SEMANTIC COMPLEXITY INFLUENCES QUANTIFIER DISTRIBUTION IN CORPORA

Jakub Szymanik
joint work with Camilo Thorne

SEARCHING FOR VARIOUS WAYS TO ESTIMATE COMPLEXITY AND EXPRESSIVITY OF NATURAL LANGUAGE

“

Linguists and non-linguists alike agree in seeing human language as the clearest mirror we have of the activities of the human mind, and as a specially important of human culture, because it underpins most of the other components. Thus, if there is serious disagreement about whether language complexity is a universal constant or an evolving variable, that is surely a question which merits careful scrutiny. There cannot be many current topics of academic debate which have greater general human importance than this one.

-Sampson et al., 2009

QUESTIONS

- What are the semantic bounds of natural languages?
- What is the ‘natural class of concepts’ expressible in NL?
- Are there differences between various languages?
- How powerful must be our linguistic theory?
- Are some natural language concepts harder to process?

PROOF-OF-CONCEPT

- Focus on one aspect of NL: quantifiers.
- Pick a measure of semantic complexity.
- Can it explain NL distributions?
- Bridging hypothesis: Zipf's principle of least effort.
- Result: distribution is skewed towards low complexity.
- Conclusion: the semantic complexity measure is relevant for NL complexity debate.

QUANTIFIERS

GENERALIZED QUANTIFIERS: EXAMPLES

1. All poets have low self-esteem.

2. Some dean danced nude on the table.

3. At least 3 grad students prepared presentations.

4. An even number of the students saw a ghost.

5. Most of the students think they are smart.

6. Less than half of the students received good marks.

7. Many of the soldiers have not eaten for several days.

8. A few of the conservatives hate each other.

DEFINITION OF GQ

A quantifier Q is a way of associating with each set M a function from pairs of subsets of M into {0; 1} (False, True).

$$\text{every}_M[A, B] = 1 \text{ iff } A \subseteq B$$

$$\text{even}_M[A, B] = 1 \text{ iff } \text{card}(A \cap B) \text{ is even}$$

$$\text{most}_M[A, B] = 1 \text{ iff } \text{card}(A \cap B) > \text{card}(A - B)$$

SPACE OF GQ

A quantifier Q is a way of associating with each set M a function from pairs of subsets of M into {0; 1} (False, True).

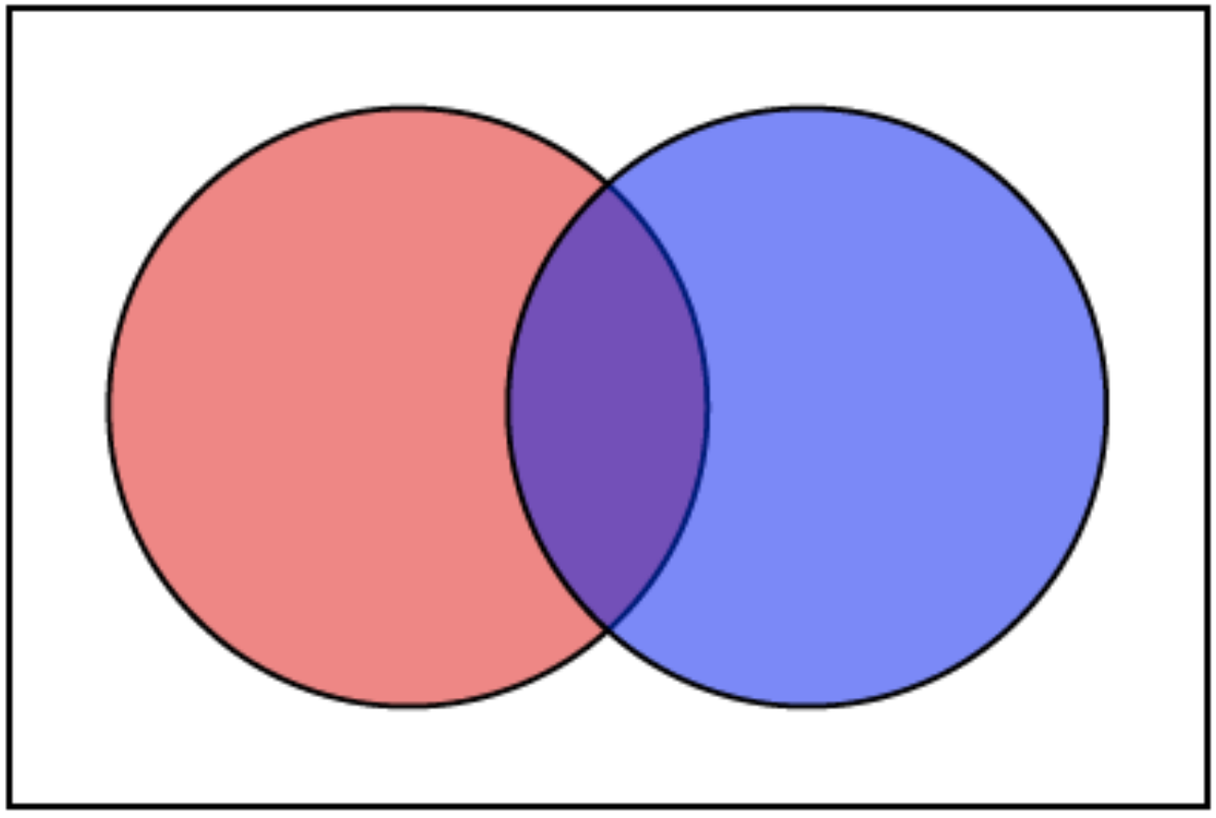
If $\mathrm{card}(M) = n$, then there are $2^{2^{2n}}$ GQs.

For $n = 2$ it gives 65,536 possibilities.

Which of those correspond to natural concepts?

Isomorphism closure

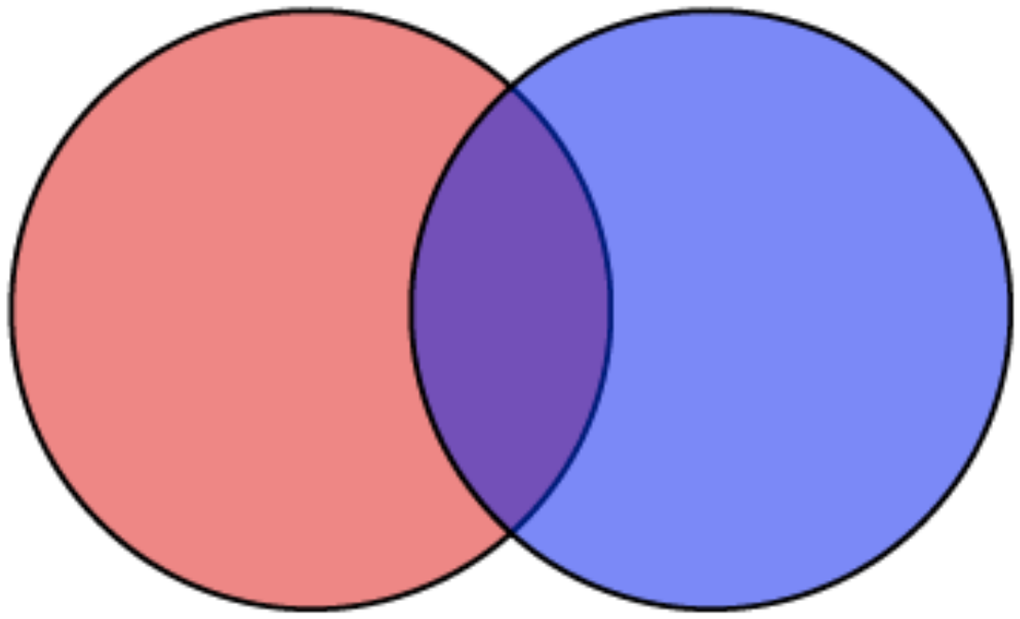
(ISOM) If $(M, A, B) \cong (M', A', B')$, then $Q_{\mathbf{M}}(A, B) \Leftrightarrow Q_{\mathbf{M}'}(A', B')$



