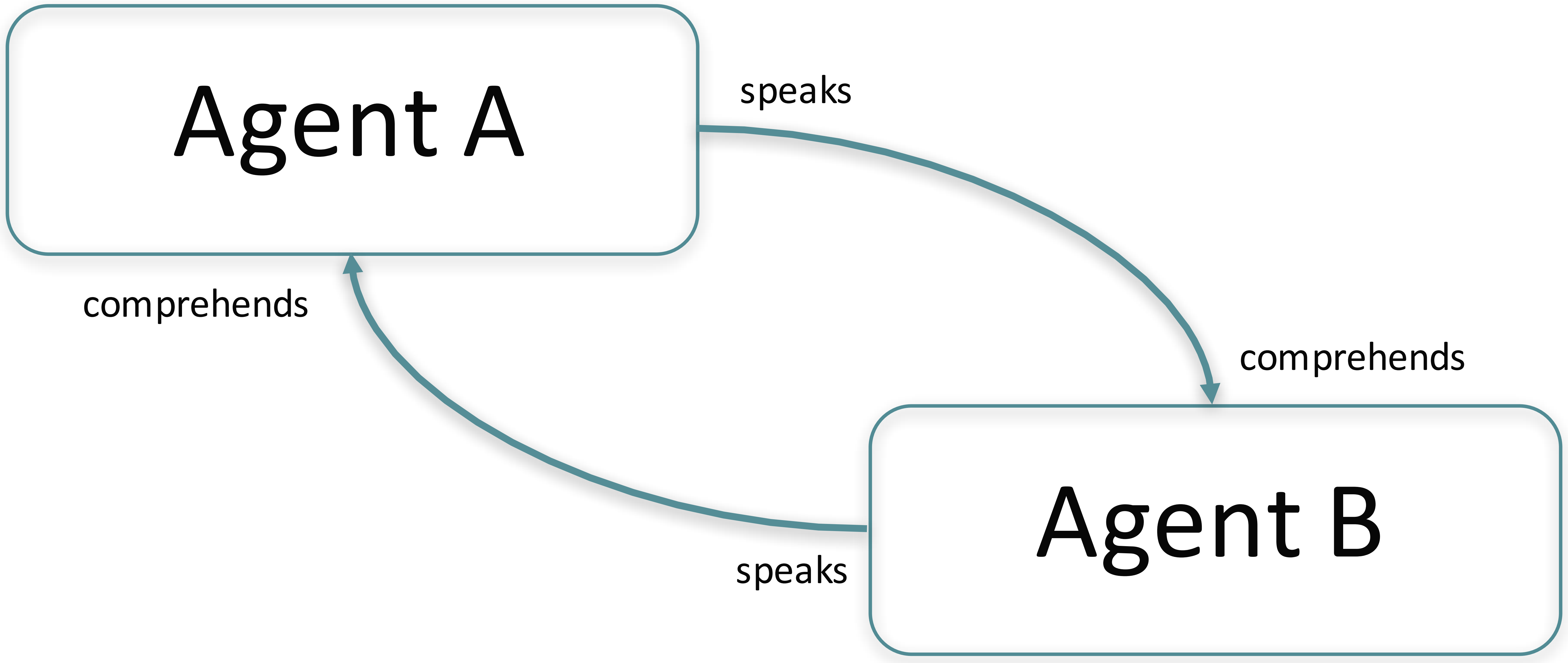
Constraints on Information Processing in Language Comprehension and Production

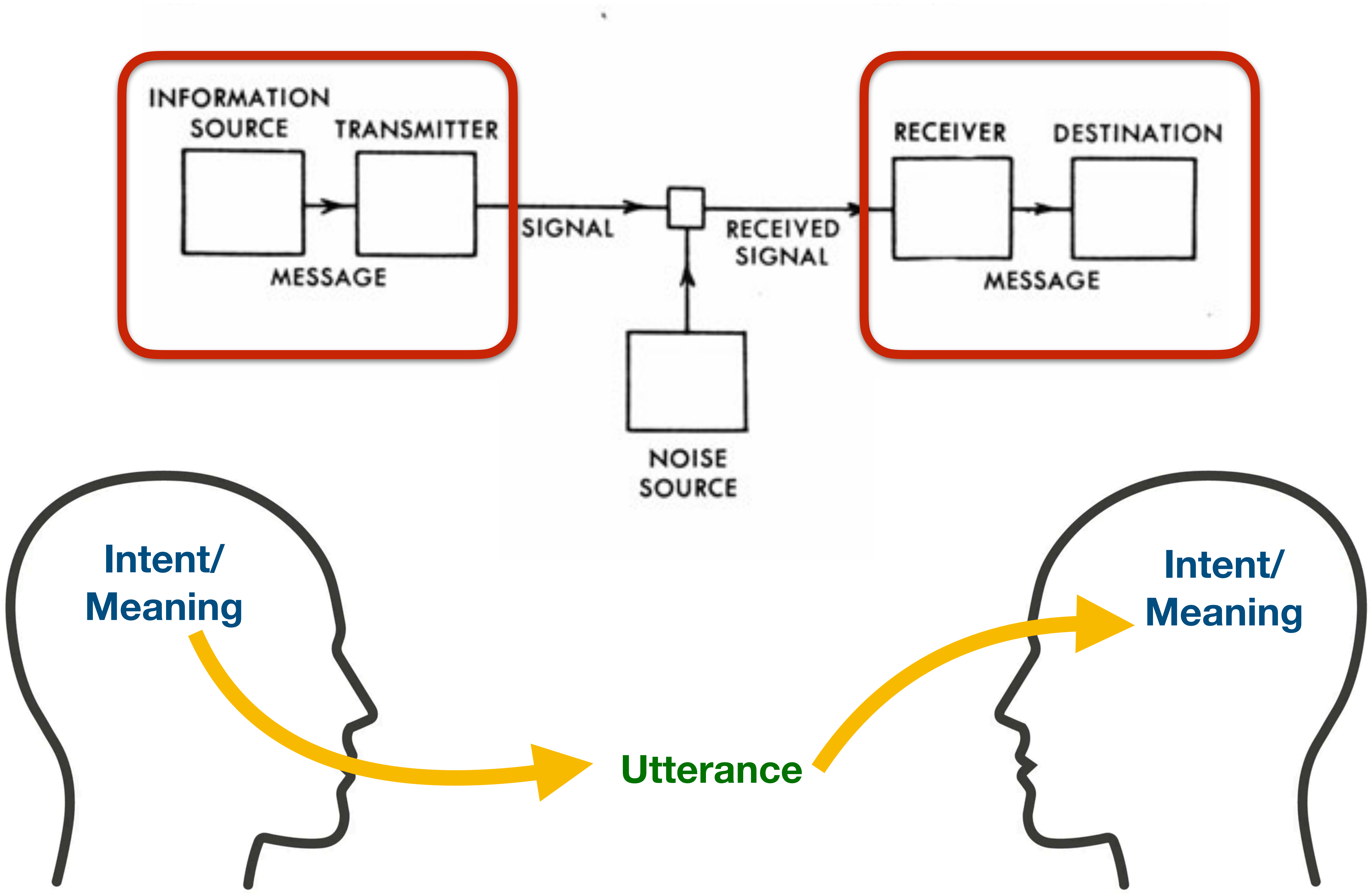
Richard Futrell
Department of Language Science
University of California, Irvine
@rljfutrell

ILFC Seminar
2023-12-13

Natural Language as an Information-Theoretic Code



Natural Language as an Information-Theoretic Code

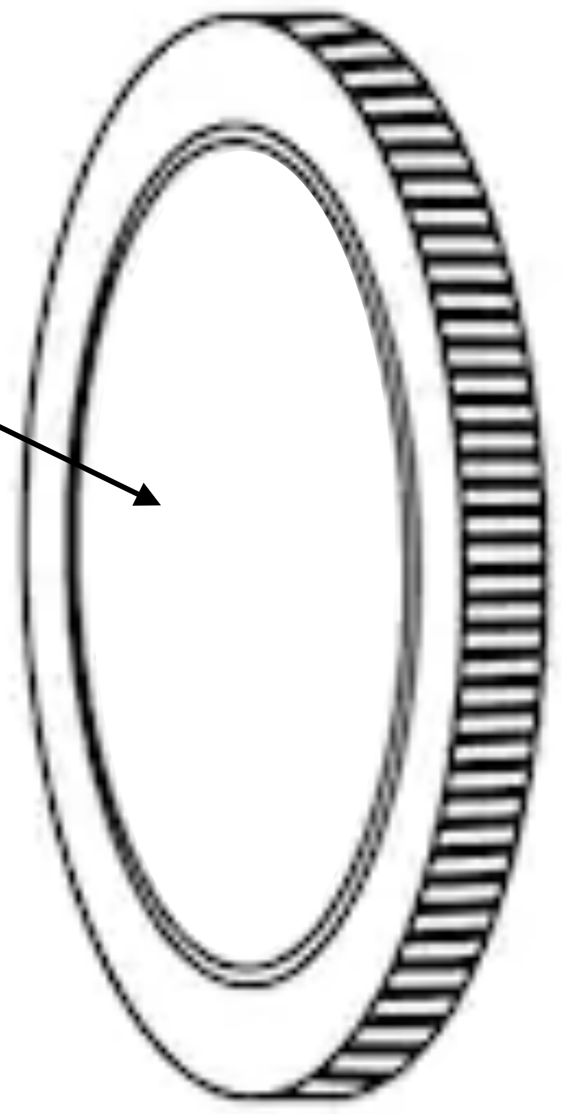


Production **Comprehension**

Claude Shannon (1948). A mathematical theory of communication. *Bell System Technical Journal.*

