Cognition, and Neuroscience* **22** (8).

6. Implications for NLG as a whole

- If an NLG project aims for utility for recipients then
  - test with recipients
  - why focus on natural language?
- If an NLG project simulates speakers then
  - test by comparing with corpora
  - simulate "bad performance" (e.g. speech errors) as well?

Implications for NLG as a whole

- REG can be regarded as a type of Computational Cognitive Modelling
- More generally, NLG can be regarded as an attempt to understand language better
- Focus can be on
  - Content Determination
  - Microplanning (Lexicalisation, REG,..)
  - Surface Realisation

(Wrapping up this lecture)
Another way to put this

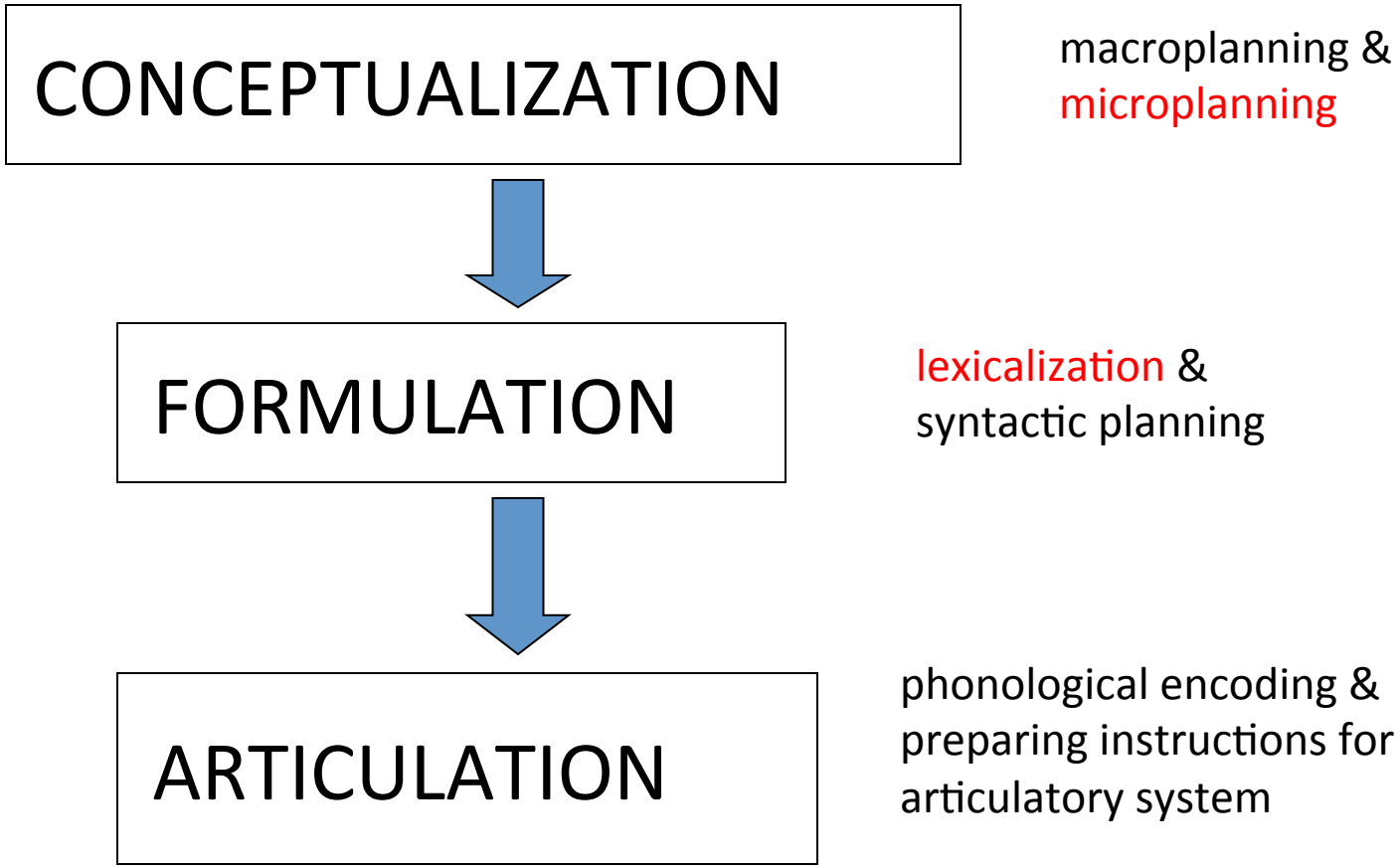
- From 1980, psycholinguists have developed models of human language production