Topic neutrality

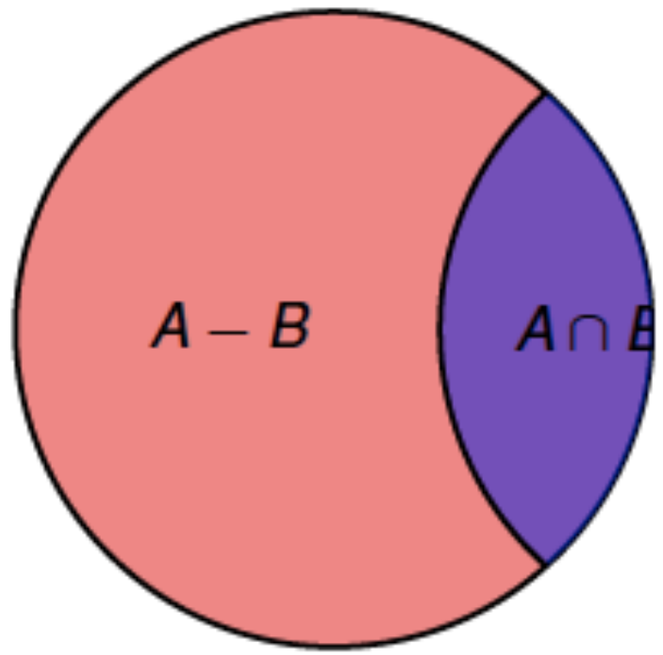
Extensionality

(EXT) If $M \subseteq M'$, then $Q_{\mathbf{M}}(A, B) \Leftrightarrow Q_{\mathbf{M}'}(A, B)$



Conservativity

(CONS) $Q_{\mathbf{M}}(A, B) \Leftrightarrow Q_{\mathbf{M}}(A, A \cap B)$



QUESTIONS

- Do all NL determiners satisfy ISOM, EXT and CONS?
- What about `only', `every third'?
- But why majority do?
- Can complexity be another semantic universale?

COMPLEXITY

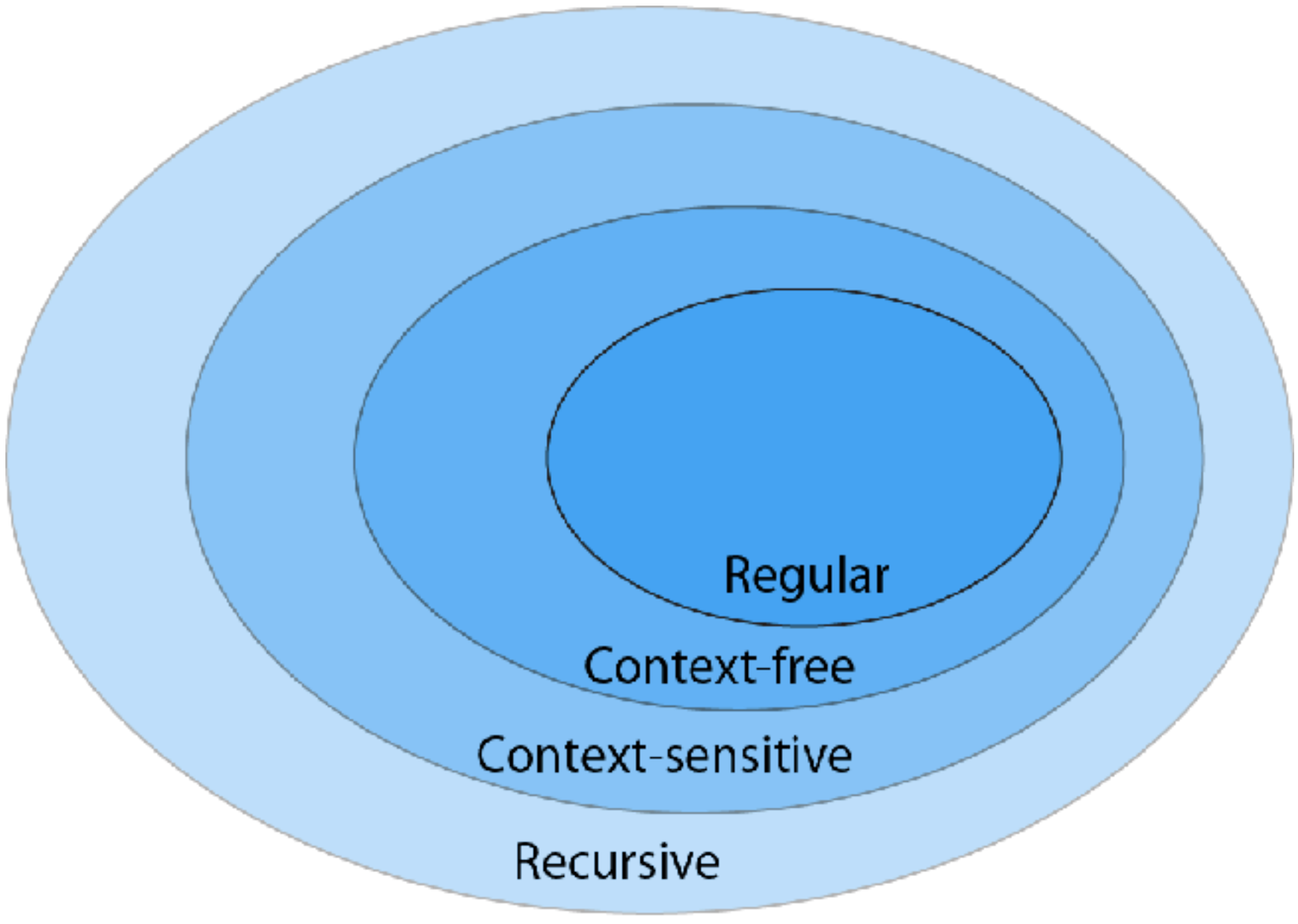
HOW DO WE MEASURE COMPLEXITY?

- Existing approaches depend on implementation/theory:
  - Chomsky hierarchy
  - Typological approach (McWhorther, 2001; Everett, 2008)
  - Information-theoretic approach (Juola, 2009)

INHERENT COMPLEXITY

- Inherent complexity of the concept
- and not the particular implementation.