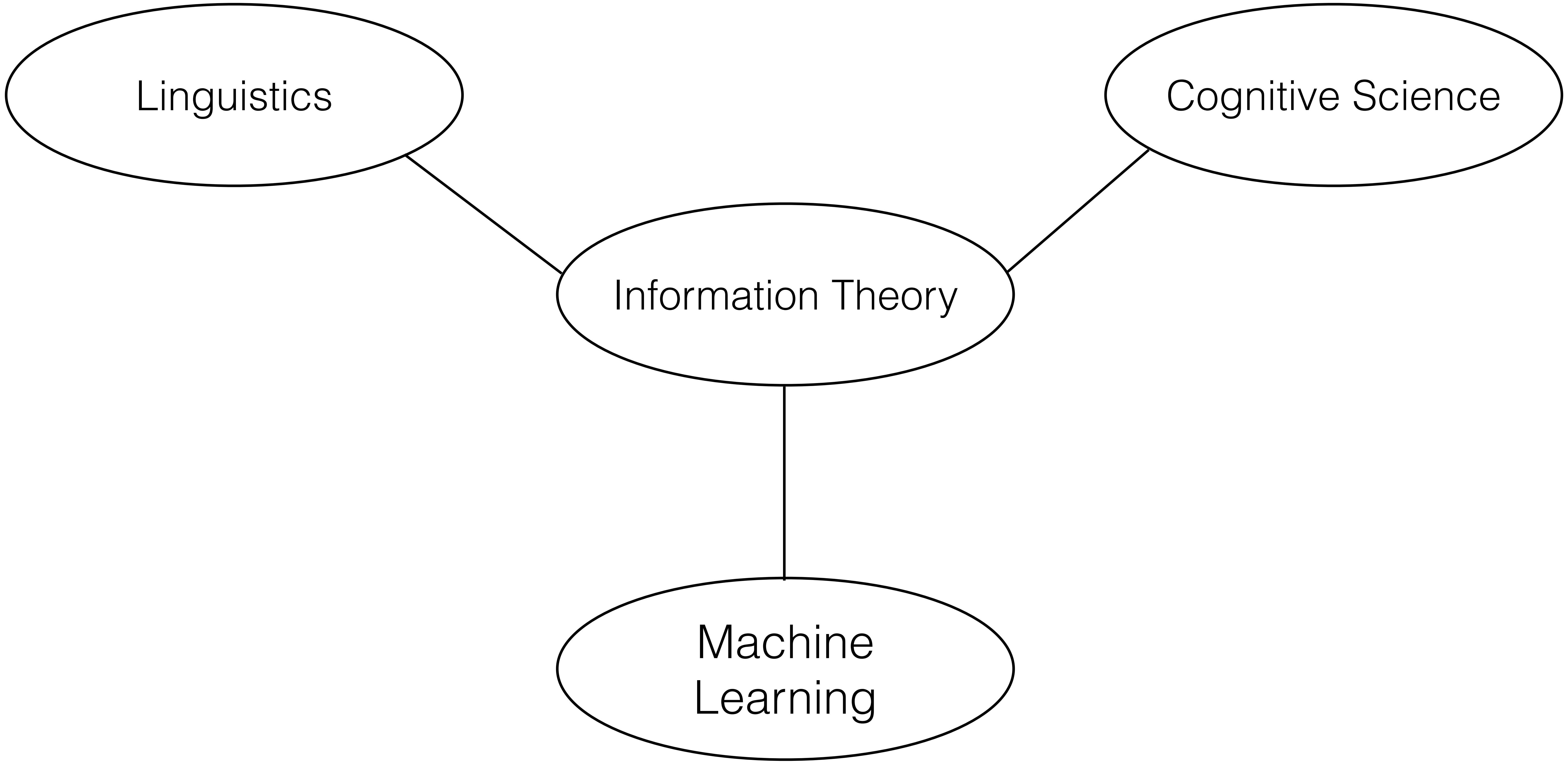
Research Program

Models of language processing



Models of how processing shapes language structure

A Nexus Between Fields



Goals Today

- Develop and test models of **language comprehension** and **language production** based on **maximizing efficiency *subject to constraints*.**
    - Show that a **bottleneck on memory** yields detailed patterns of comprehension difficulty for nested clauses.
    - Show that a **bottleneck on control** yields accessibility effects in incremental production of words.
- On both sides, a **predictive language model** ends up playing a central role.

Outline

- Introduction
- Information Theory for Language Processing
- Memory Bottleneck in Language Comprehension
- Control Bottleneck in Language Production
- Conclusion

What is Information?

- The children went outside to… *play*

011101

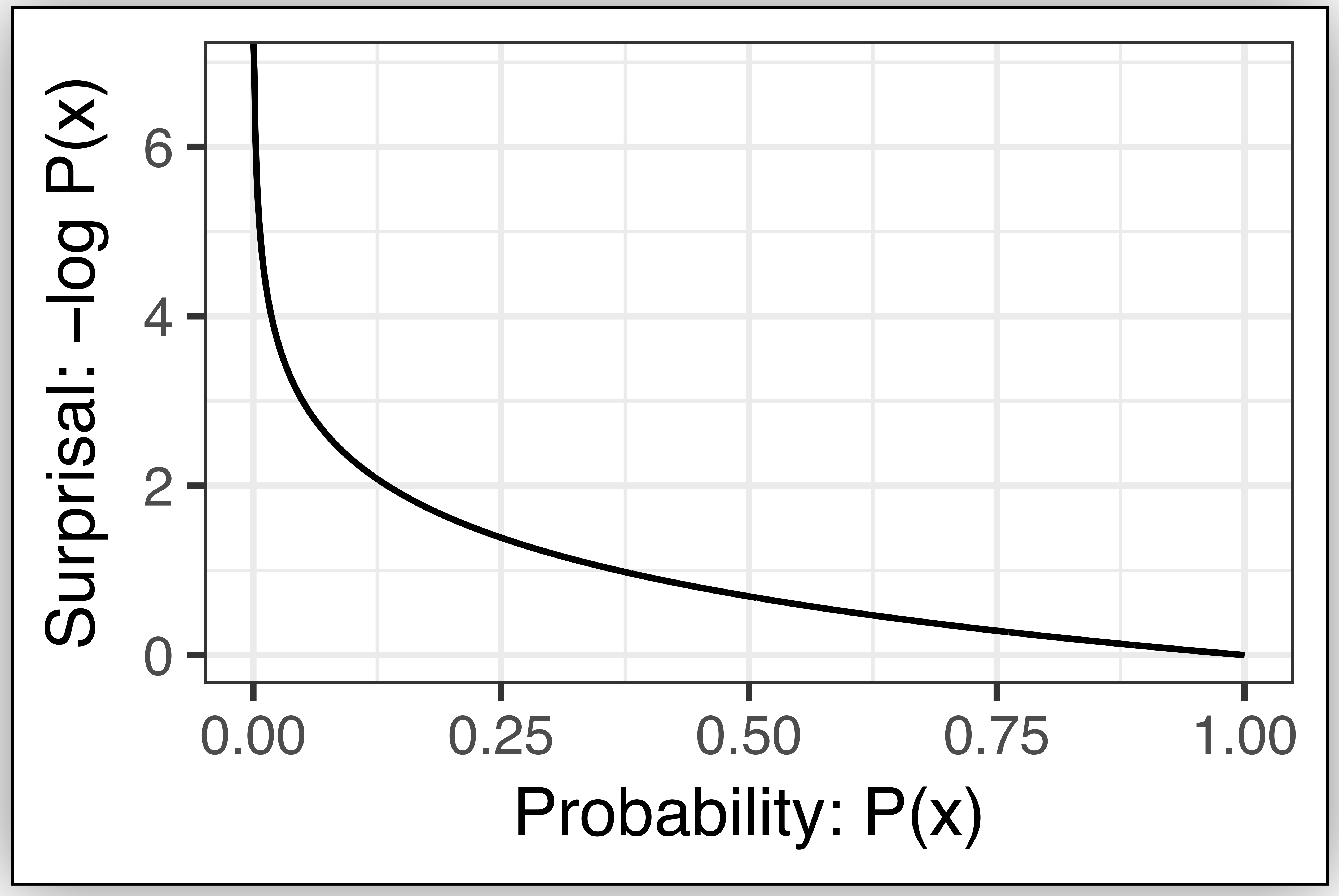
Amount of information ~ Amount of surprise

- The children came inside to… *play*

0111011100001011101

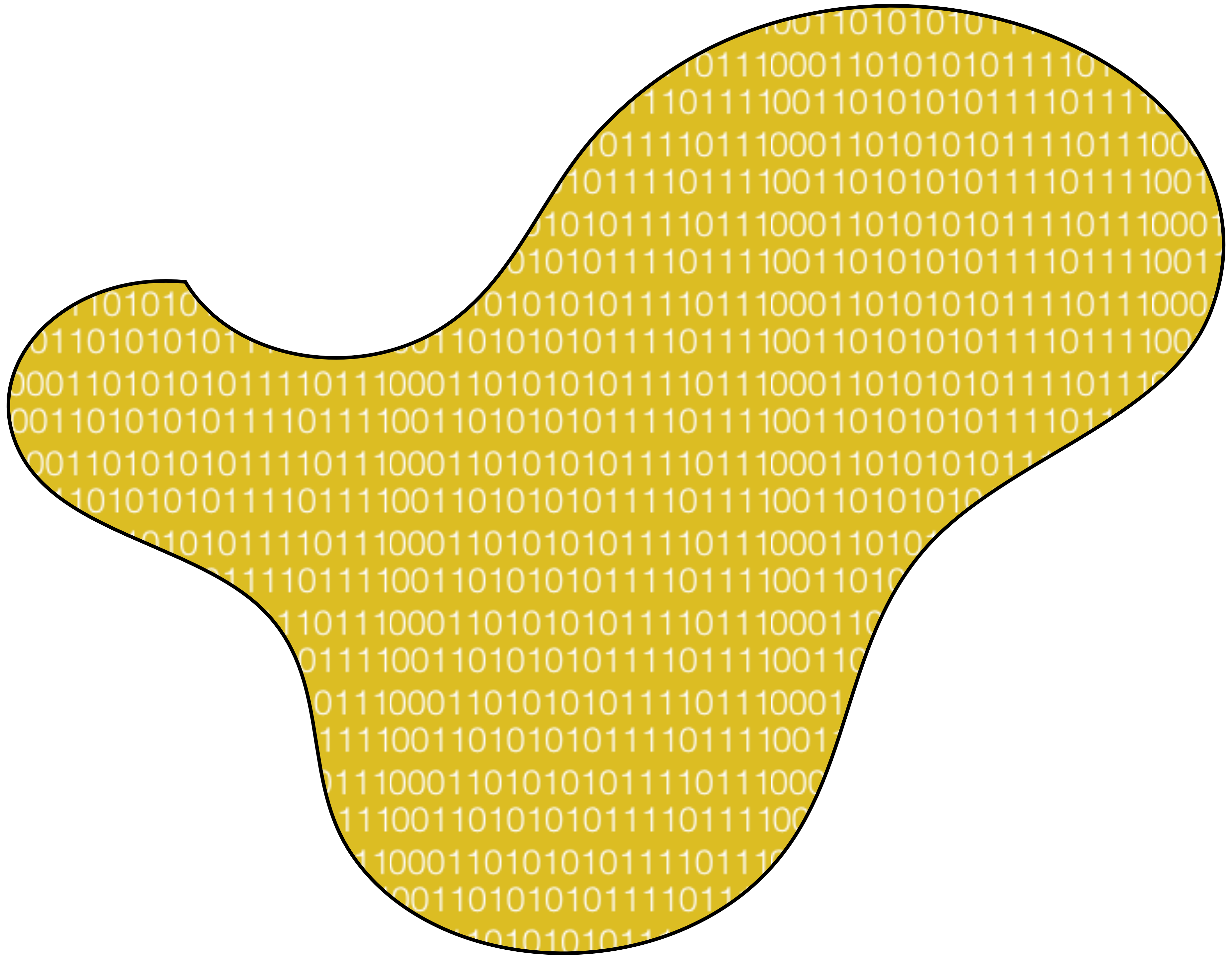
Basic Information Theory: Surprisal

- The **amount of information** in a word (or anything!) depends on how *surprising* it is in context.
- Information content is quantified as **surprisal**:
  - $S(word \mid context) = -\log_2 P(word \mid context)$ (measured in bits)
- Surprisal is also the **length of the shortest binary representation** that encodes the word in context.