- Perhaps the most famous model is from Levelt (1989) "Speaking: From Intention To Articulation" (chapter 12)

  More recent models include
  Dell et al. (1997), Vigliocco and Hartsuiker (2002)

Psycholinguistics

the language production pipeline (Levelt 1989)



Levelt's model in more detail

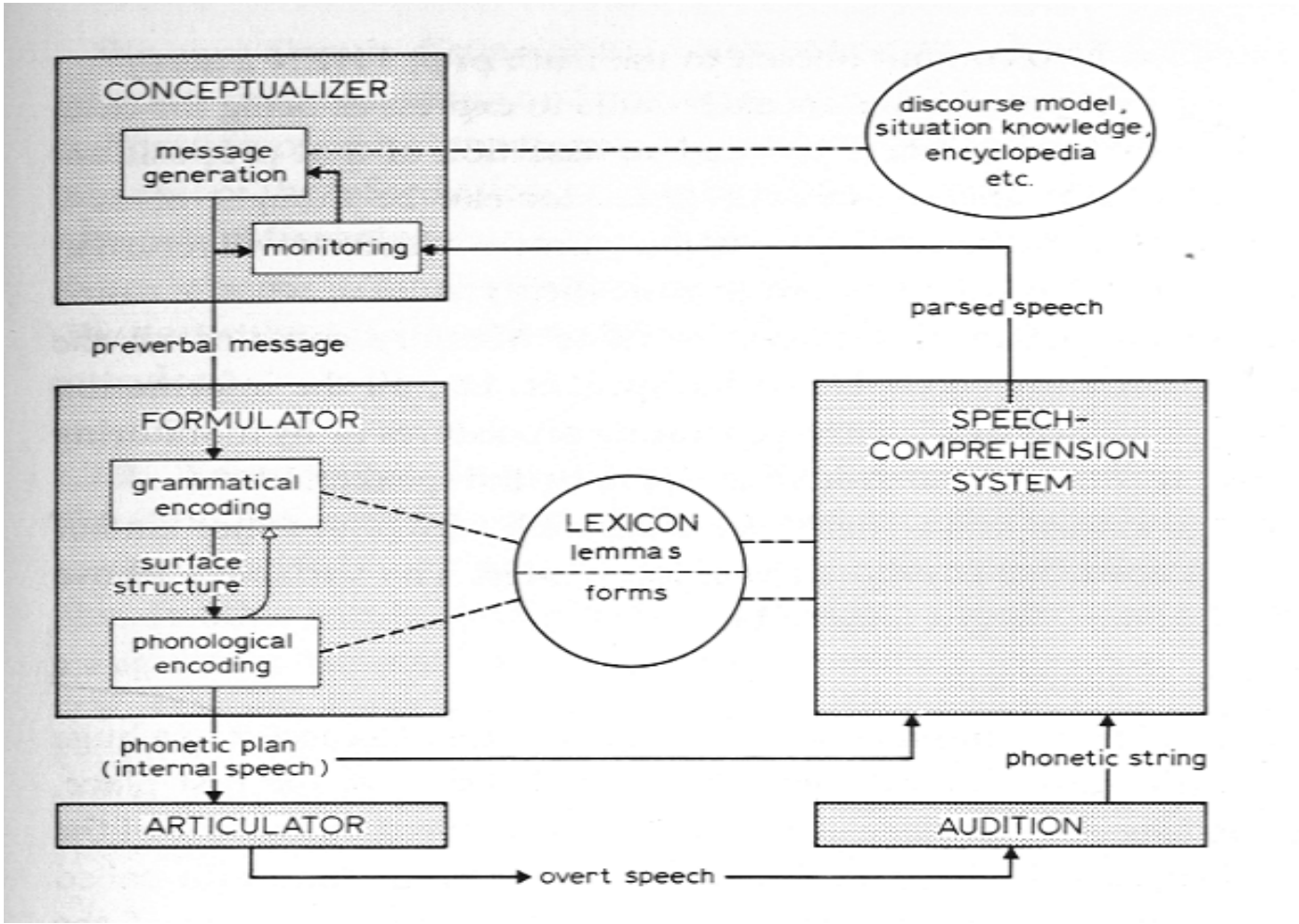
The NLG pipeline resembled this one, but looks at entire texts (> 1 sentence)