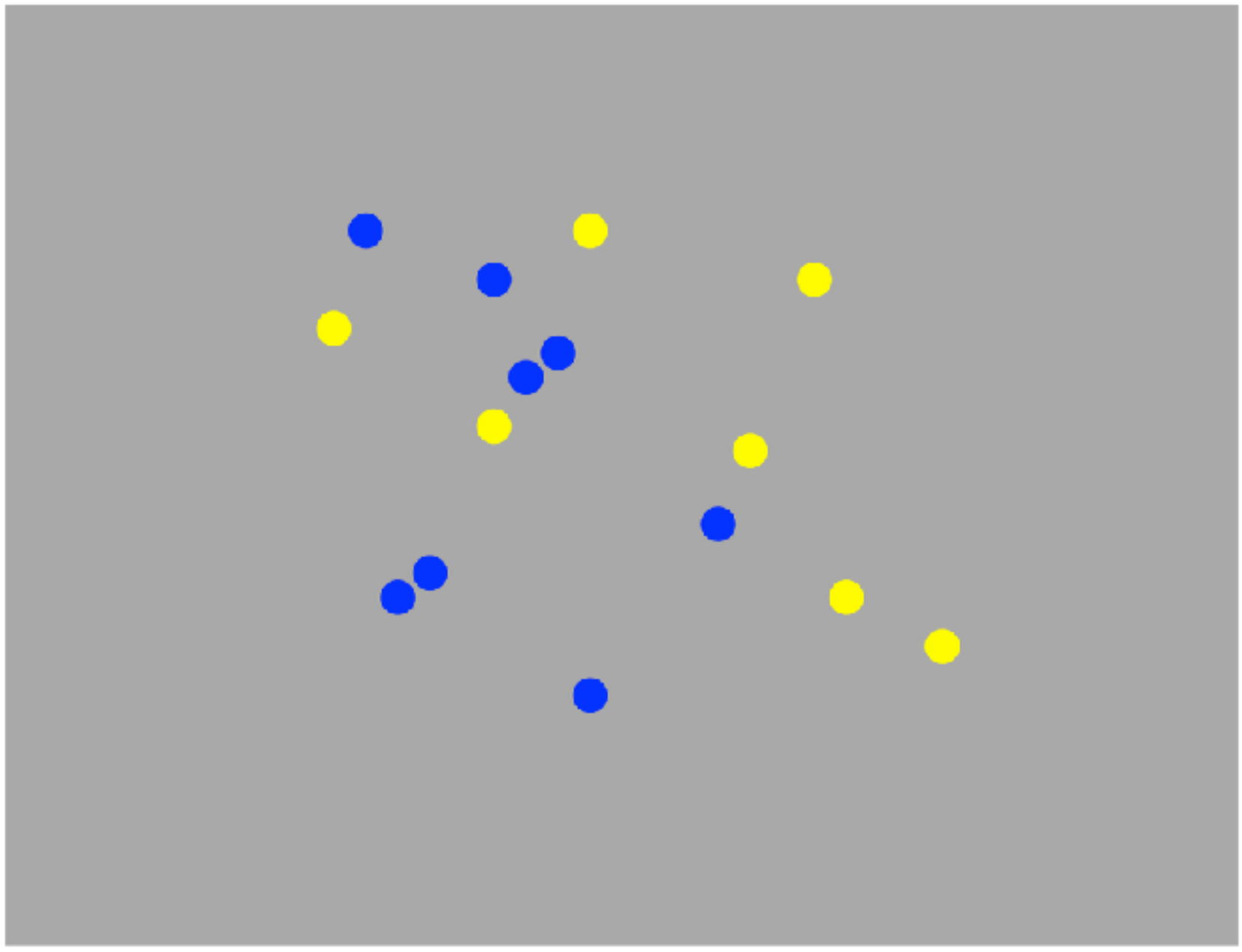
E.G., IN TERMS OF CHOMSKY HIERARCHY



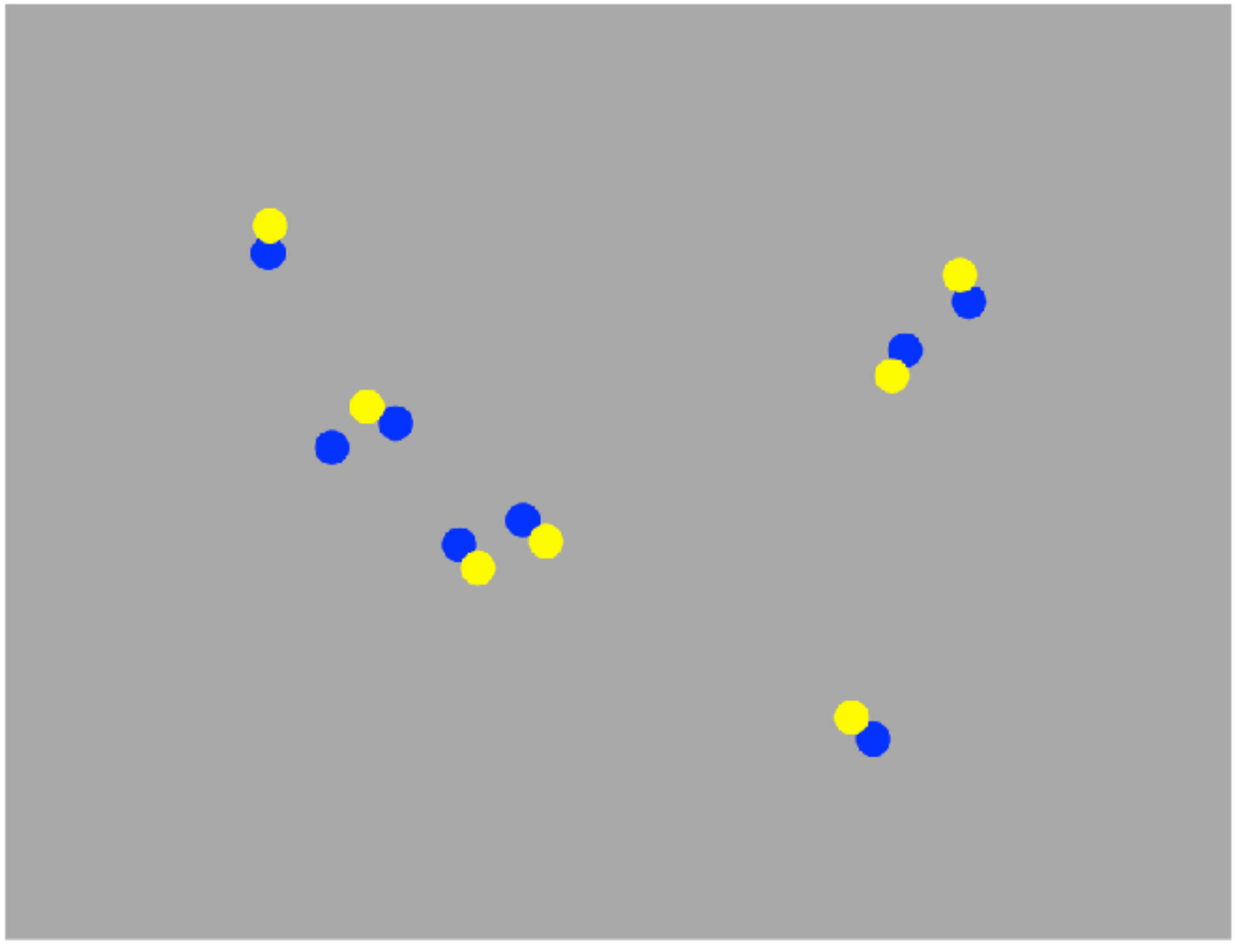
COMPLEXITY OF WHAT?

- Various semantic problems:
  - inferential meaning (complexity of reasoning)
  - referential meaning (complexity of verification)