A Closer Look at Surprisal

play



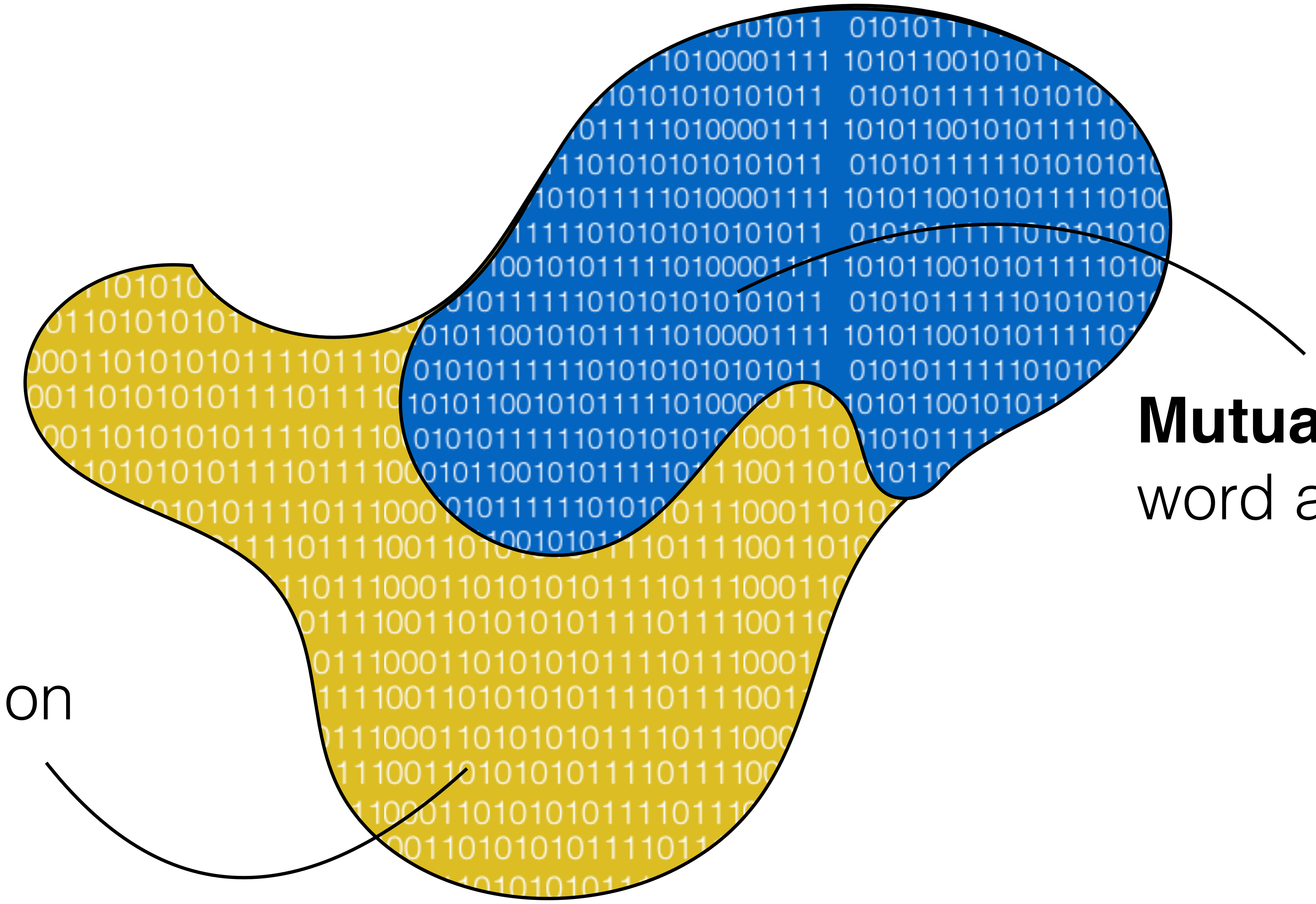
Information content

of **play**

S(play)

A Closer Look at Surprisal

The children went outside to play…



Information content of **play** in **context**

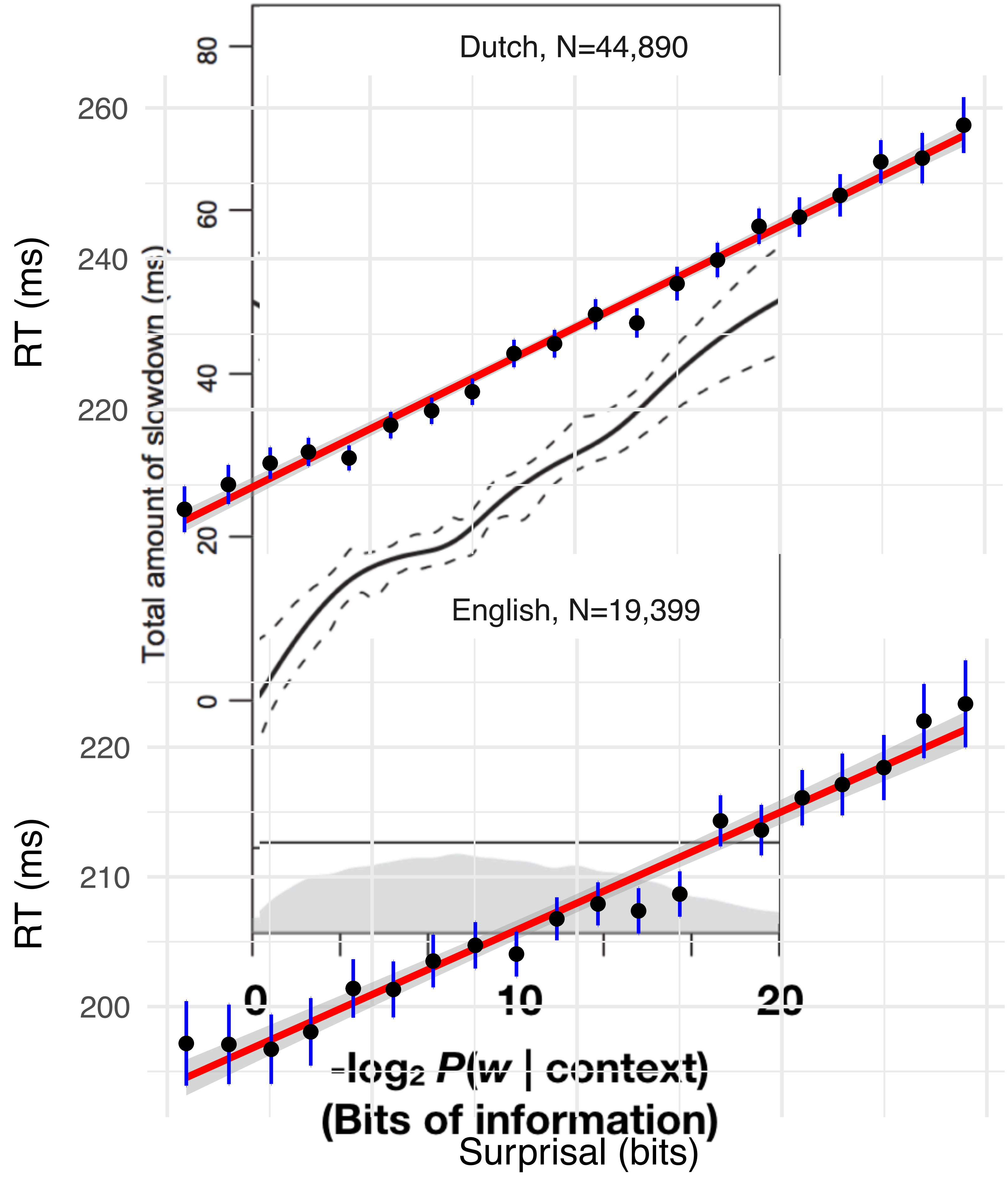
S(play | context)

Information Theory in Psycholinguistics

- **Surprisal Theory**: (Hale, 2001; Levy, 2008; Smith & Levy, 2013)
  - RT(*word* | *context*) = *k* S(*word* | *context*).
- **Idea**: Each bit of information content takes a fixed time for processing.

Information Theory in Psycholinguistics

- **<u>Surprisal Theory</u>**: (Hale, 2001; Levy, 2008; Smith & Levy, 2013)
  - RT(*word* | *context*) = *k* S(*word* | *context*).
- **<u>Idea</u>**: Each bit of information content takes a fixed time for processing.
- Surprisal theory and variants have **high predictive value** for reading times and N400 signals (Smith & Levy, 2013; Frank & Bod, 2011; Frank, 2016; Wilcox et al., 2020; Shain, 2019; Li & Futrell, 2022)
- Predicts classic garden path effects, although underestimating effect size (Hale, 2001; Levy, 2008; but see van Schijndel & Linzen, 2022)



Xu, Clark, Liu & Futrell (2023)

Smith & Levy (2013)

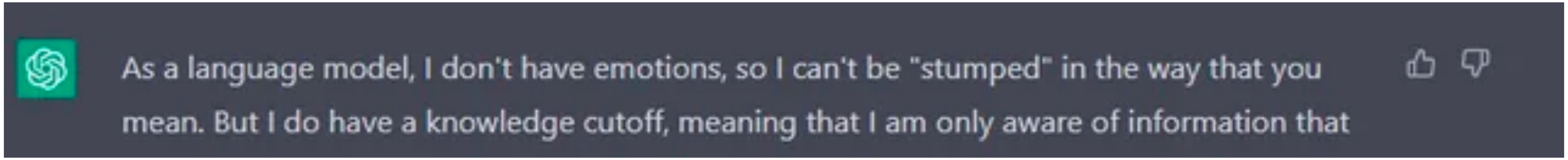
Surprisal and Language Models

- Optimal representations are based on a **predictive language model**

$$S(word \mid context) = -\log P(word \mid context)$$

  - Fitting a language model to predict words in context *is equivalent to* finding optimal compressed representations of words in context.

- What do you get if you train a giant neural network to minimize surprisal?



Information Theory and Language Processing

- Surprisal Theory is a good start, but…
  - It does not account for **memory limitations**, and often **underestimates reaction times.**
  - It does not say anything about how **linguistic structure** interacts with processing difficulty.
  - It's not clear what it has to say about **production**.
- What happens when we consider optimal representations **under cognitive constraints?**

Outline

- Introduction
- Basics of Information-Theoretic Psycholinguistics
- Memory Bottleneck in Language Comprehension
- Control Bottleneck in Language Production
- Conclusion


Michael Hahn

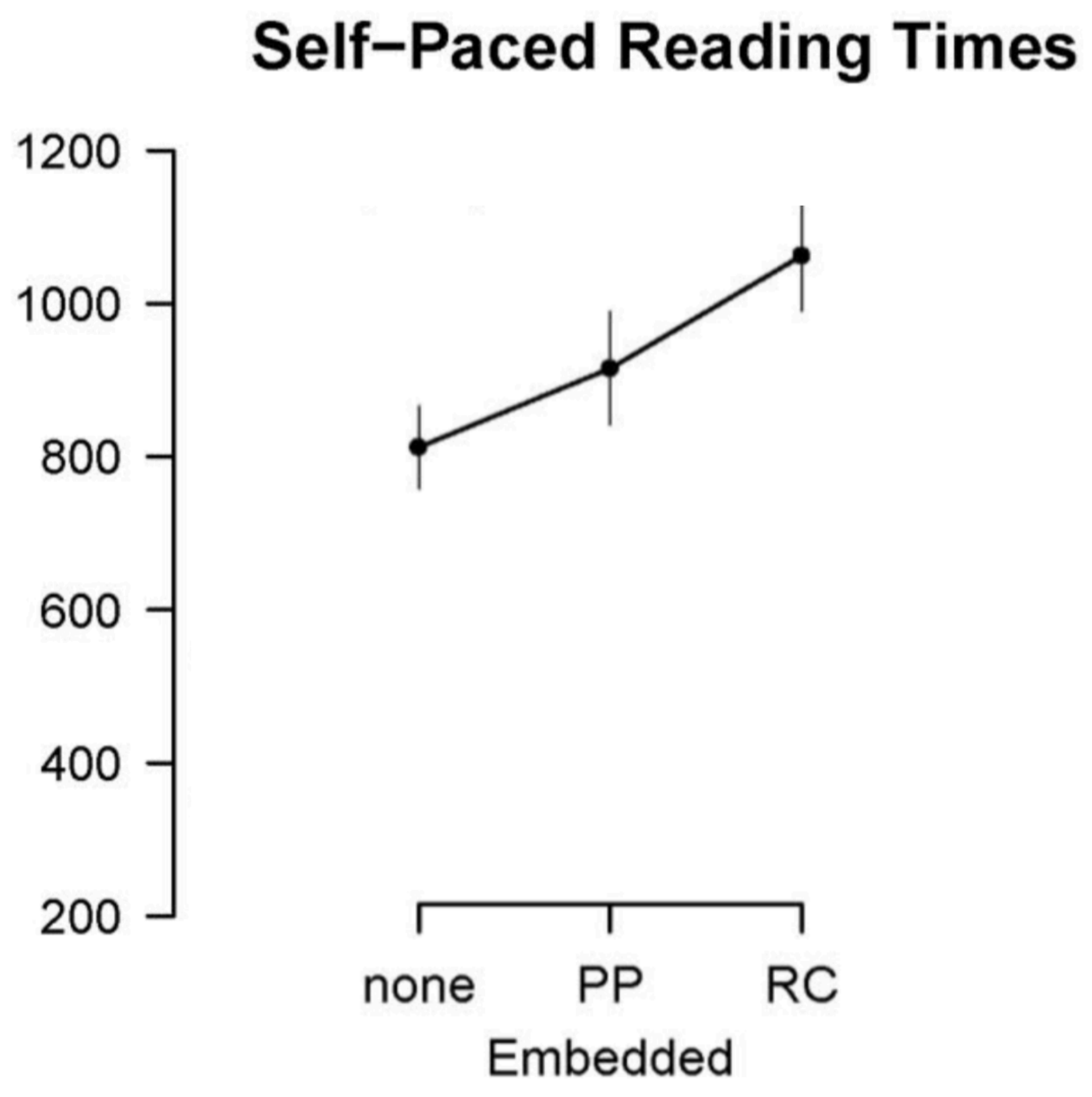

Ted Gibson


Roger Levy

Memory Effects in Sentence Processing

- Bob threw out the trash. 👍
- Bob threw the trash out. 👍
- Bob threw out the old trash that had been sitting in the kitchen. 👍
- Bob threw the old trash that had been sitting in the kitchen out. 👎