Levelt's full model is more complex because of monitoring: Speakers monitor their own

1. preverbal message (sentence plan)
2. phonetic plan
3. speech

(Evidence from speech errors & self-correction)

Language production:
Levelt 1989: Model overview



Summing up

- Algorithms and Cognitive Models have always cross-fertilised

- Algorithms and Cognitive Models have always cross-fertilised
- … and long may it last!

Postscript: Nondeterministic Models

(Roger van Gompel) Psychologist:

“Why are all your algorithms deterministic?”

Experimental Materials

Fig. 1a. Size-only fully discriminating condition

IA predicts: “the small grey candle”

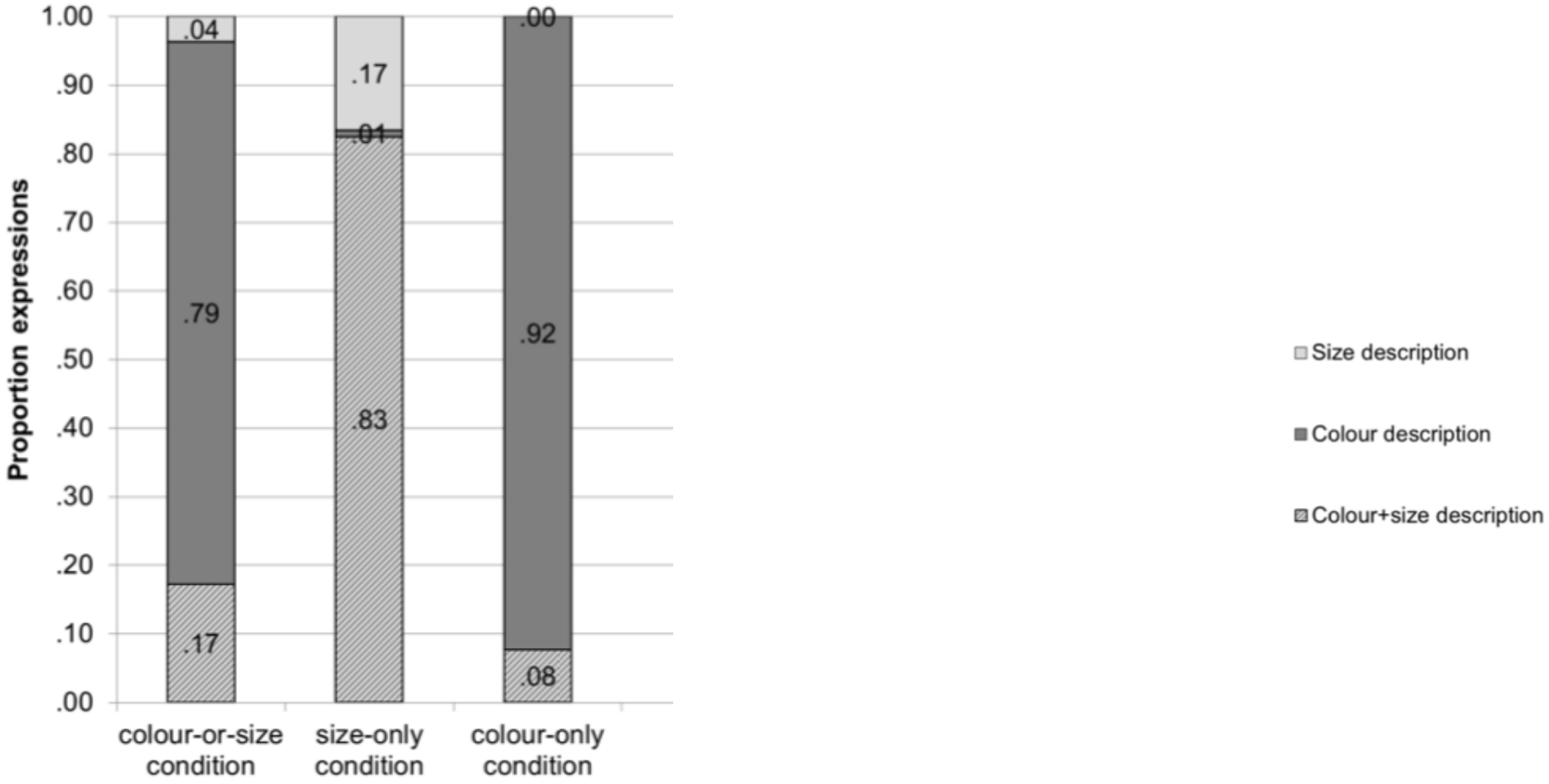
Fig. 1b. Colour-only fully discriminating condition