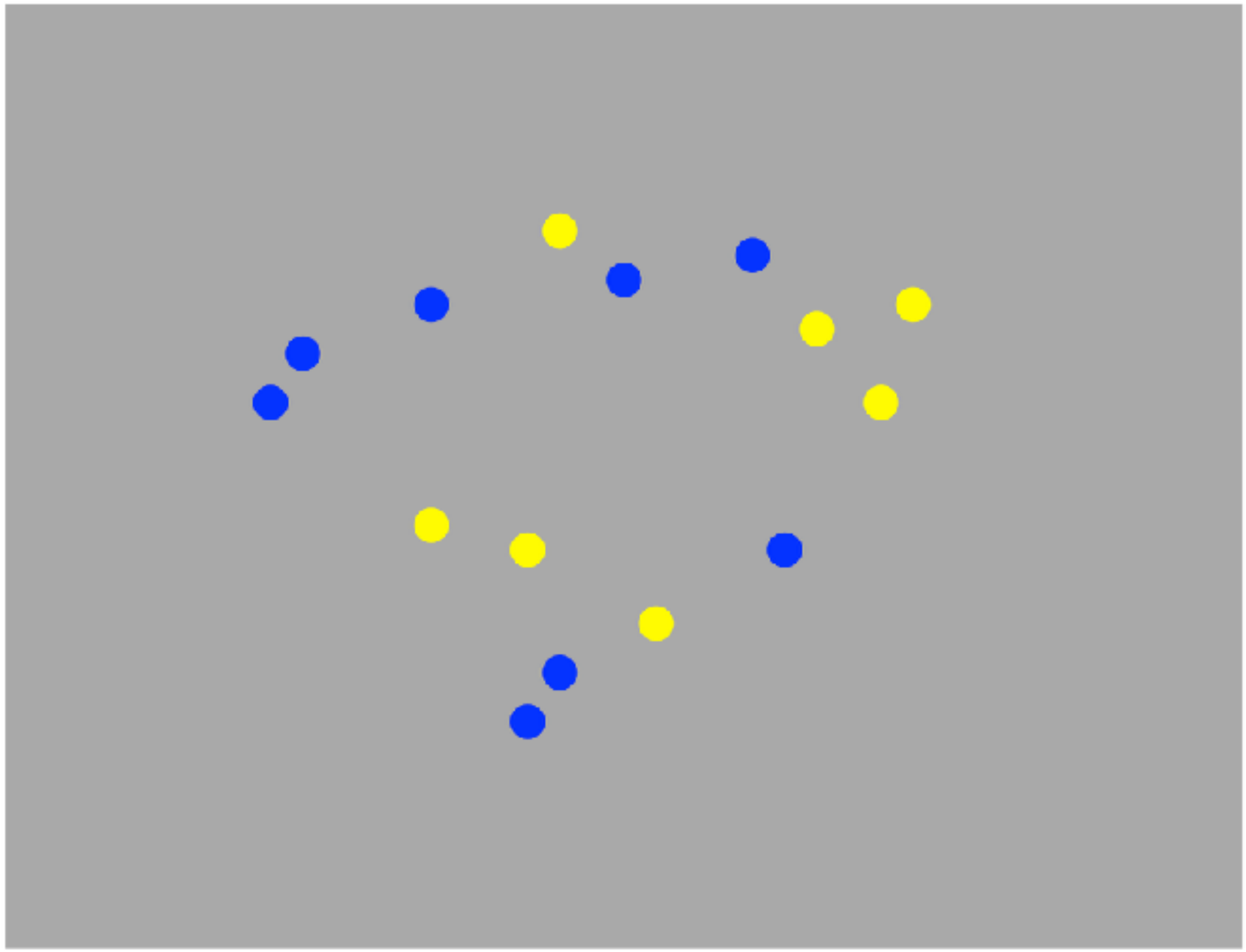
SOME OF THE DOTS ARE BLUE



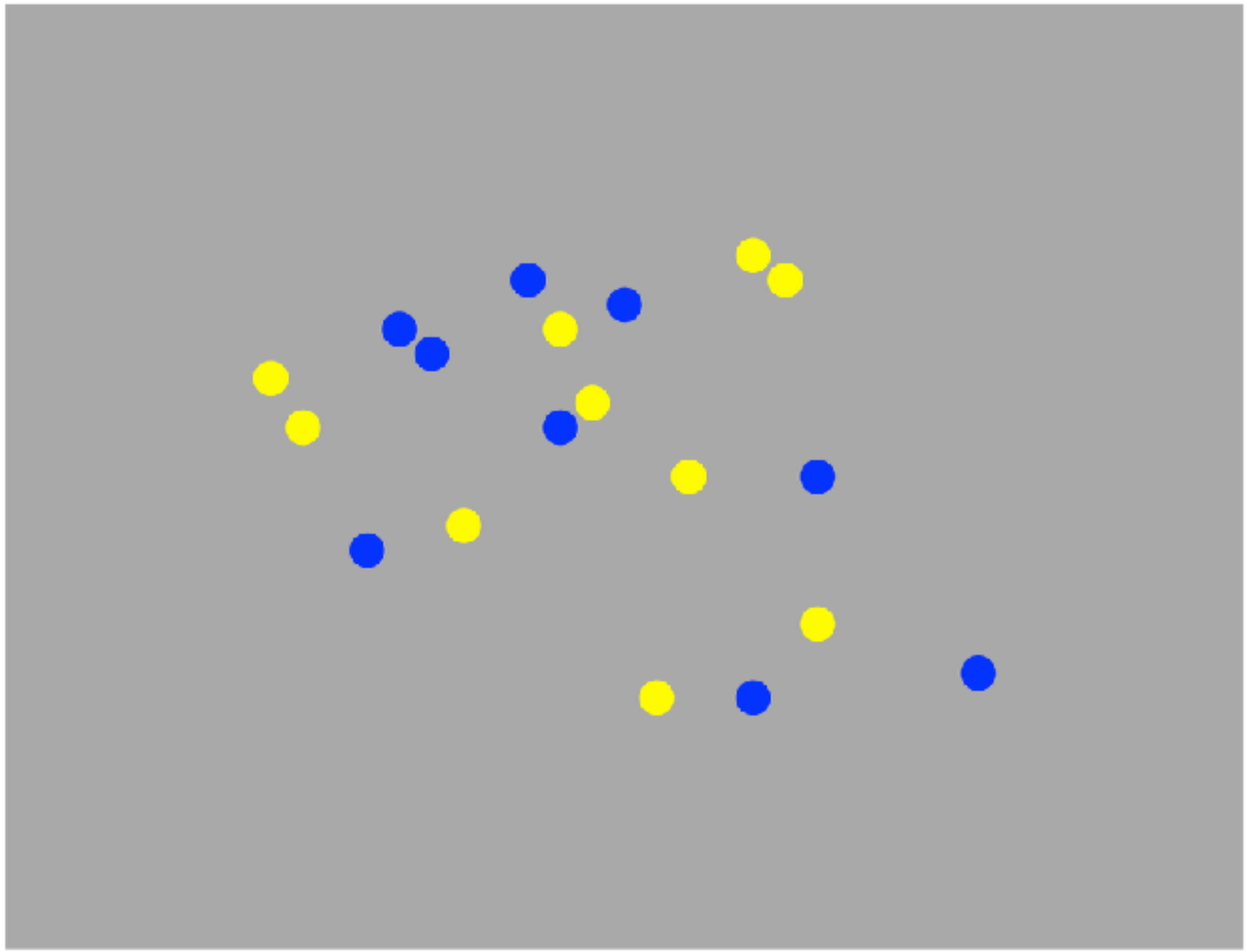
MORE THAN 5 OF THE DOTS ARE BLUE



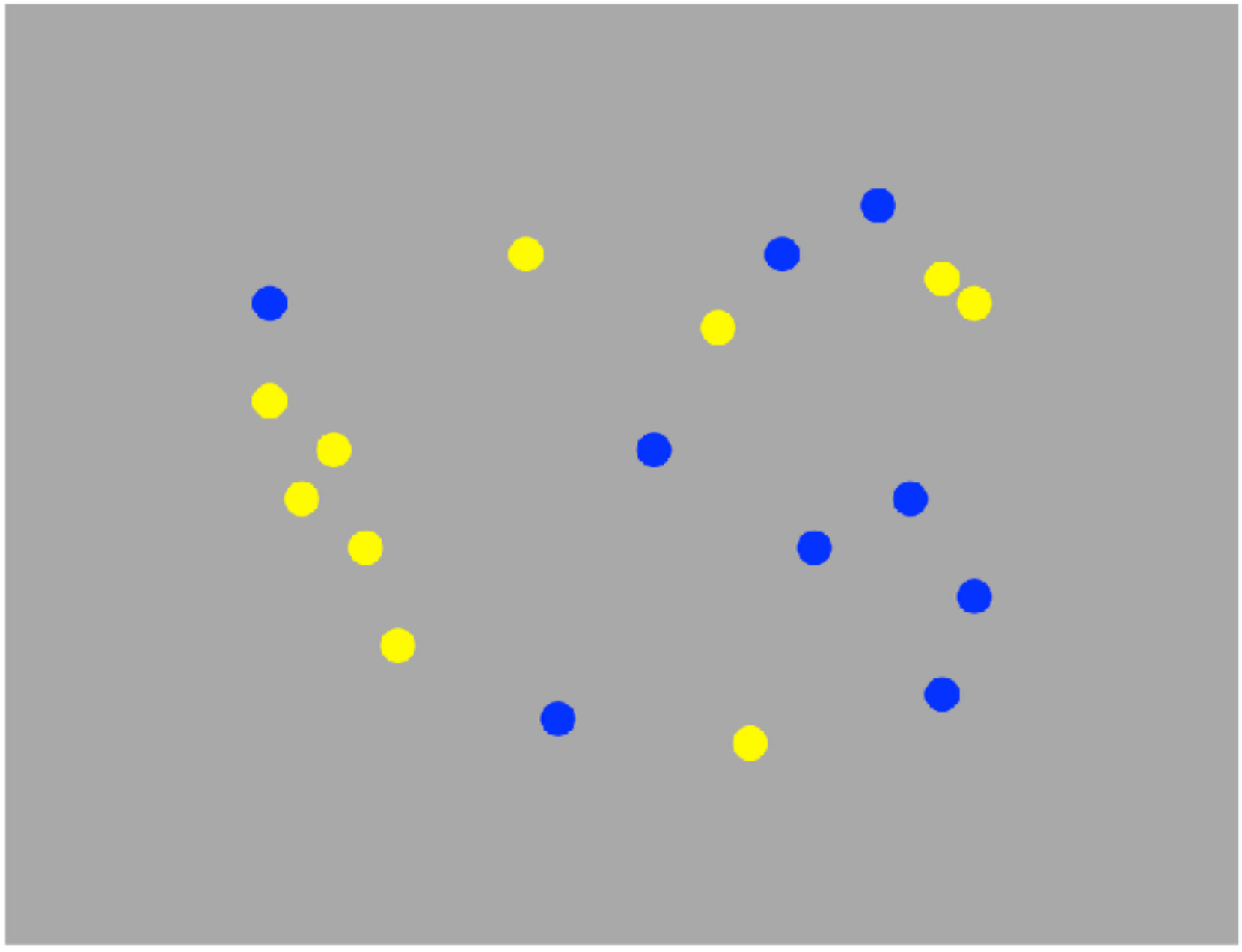
FEWER THAN 7 OF THE DOTS ARE BLUE



AN EVEN NUMBER OF THE DOTS ARE BLUE

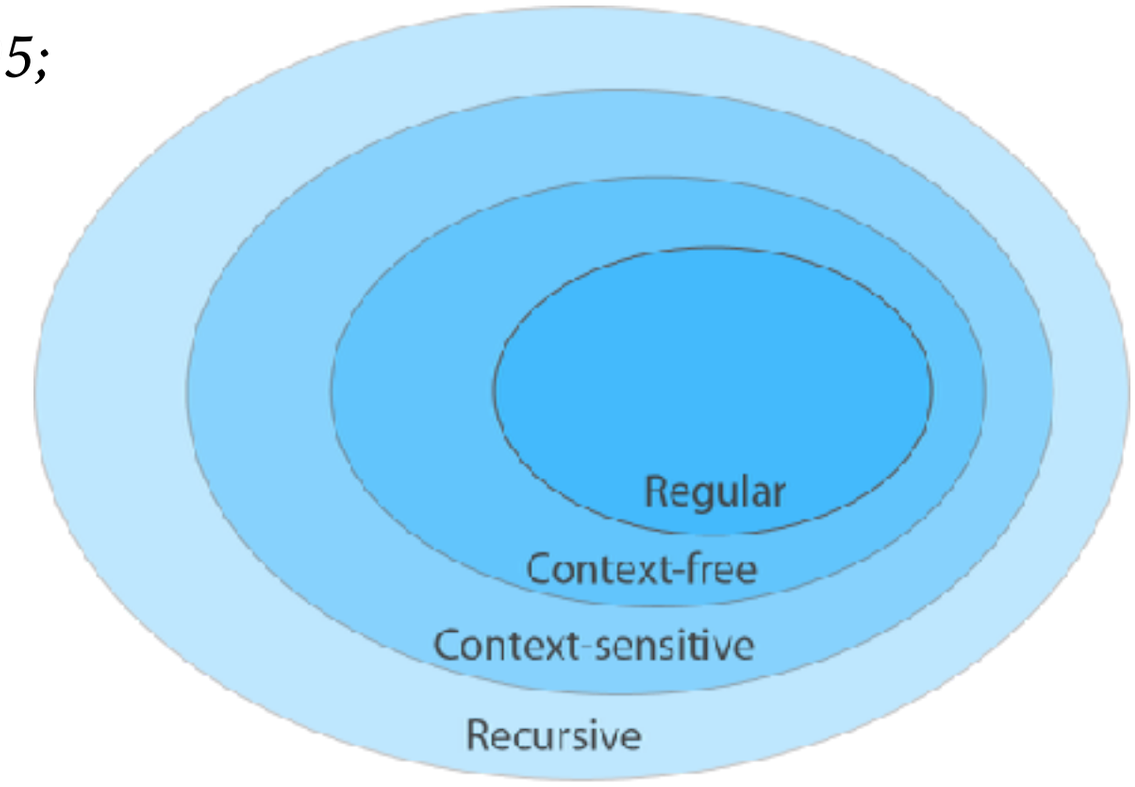


LESS THAN HALF OF THE DOTS ARE BLUE