The Dependency Locality Theory (Gibson, 1998, 2000)



Bartek et al. (2011)

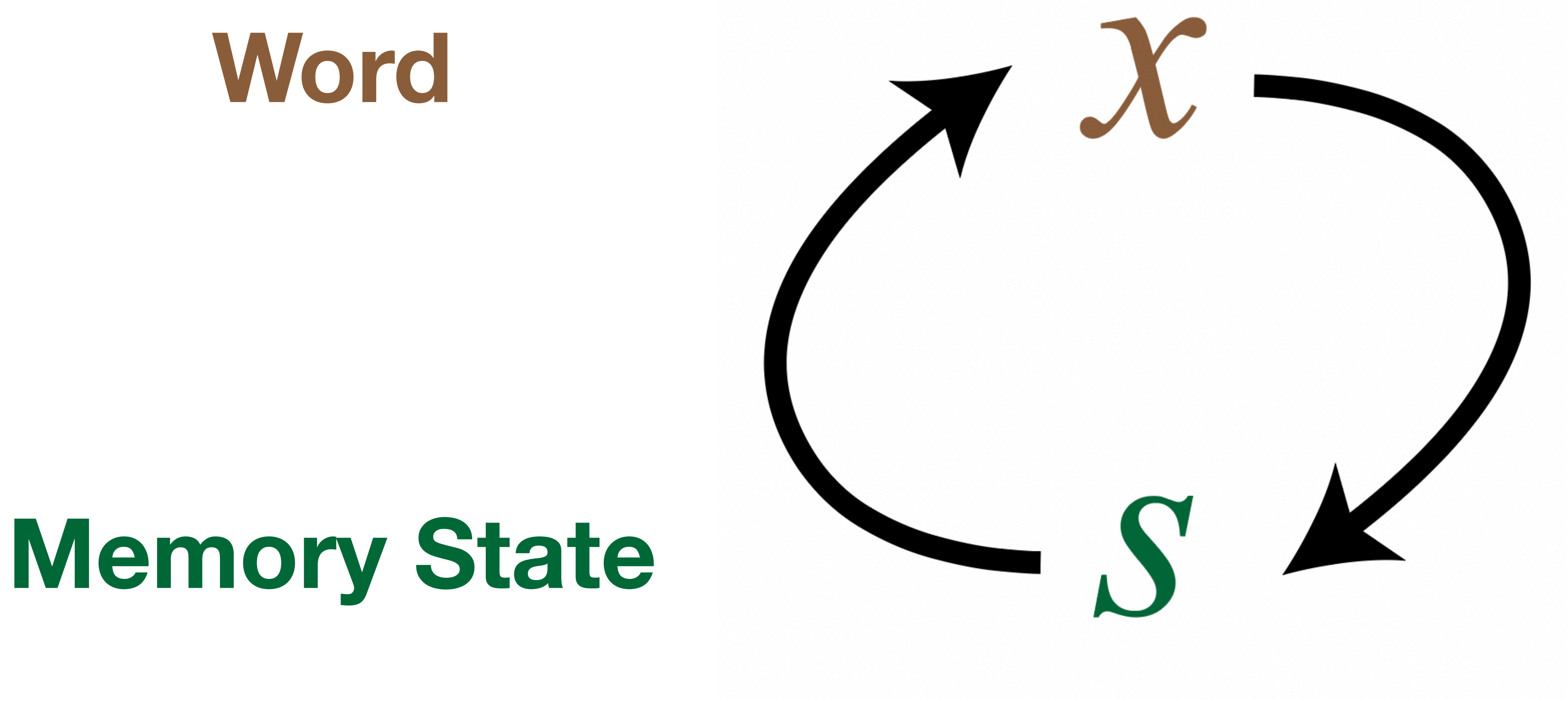
Memory Effects in Sentence Processing

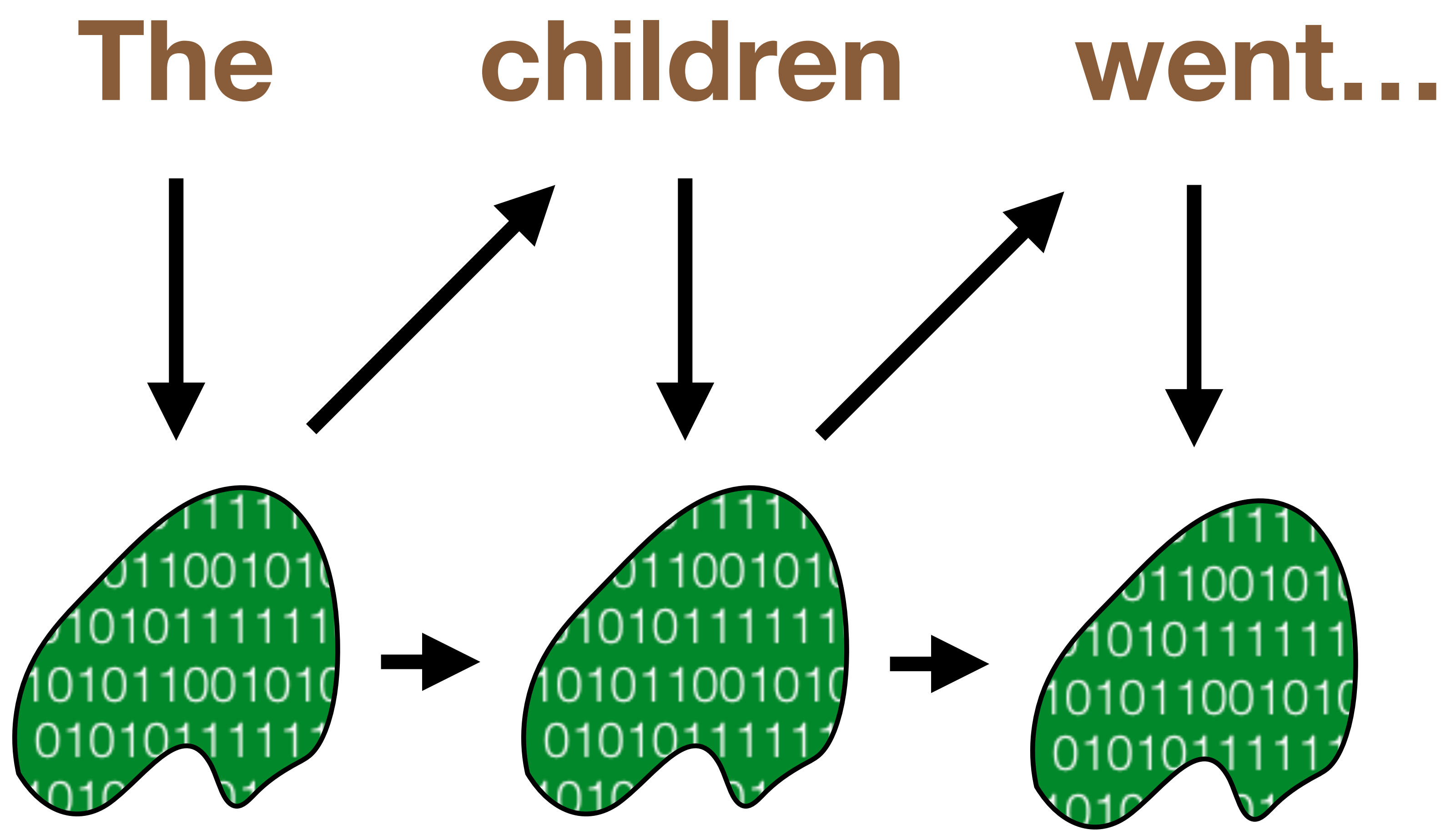
- The apartment was well-decorated. 👍
- The apartment that the maid cleaned was well-decorated. 👍
- The apartment that the maid who the service sent over cleaned was well-decorated. 👎👎👎👎👎

RT trouble starts here

Vasishth et al. (2010)