IA predicts: “the grey candle”

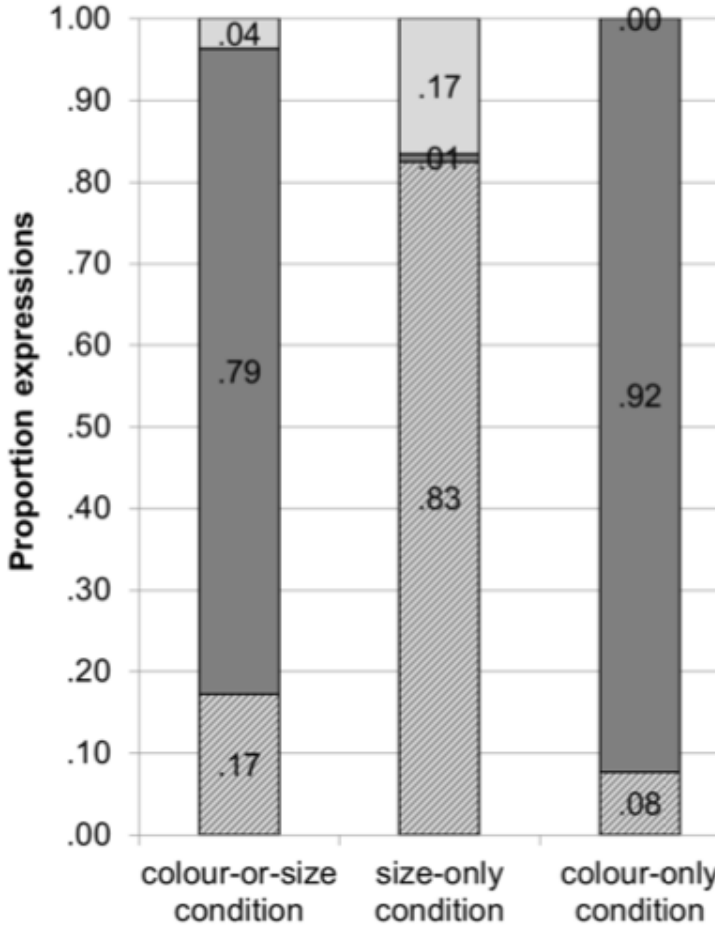
Fig. 1c. Colour-or-size fully discriminating condition