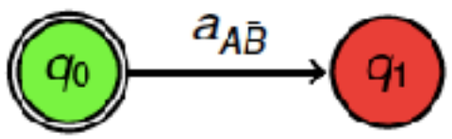


LOGIC & COMPLEXITY CLASSIFICATIONS

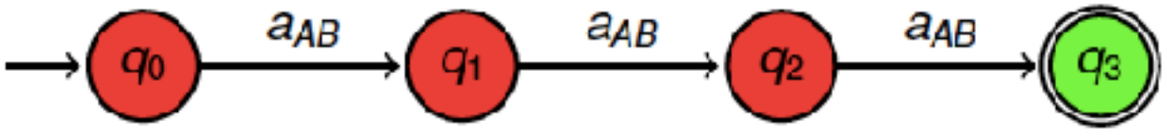
Barwise & Cooper, `81; van Benthem, '86; Mostowski, '95;
Stanley & Westerståhl, '06;
Szymanik, `16…



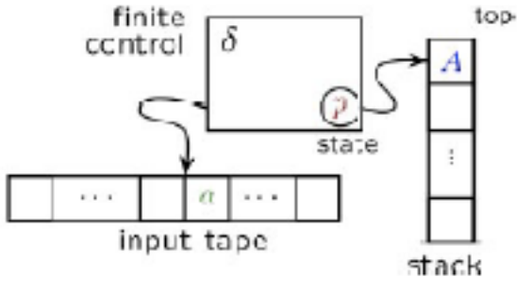
All As arə B.

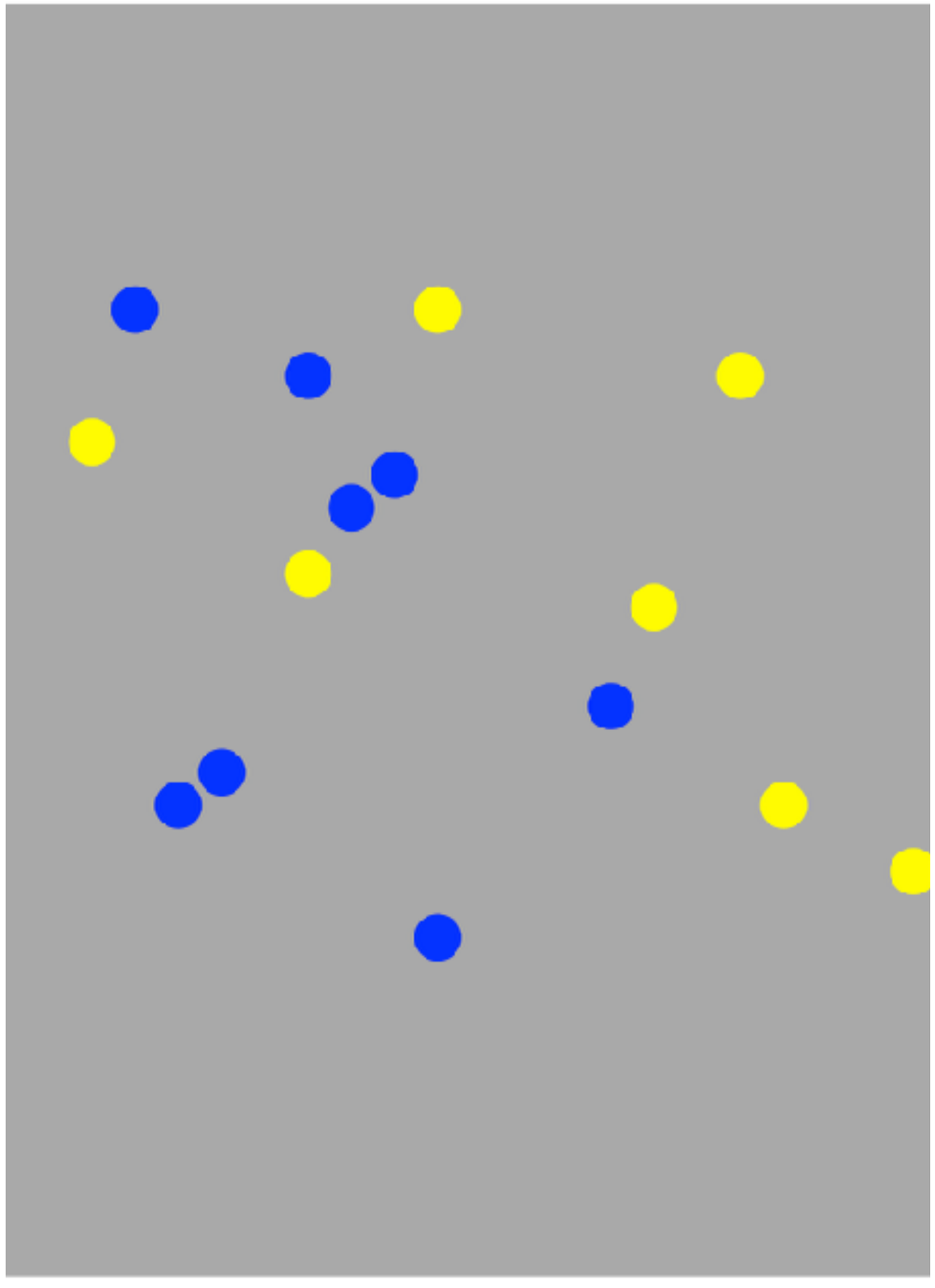


More than 2 As are B.