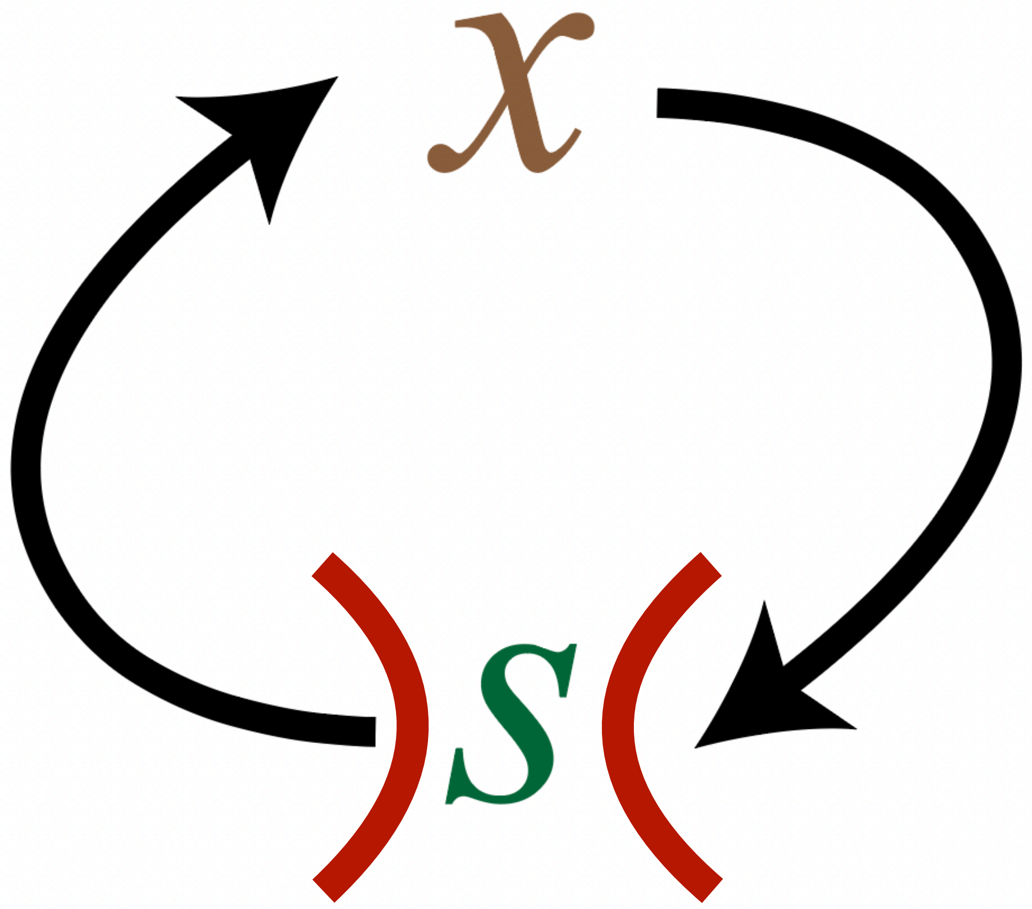
Memory in Language Comprehension





Memory in Language Comprehension

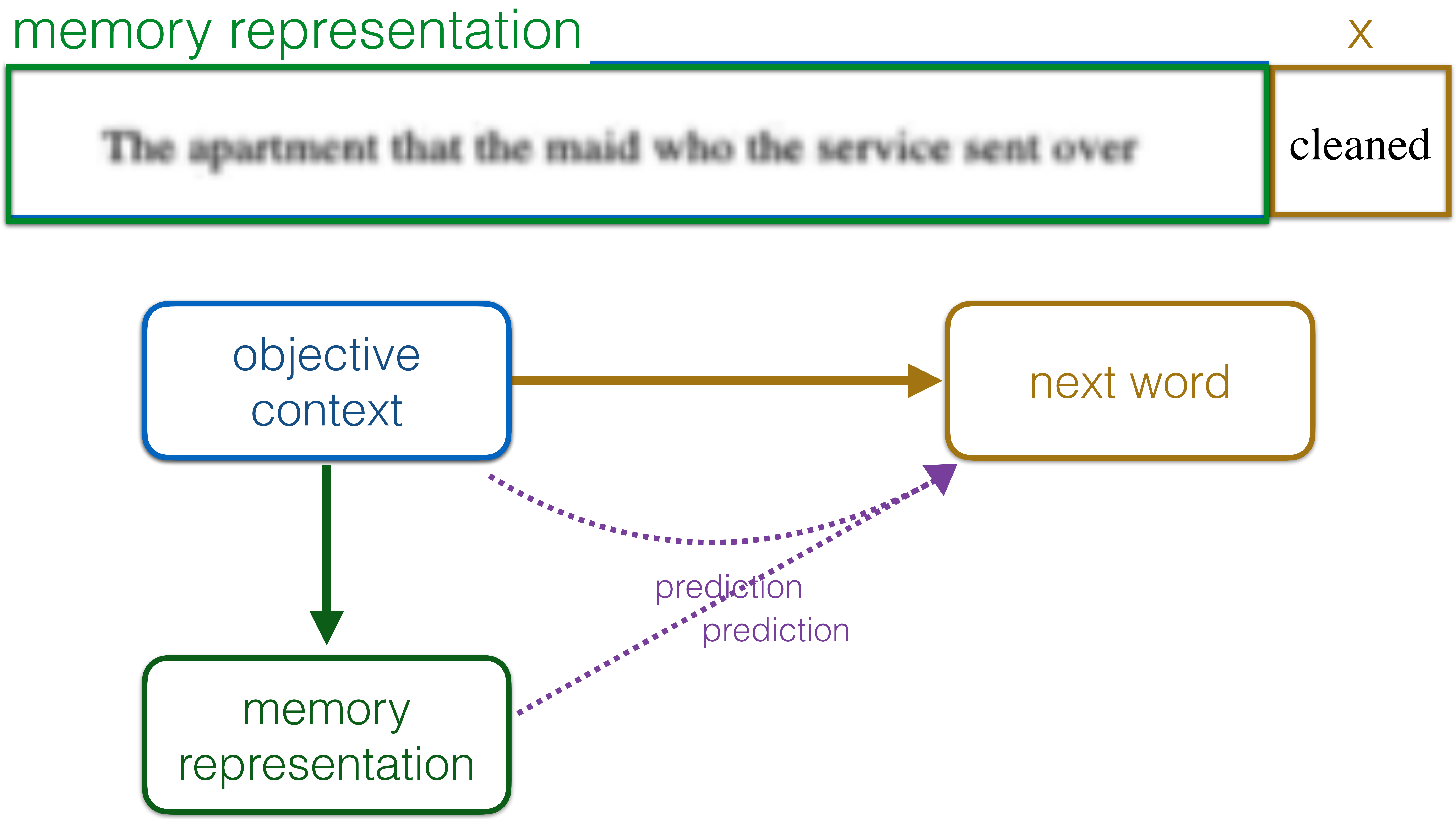
Word



Memory State

How to fit a memory bottleneck into Surprisal Theory?

- Surprisal: RT(w | context) = S(w | context)
- Lossy-context surprisal: RT(w | context) = S(w | memory representation)



Futrell, Gibson & Levy (2020)

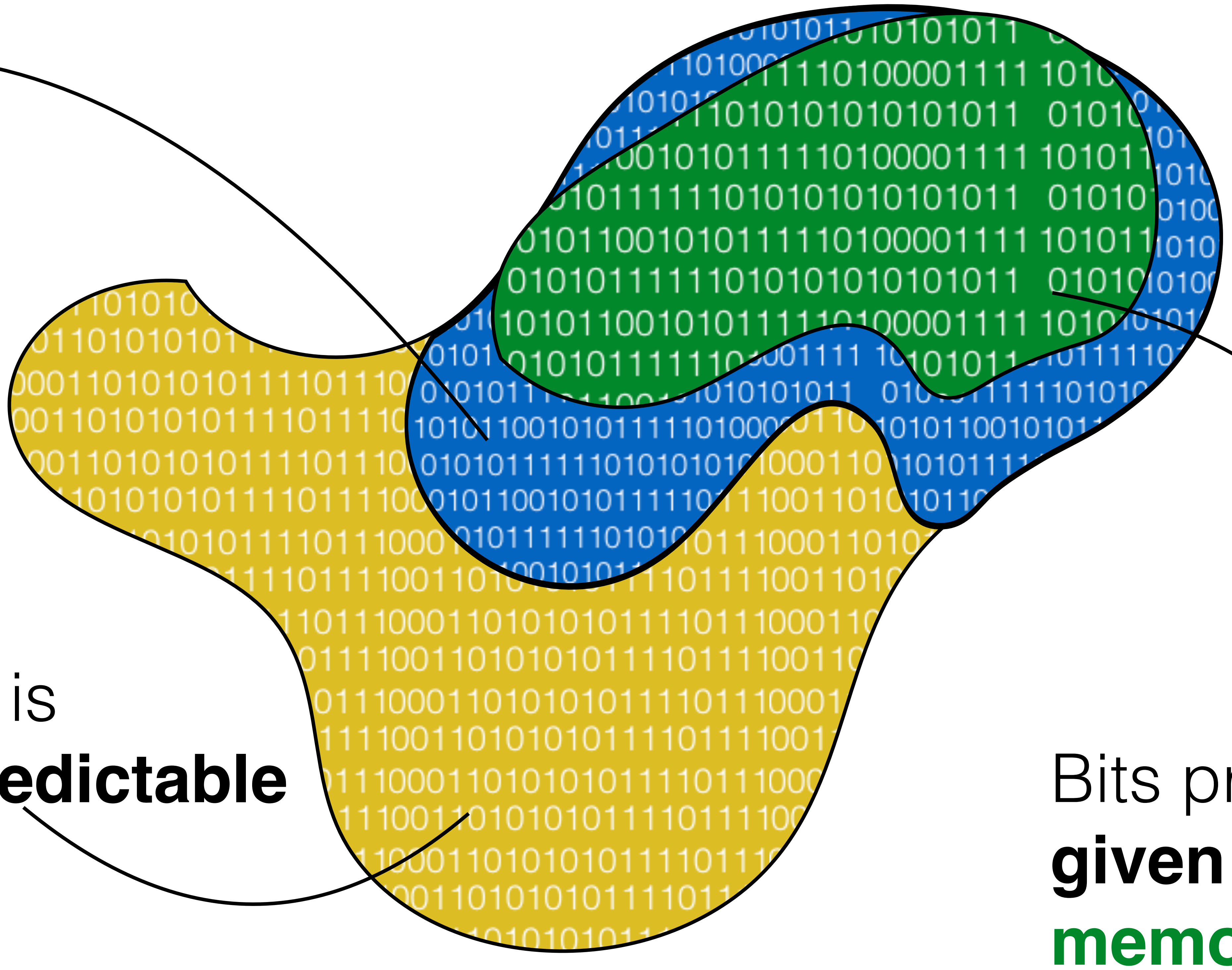
Lossy-Context Surprisal

The maid cleaned...

Lossy-context surprisal **S(cleaned | The maid)**

Objective surprisal **S(cleaned | memory)**

Memory cost due to memory limitations

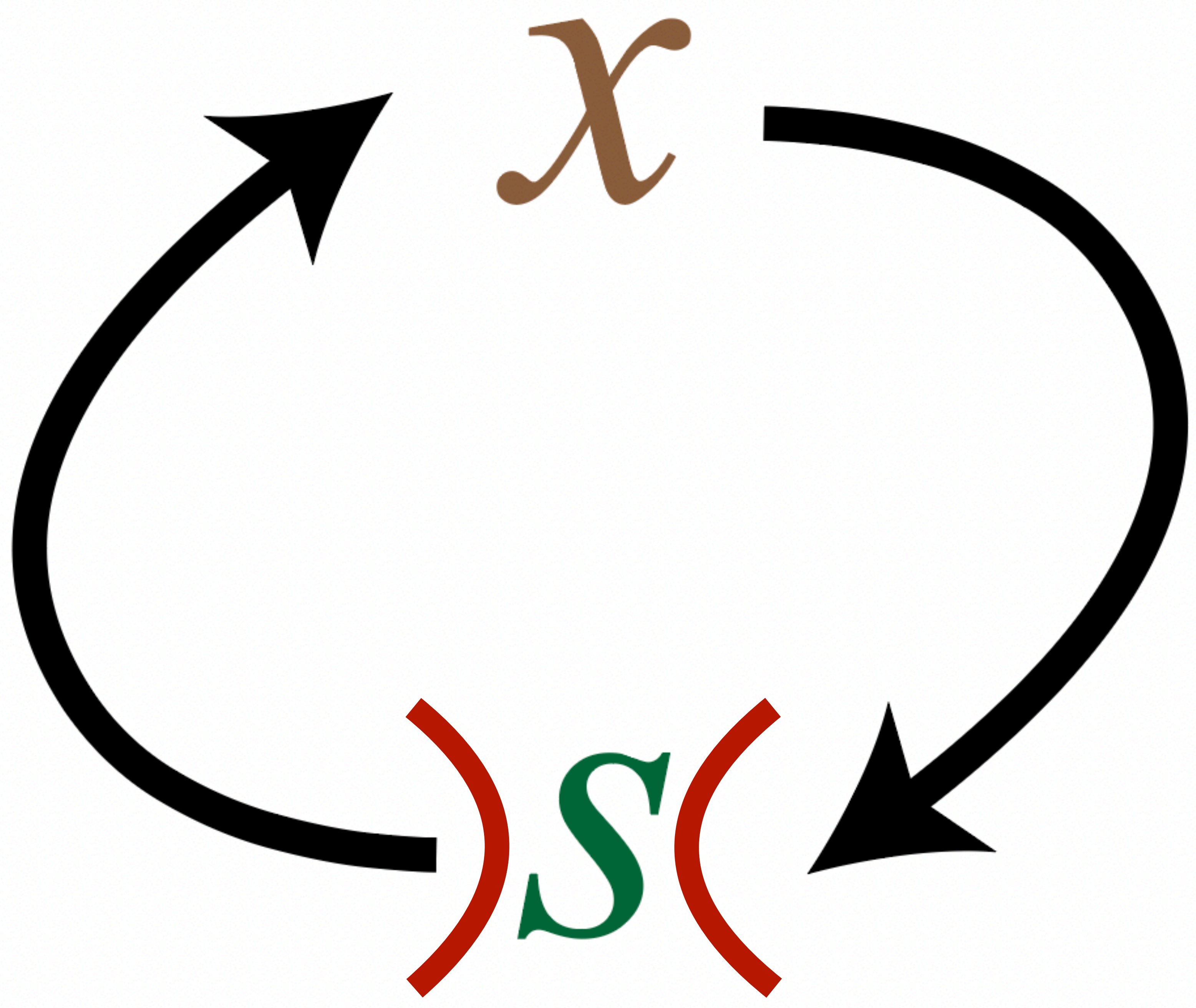


Processing difficulty is the **number of unpredictable bits**.