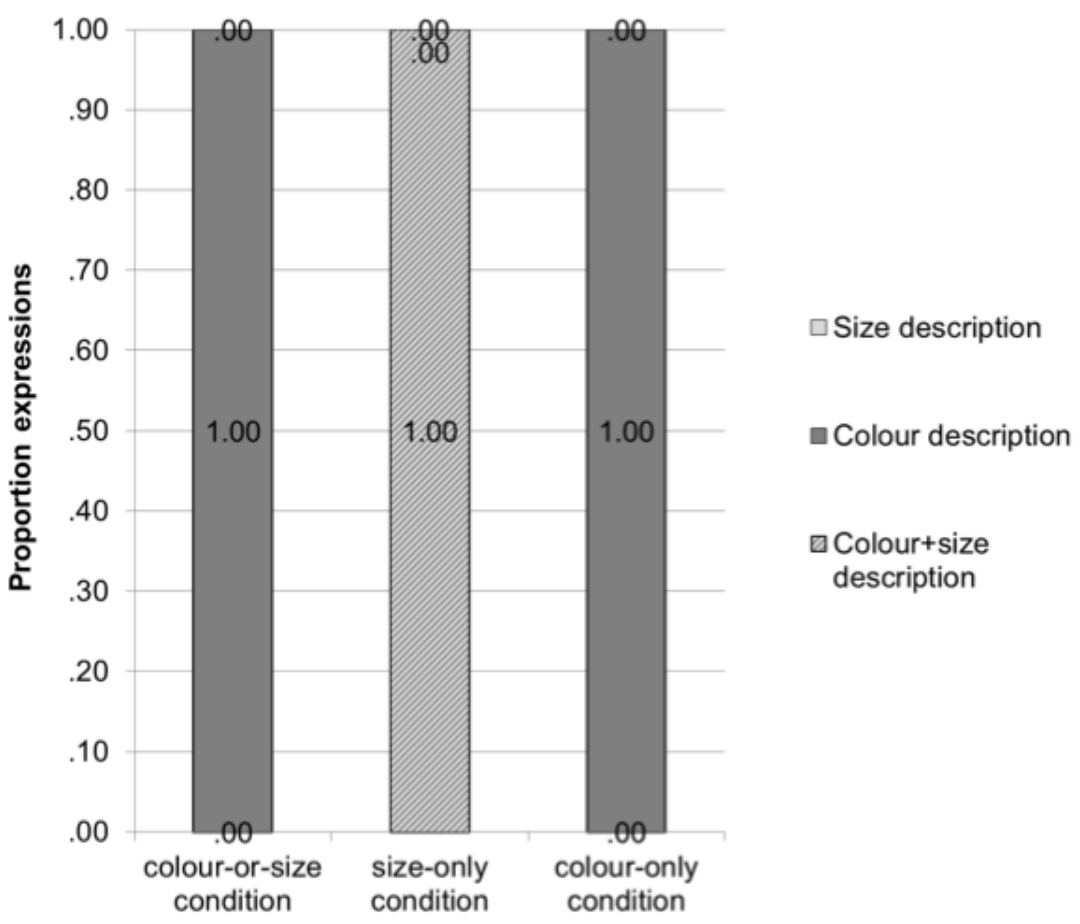
IA predicts: “the grey candle”