Most As are B.





DRAW AUTOMATA

- **Some dots are blue.**
- All dots are blue.
- No dots are blue.
- Some dots are not blue.
- **More than 3 dots are blue.**
- Fewer than 4 dots are blue.
- **An even number of dots are blue.**
- An odd number of dots are blue.
- **Most dots are blue.**
- Less than half dots are blue.

CLASSIFYING MINIMAL COMPLEXITY

- Aristotelian quantifiers, e.g, all, some, no, some-not (2-state FA)
- Numerical quantifier, e.g, more than 5 (FA)
- Proportional quantifier, e.g., most (PDA)

van Benthem'86, Mostowski'98,

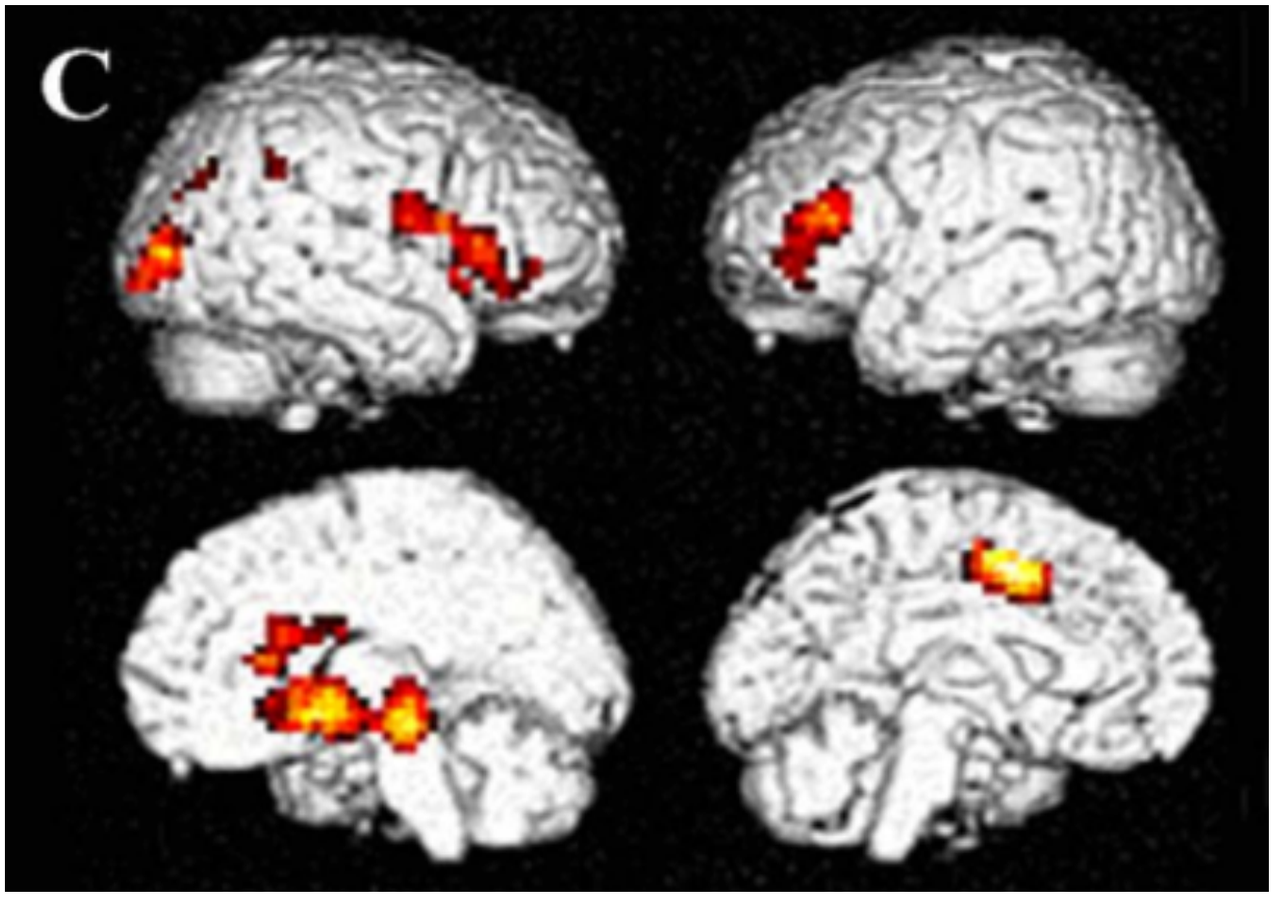
Kanazawa'13; Steinert-Threlkeld & Icard'13, Szymanik'16

ARE LOGICAL DISTINCTIONS PLAUSIBLE?

Differences in brain activity:

A. All quantifiers are associated with numerosity: recruit right inferior parietal cortex.

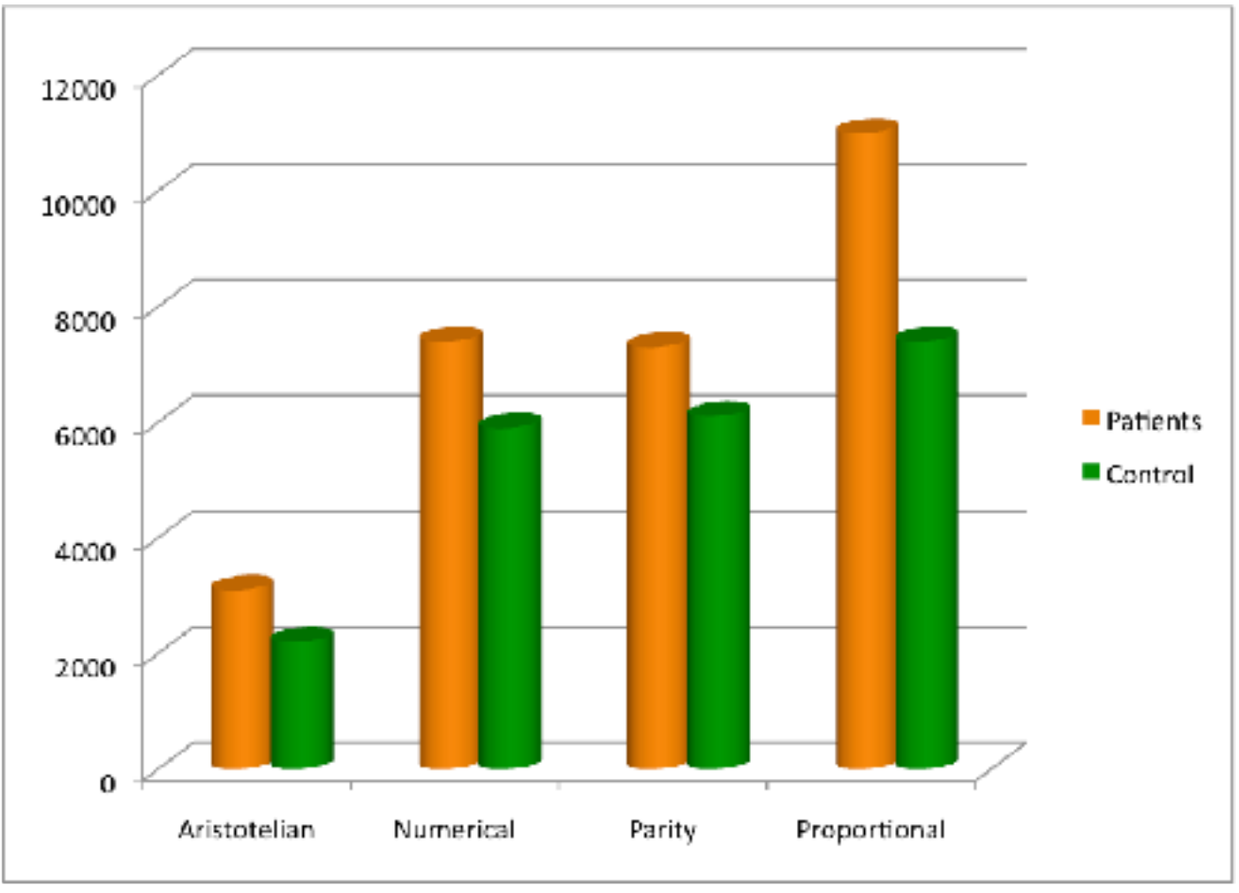
B. Only higher-order activate working-memory capacity: recruit right dorsolateral prefrontal cortex.