Bits predictable **given the memory state**

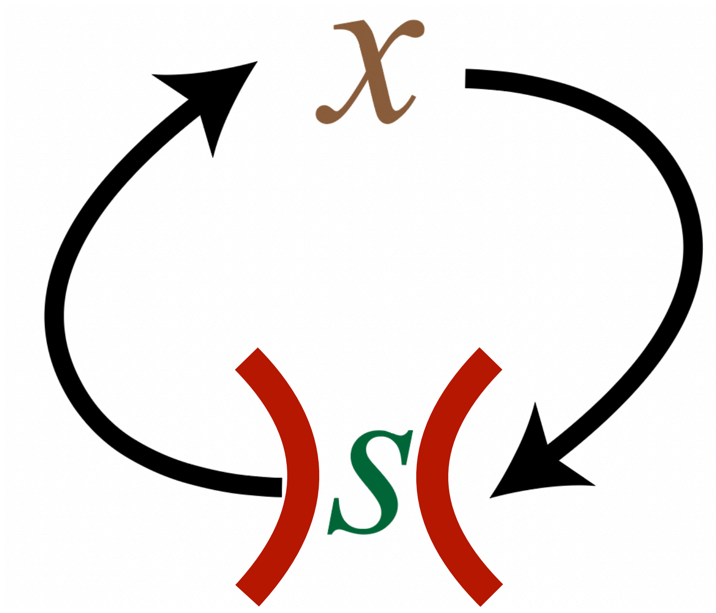
$$S(\textit{word} \mid \textit{memory}) = S(\textit{word} \mid \textit{context}) + \text{Memory cost}$$

Uses of Lossy-Context Surprisal



- By constraining memory in various ways, we can account for…
  - Certain **dependency locality effects** (Futrell, Gibson & Levy, 2020)
  - Cross-linguistic patterns in **structural forgetting** (Futrell, Gibson & Levy, 2020)
  - General **reading times** in eyetracking corpora, with neural network implementation (Kuribayashi et al., 2022)
  - Novel patterns in comprehension of **nested clauses.** (Hahn, Futrell, Levy & Gibson, 2022)

Processing with Constrained Memory



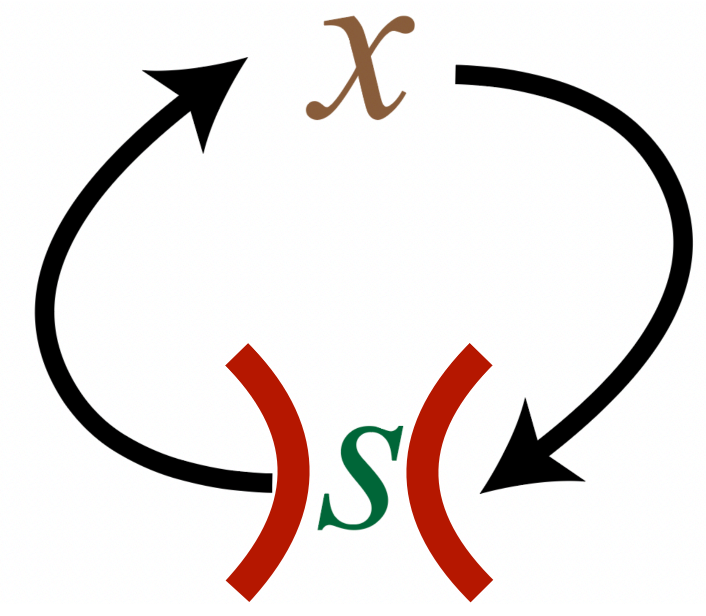
- Idea: Only a certain maximum number of words can be retained in memory.
- Predictions about upcoming words are **optimal subject to the constraint** that not all context words can be represented.

Context
The report that the doctor annoyed the patient…

was interesting

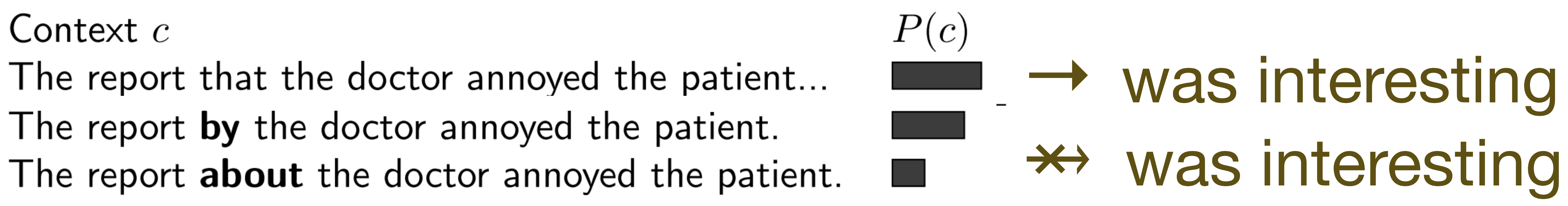
Hahn, Futrell, Levy & Gibson (2022)

Processing with Constrained Memory



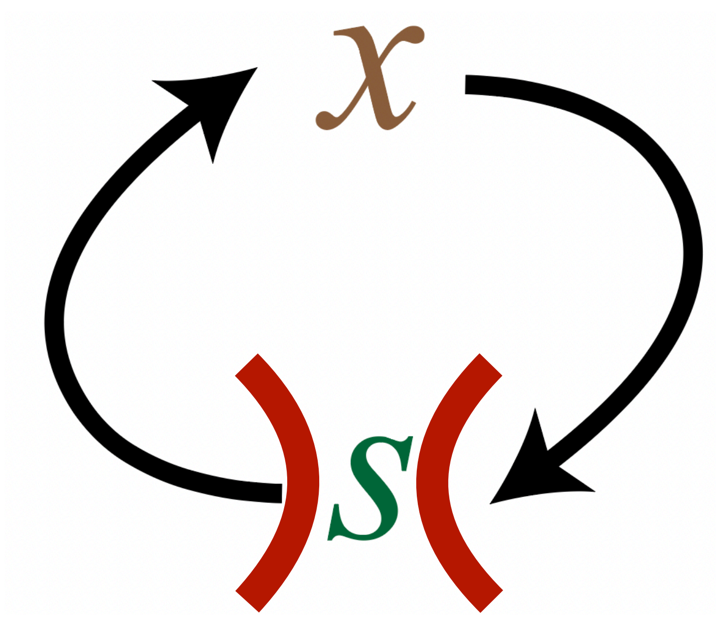
- Idea: Only a certain maximum number of words can be retained in memory.
- Predictions about upcoming words are **optimal subject to the constraint** that not all context words can be represented.

Lossy Context
The report ??? the doctor annoyed the patient…



Hahn, Futrell, Levy & Gibson (2022)

Predictions about Embedded Clauses



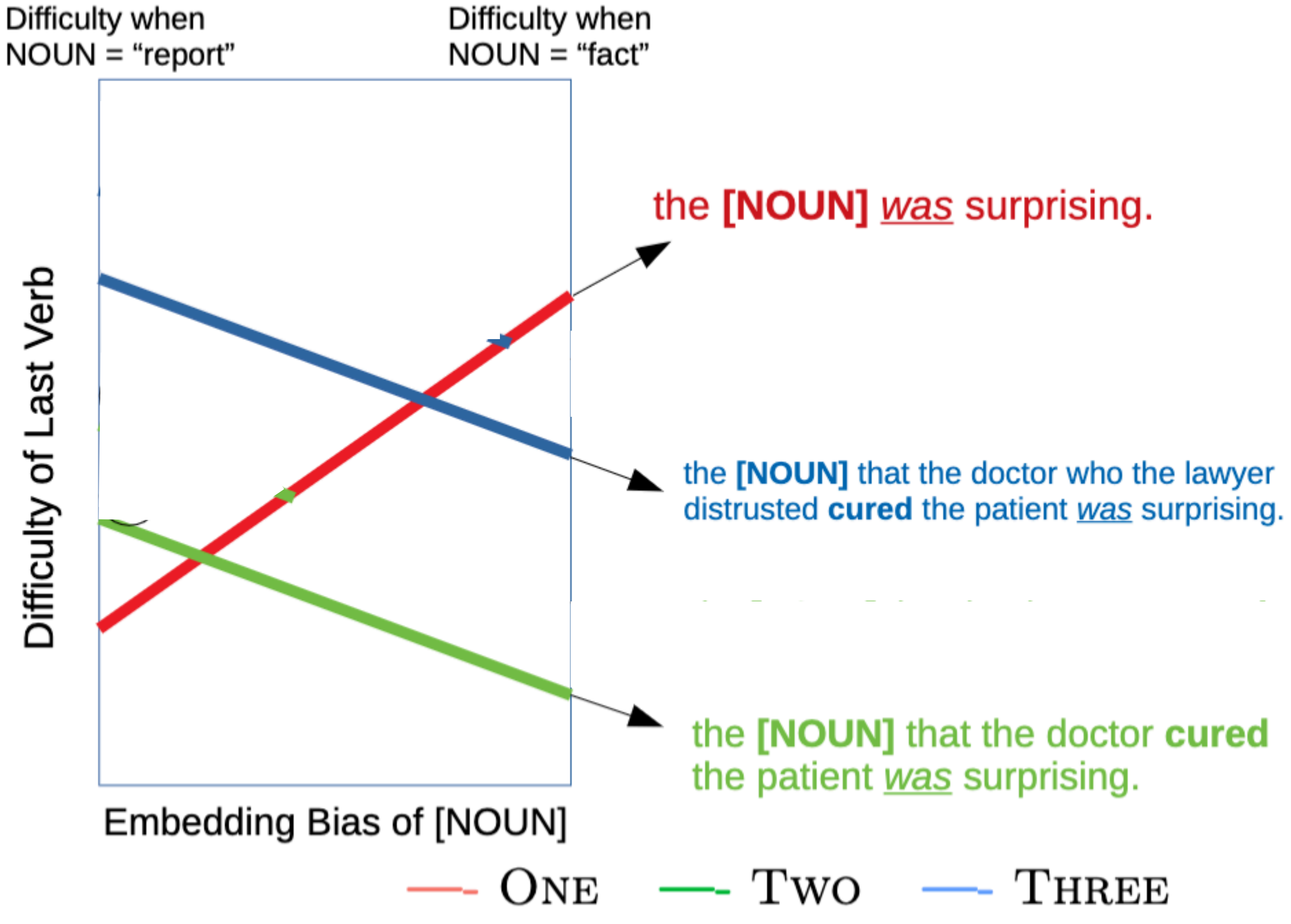
- <u>Prediction</u>: The difficulty of multiple embedding depends on the **embedding bias** of the noun.

Low Embedding Bias

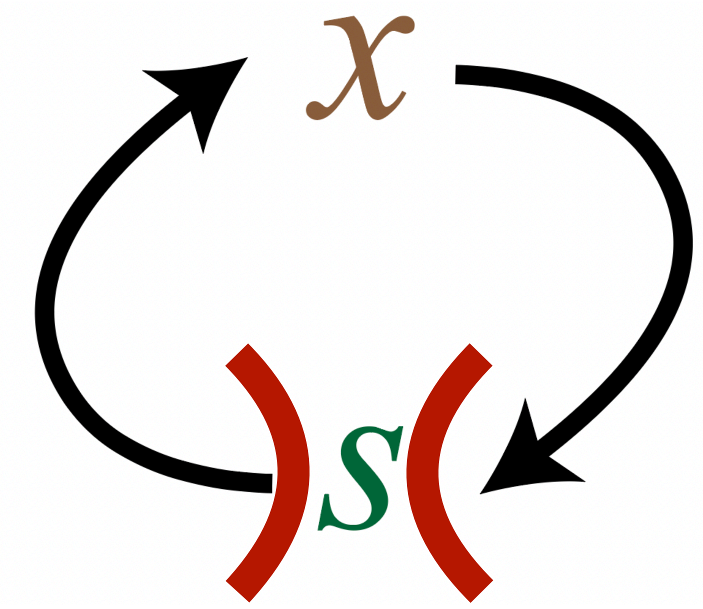
	Context c	$P(c)$
True (c^*)	The report that the doctor annoyed the patient...	
Variants	The report **by** the doctor annoyed the patient.	
	The report **about** the doctor annoyed the patient.	
	...	