Human speakers



Human speakers vs Incremental Algorithm

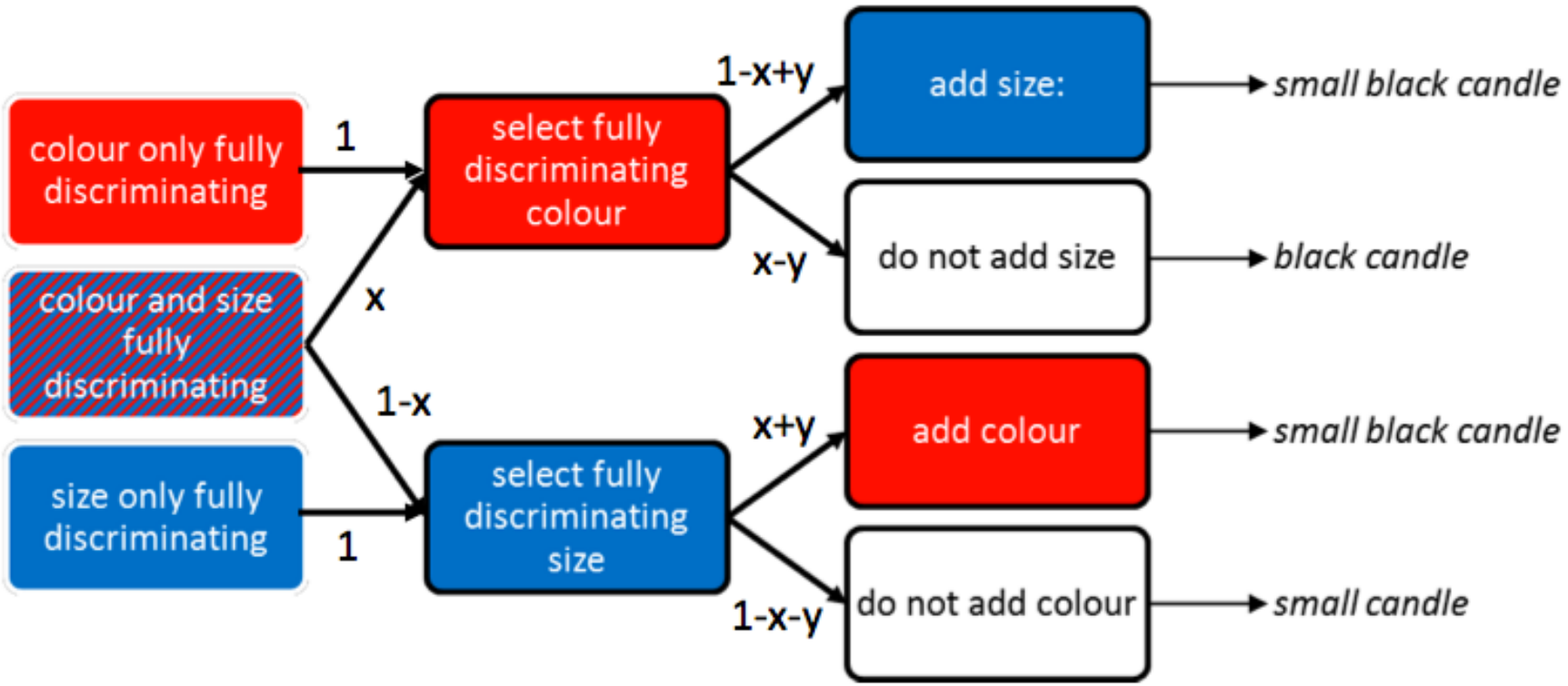




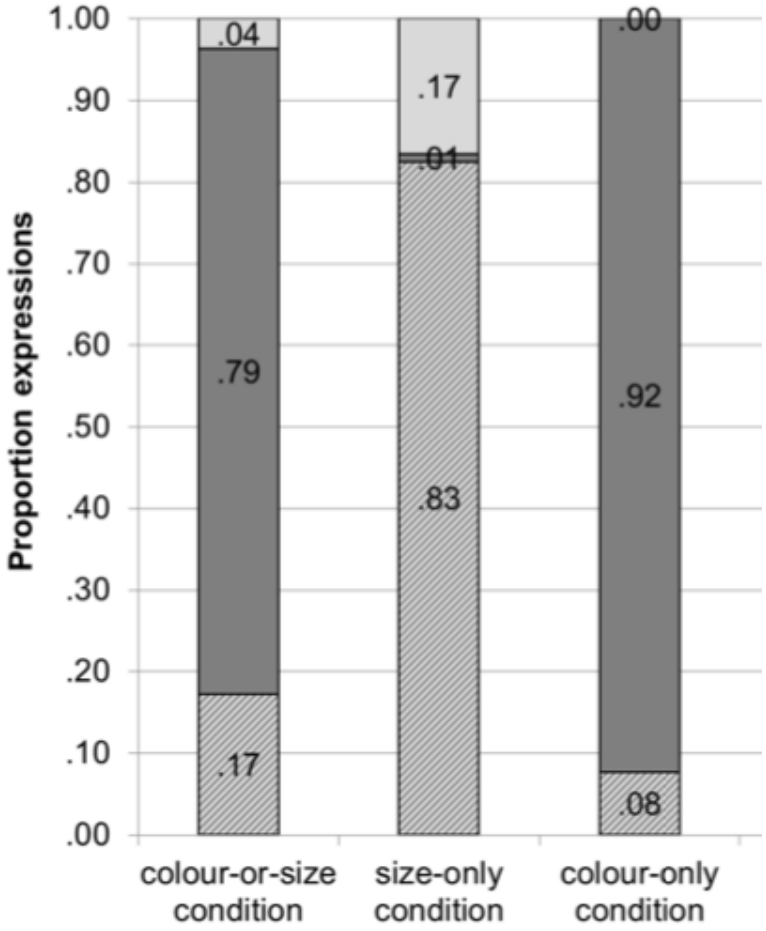
New algorithm: **P**robabilistic **R**eferential **O**verspecification (PRO)

- Select properties with probability x
  Higher x for more preferred properties
- Further properties may be added, based on parameter y for eagerness to over-specify
  Higher y for greater eagerness
- Parameters x and y estimated on held-out data