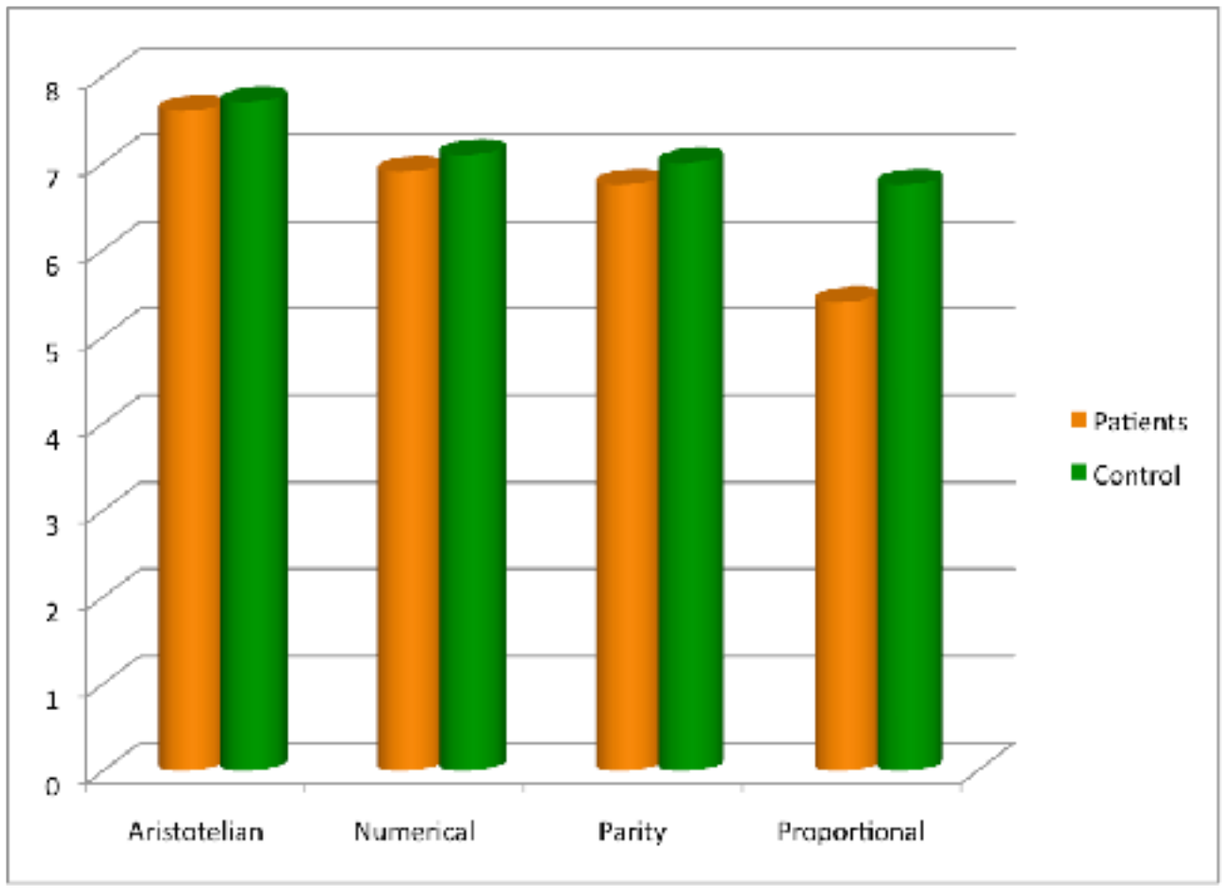


McMillan et al.'05, '06, Szymanik'07

ARE LOGICAL DISTINCTIONS PLAUSIBLE?

Behavioral differences:





Szymanik & Zajenkowski'10, Zajenkowski et al.'11, Szymanik'16

OTHER SEMANTIC FACTORS

MONOTONICITY

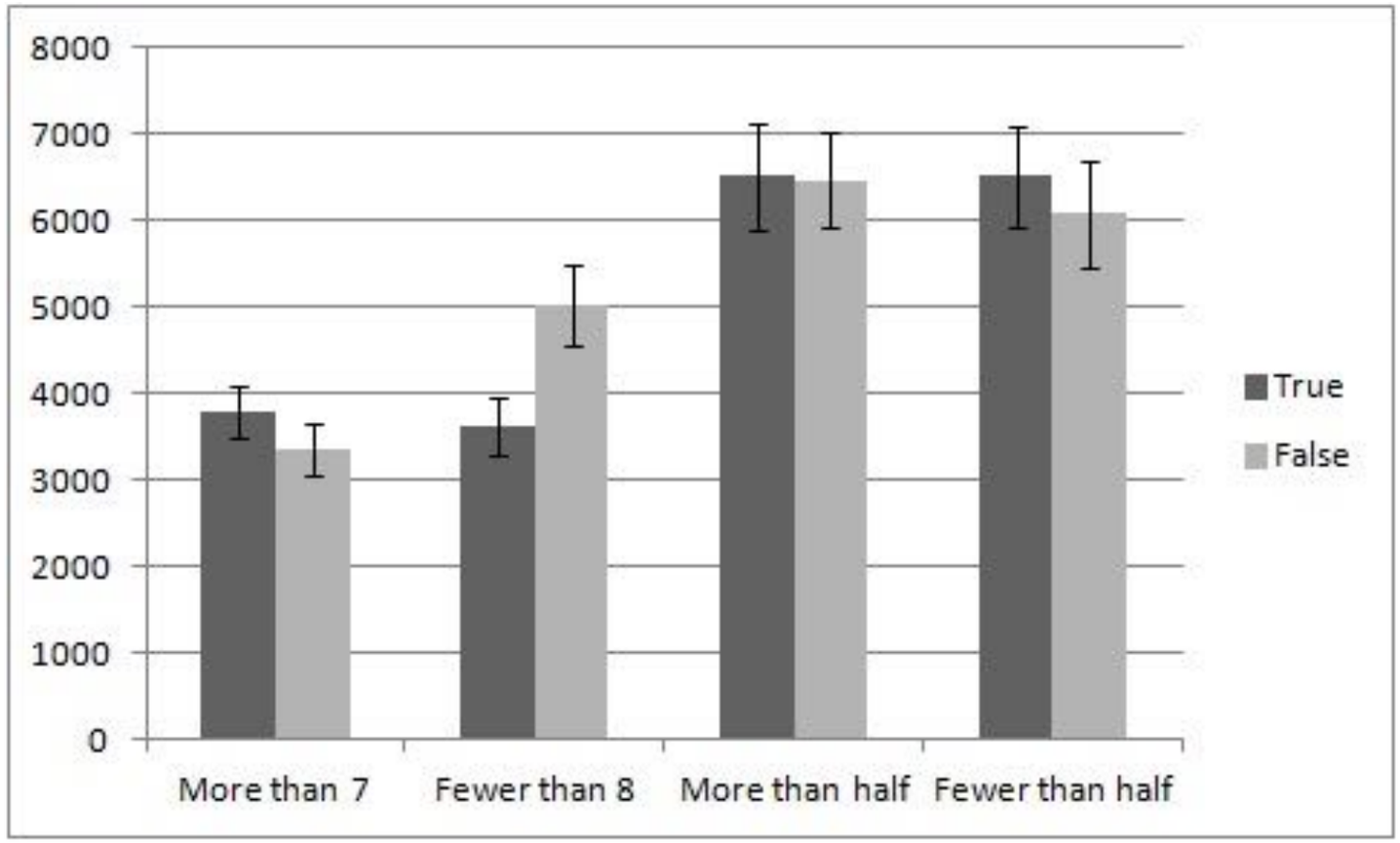
A quantifier Q is increasing/decreasing in its right argument if and only if, for any sets A, B and B', if B is a subset of B', then:

Q(A, B) entails Q(A, B') (resp, Q(A, B') entails Q(A, B)).