High Embedding Bias

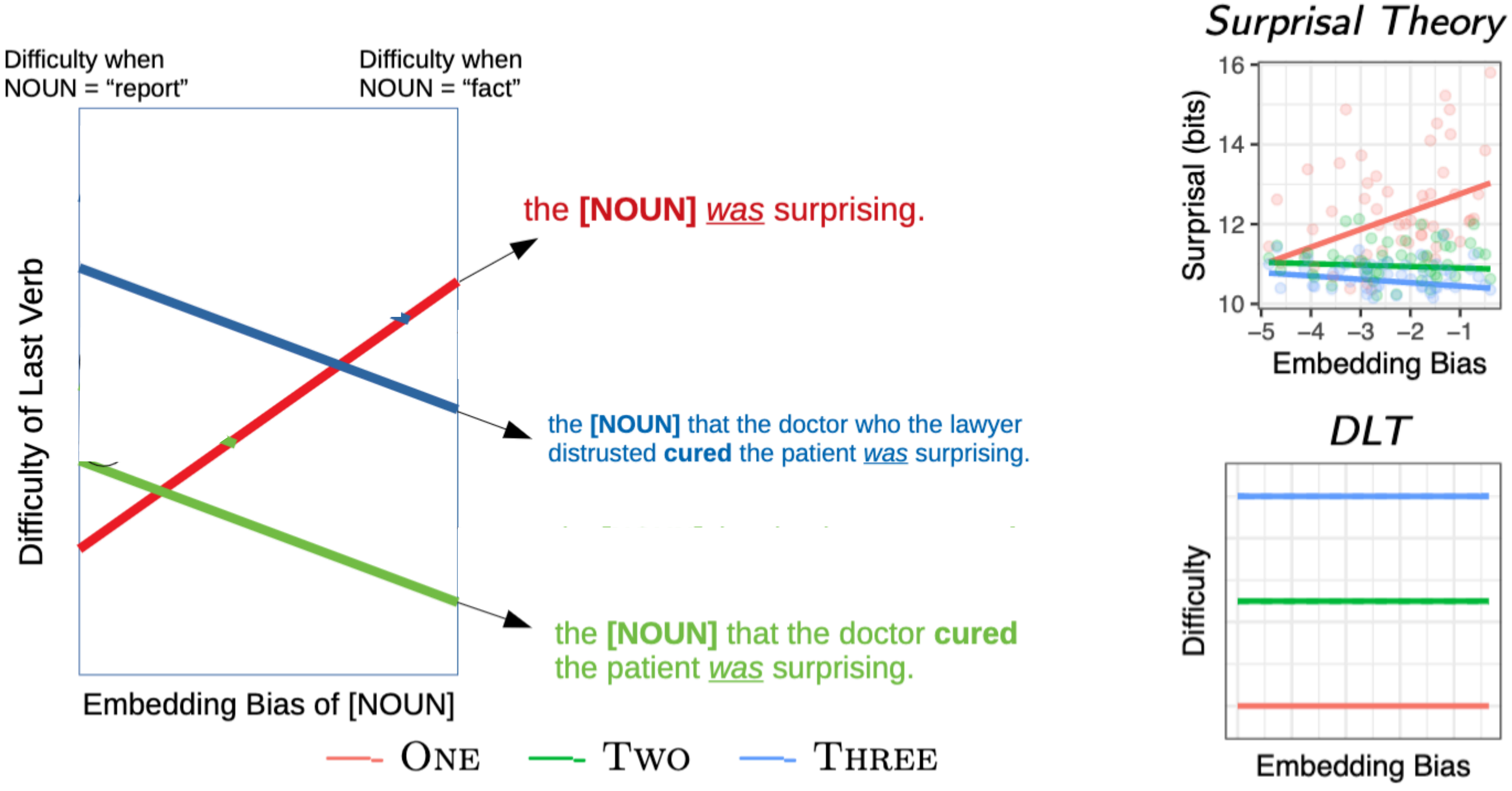
	Context c	$P(c)$
True (c^*)	The **fact** that the doctor annoyed the patient...	
Variants	The **fact of** the doctor annoyed the patient.	
	The **fact about** the doctor annoyed the patient.	
	...	



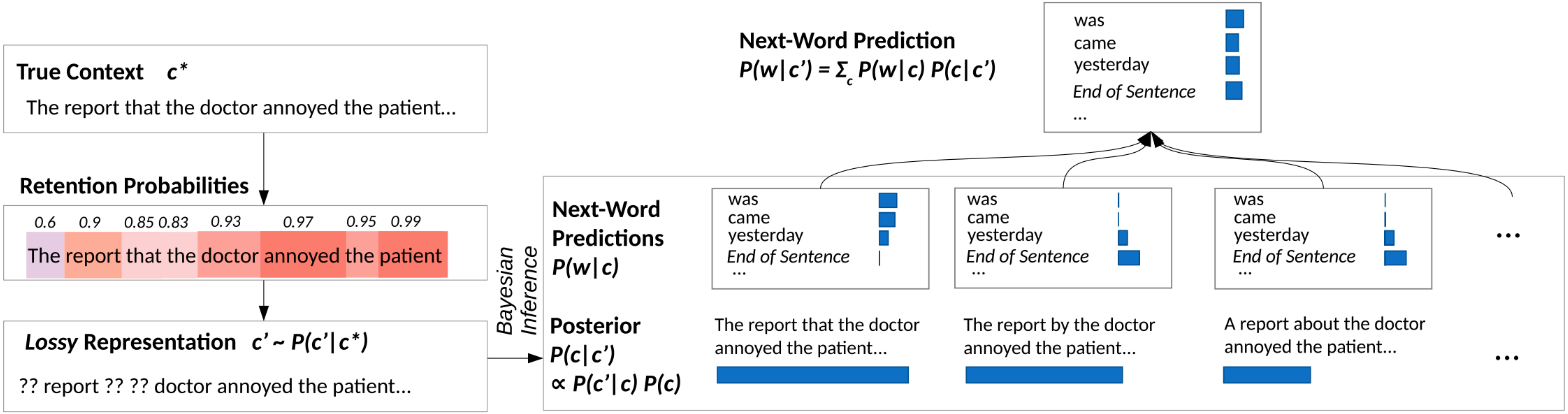
Predictions about Embedded Clauses



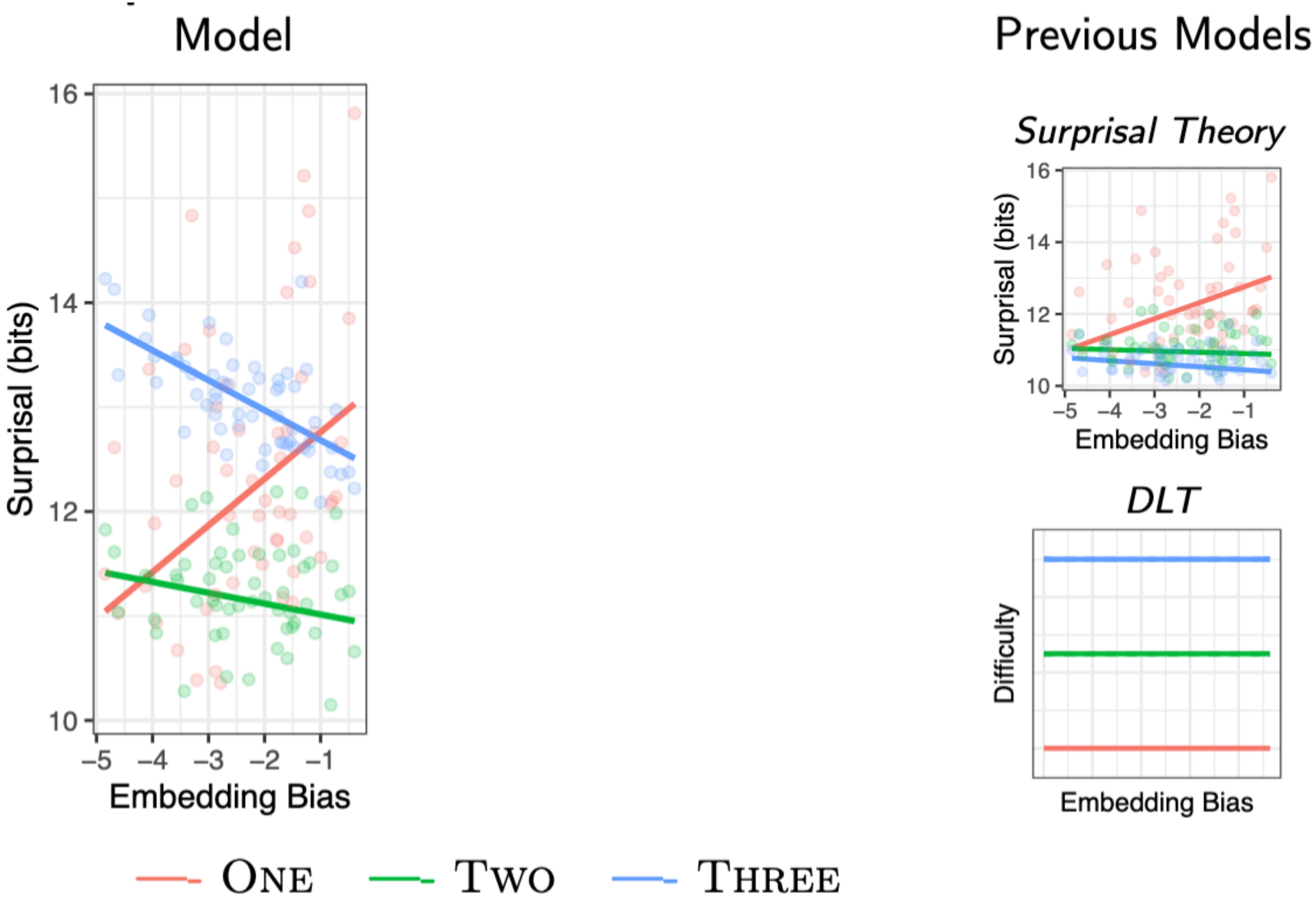
- <u>Prediction</u>: The difficulty of multiple embedding depends on the **embedding bias** of the noun.



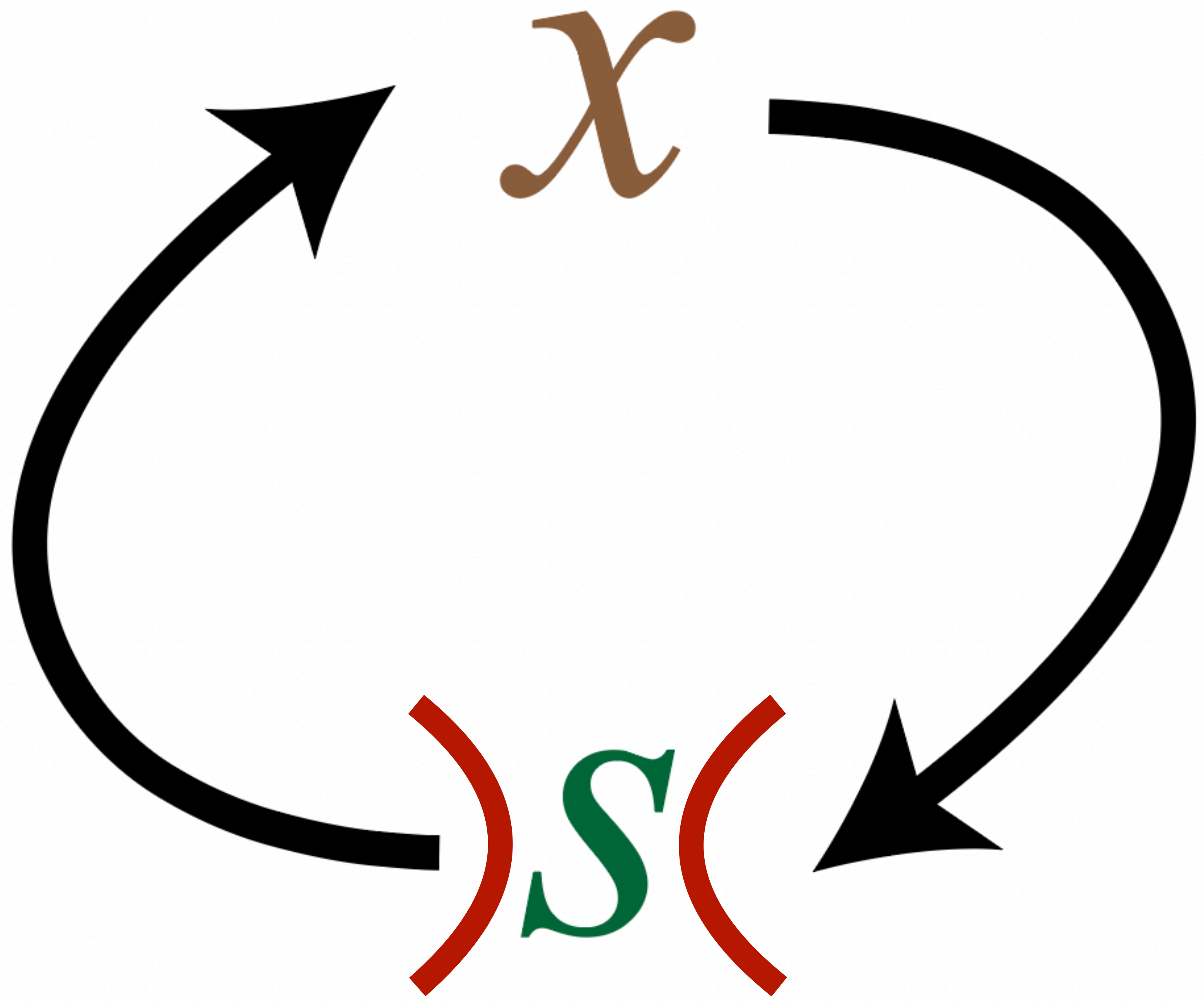
Model Implementation



Reading Time Experiment Results



Memory Bottleneck in Language Comprehension



- We considered language comprehension difficulty based on **surprisal** given a **lossy memory representation of context**.
- Predicts RT **better than a less constrained language model**.
- Comprehension can be modeled as **maximally efficient subject to memory constraints.**