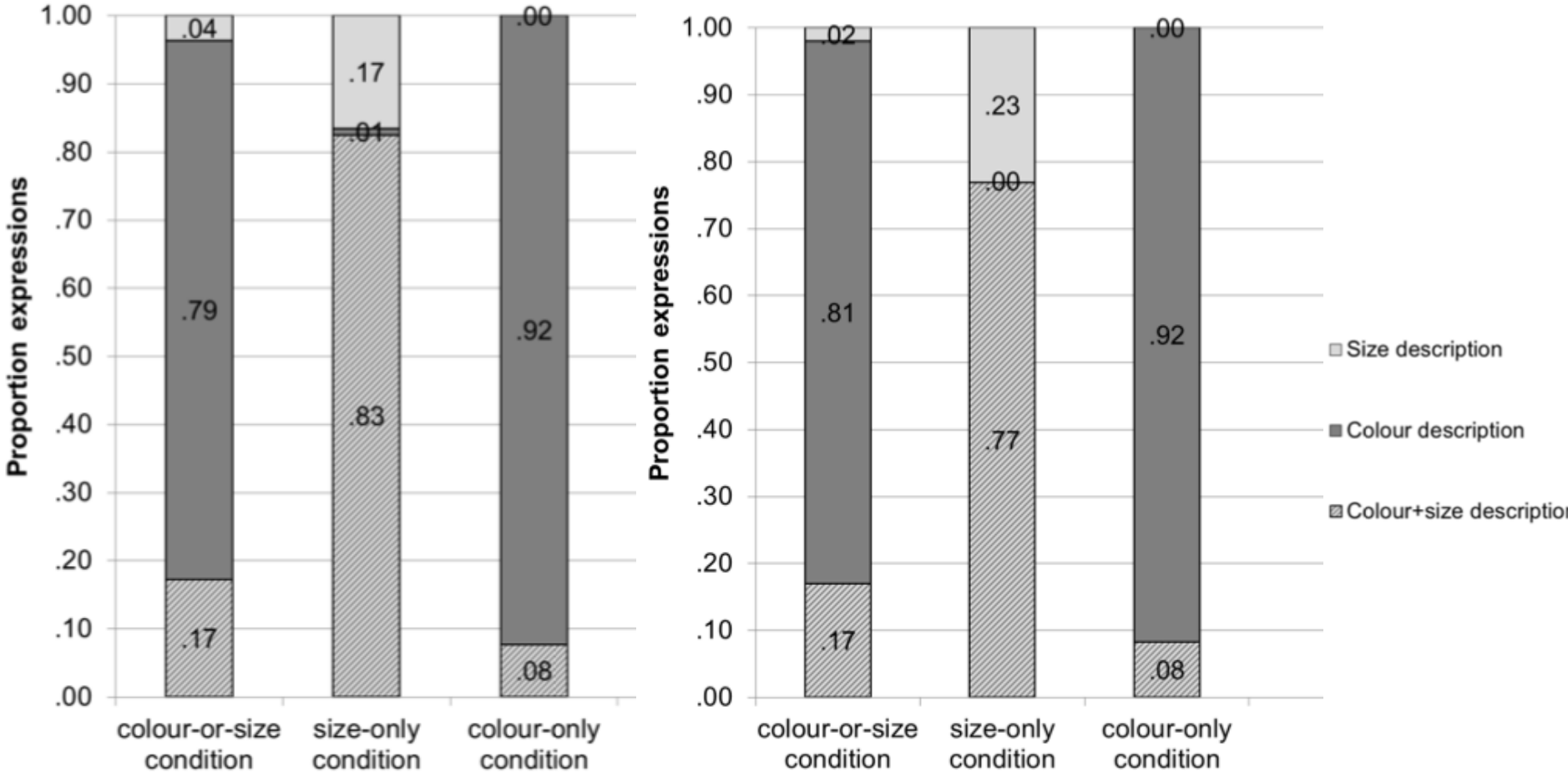
PRO in the situations investigated in the experiment



Human speakers vs PRO



Human speakers vs PRO



Factors influencing x and y

Earlier research suggests:

- y is influenced by whether the domain is fault critical (Arts et al. 2012)
- y is influenced by the amount of clutter in the scene (Koolen et al. 2013)