1. Most boys are very happy.
2. Most boys are happy.
3. No children are happy.
4. No children are very happy.

MONOTONICTY INFLUENCES PROCESSING

Downward monotone quantifiers are harder to process than upward monotone ones.



Szymanik & Zajenkowski'13

SUPERLATIVES VS COMPARATIVES

`more than 3' = `at least 4'

- Geurts and Nouwen, 2007: superlatives have richer meaning
- Geurts et al., 2009:
  - superlative are mastered later in the developmental process
  - superlatives are harder to process than comparatives

EXPERIMENT

Is GQ distribution influenced by semantic complexity?

QUANTIFIERS

Class	Examples	Quantifier	Complexity
Aristotelian	'every', 'some'	all, some	2-state acyclic FA
counting	'more than 4', 'at most 5'	>k, <k	k+2-state FA
proportional	'most', 'less than half' 'few', 'more than half' 'less than three-fifths' 'more than two-thirds'	most, <1/2, few, >1/2, <p/k, >p/k	PDA

CORPUS

- WaCky corpus (Baroni, 2009)
- Sentences ~ 43 millions
- Tokens ~ 800 milions
- POS-annotated

PROBLEM: HOW TO FIND QUANTIFIERS?

E.g. word `most' can be:

- a determiner (tag DT), or
- adverb (RBR/RBS)

When does it denote proportional quantifier?

When followed by a plural noun (NNS) as in:

- 'most/DT men/NNS',

rather than an adjective (JJ) as in:

- 'most/DT grateful/JJ.

LINGUISTICALLY PLAUSIBLE PATTERNS

E.g., GQ >k we matched with:

- `'at/in least/jjs [a-z]{1,12}/cd', viz., the preposition 'at' followed by the superlative adjective 'least' and a cardinal comprising up to 12 characters;`
- `'more/rbr than/in [a-z]{1,12}/cd', viz., the comparative adverb 'more' followed by the preposition 'than' and a cardinal;`
- `'more/jjr than/in [a-z]{1,12}/cd', viz., the same as before, but with 'more' a comparative adjective.`

In total we counted occurrences of 36 patterns.

FEATURES

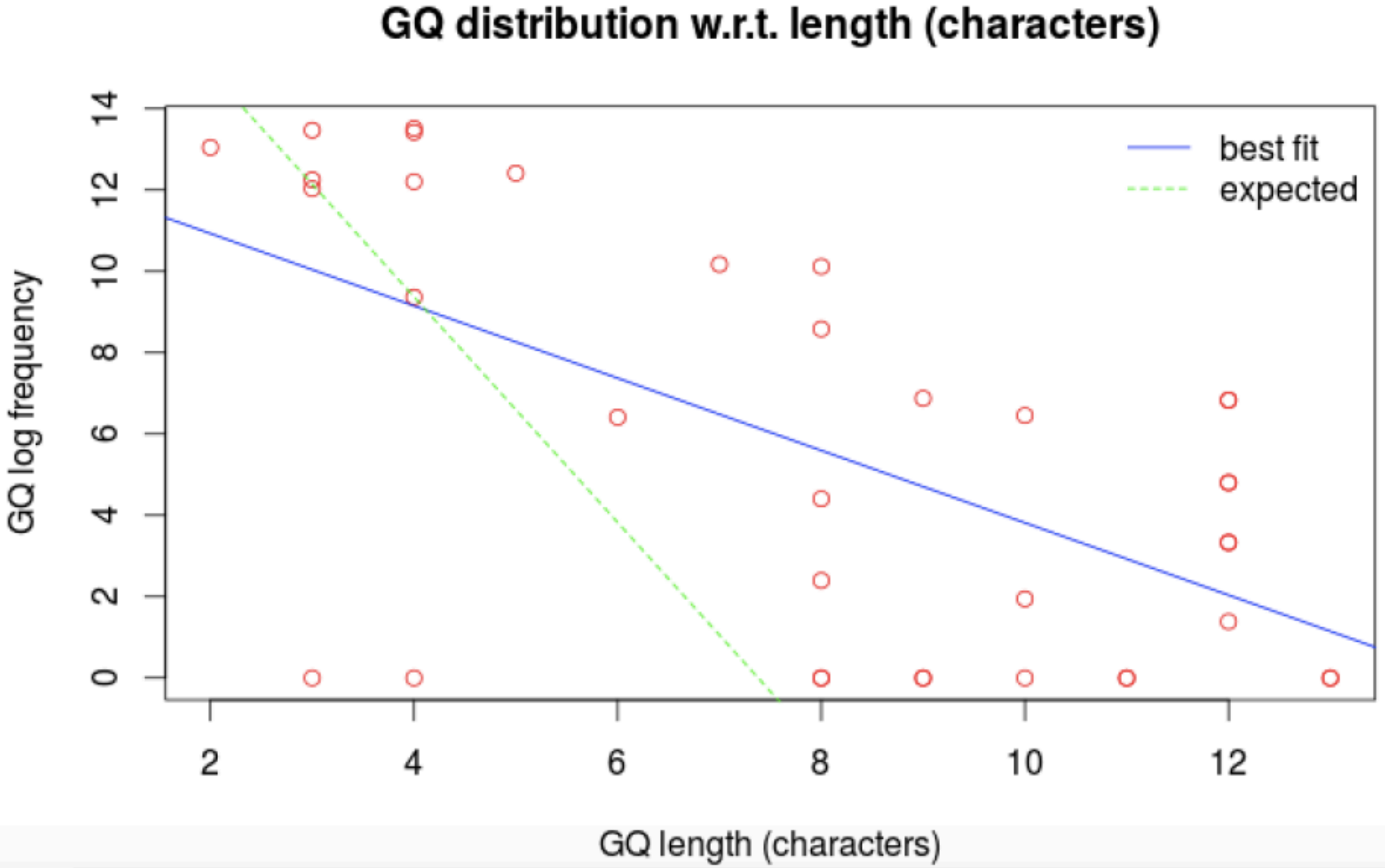
- Frequency: counts for each pattern.
- Frequency rank: GQ patterns by frequency order.
- Class: Aristotelian, counting, proportional.
- Monotonicity: ‘up’, ‘down’, and ‘none’.
- Type: comparative or superlative.
- Length in words: number of word tokens.
- Length in characters: number of characters.

Enough to make the conceptual point.

DESCRIPTIVE ANALYSIS

- Aristotelian > counting > proportional, but also:
- short > multiword
- Both syntax and semantics influence.

BUT LENGTH SEEMS NOT TO BE EVERYTHING



GLM REGRESSION ANALYSIS

- GQ frequency as a complex function of various factors.
- Model selection
- 4 independent factors:
  - class (~27%),
  - type,
  - **length** (~47%),
  - right monotonicity (~26%).

semantic complexity underlies GQ distributions

LIMITATIONS OF THE STUDY

- How semantic complexity relates to language production?
- Ignore pragmatics, e.g., round numbers.
- Sacrifices coverage at the expense of precision.
- Annotation quality is limited.
- Relatively small number of observations (i.e., 36 quantifiers)

CONCLUSIONS & OUTLOOK

- Semantic complexity influences linguistic distributions.
- Semantic complexity as a semantic universal.

- We need a quantifier corpus, anyone?