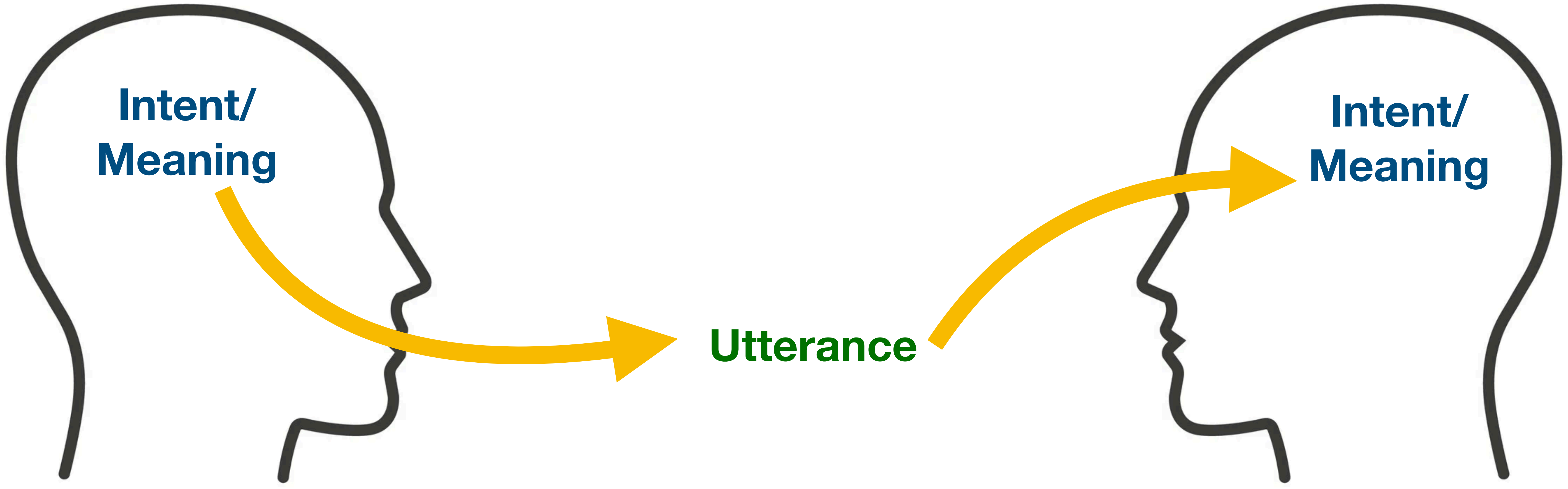
Outline

- Introduction
- Basics of Information-Theoretic Psycholinguistics
- Memory Bottleneck in Language Comprehension
- Control Bottleneck in Language Production
- Conclusion

Information Theory and Language Production

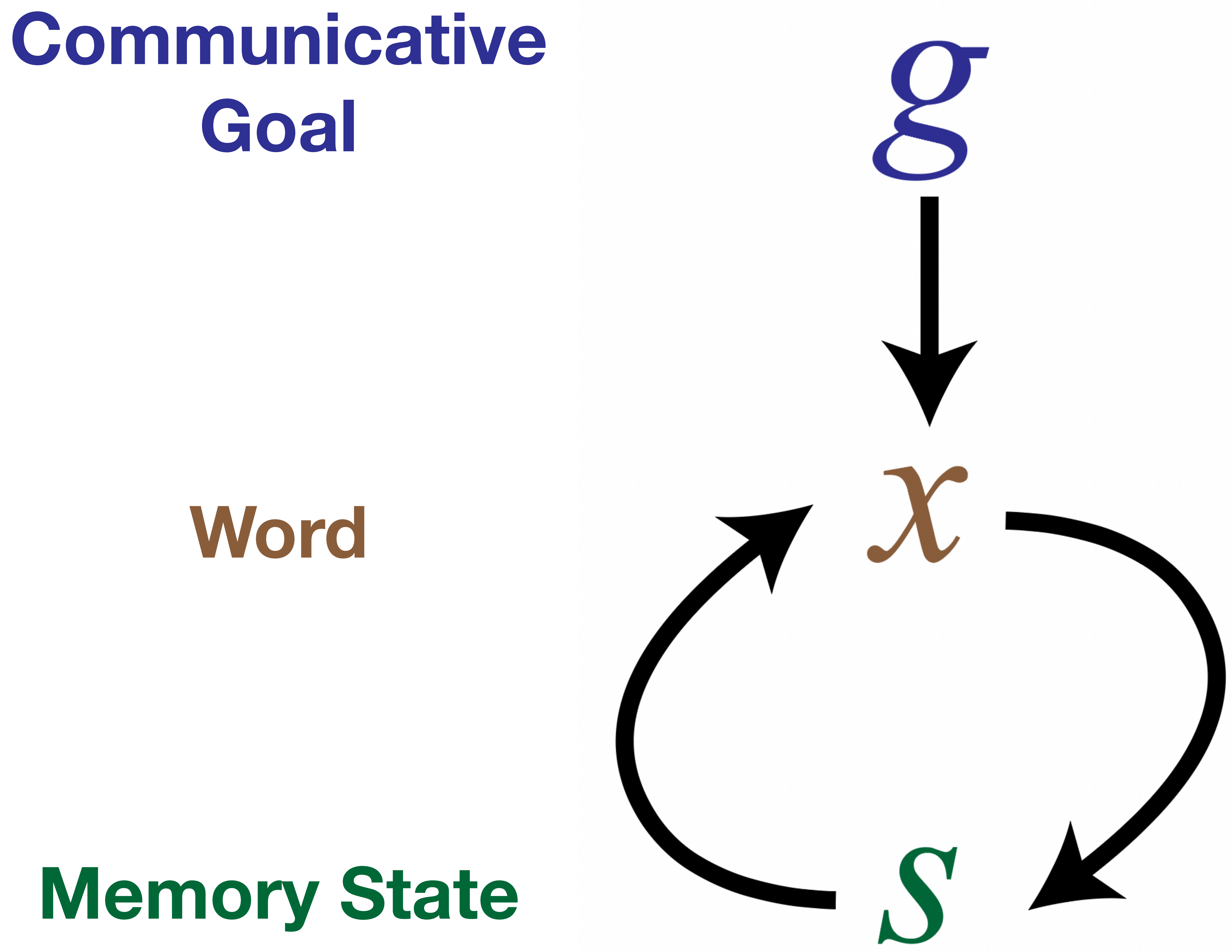
- Information-theoretic models of language processing have mostly focused on **comprehension**.

- What can we say about **production**?



Production **Comprehension**

From Comprehension to Production



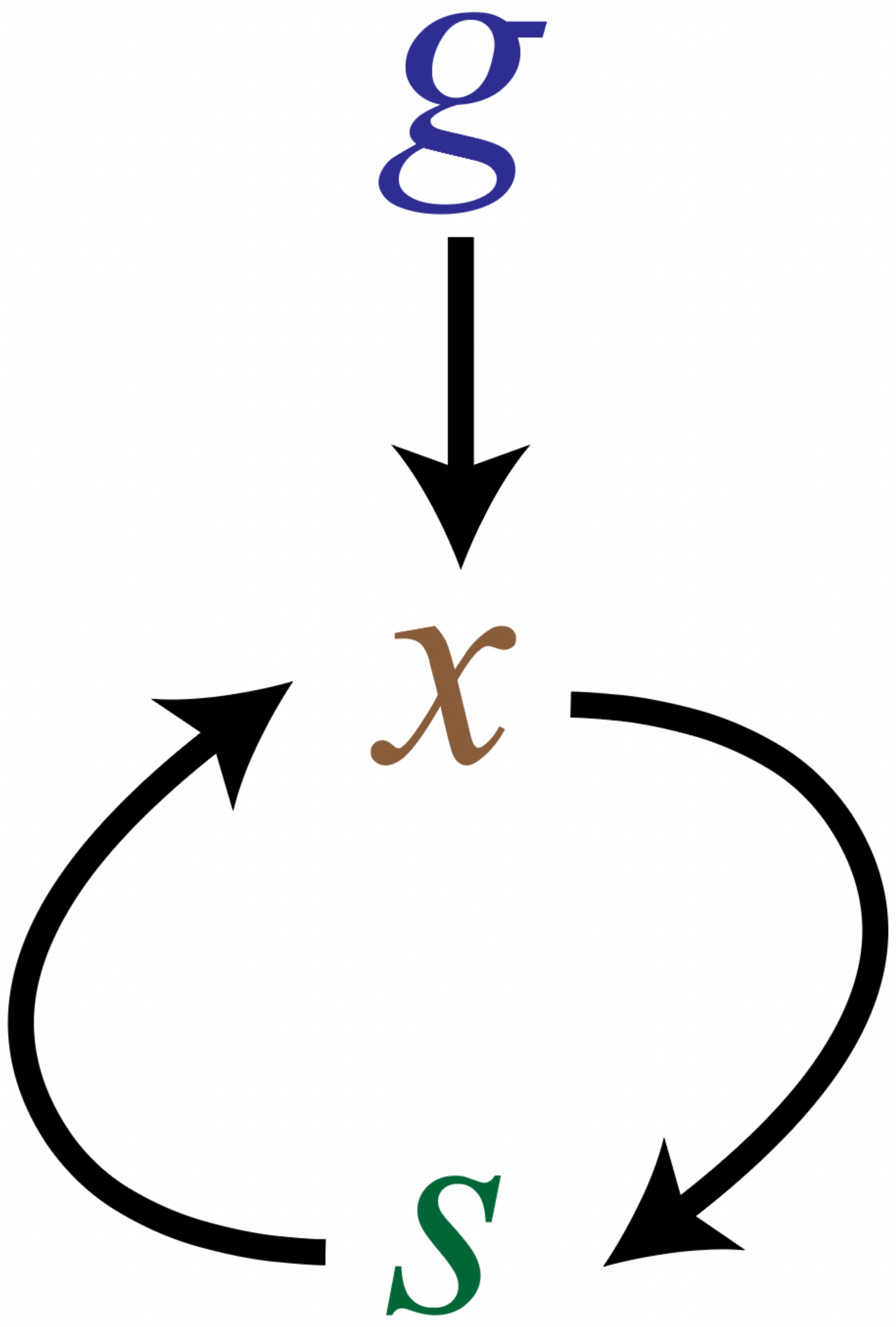
Picture of Language Production

Goal =

Word x cat $\sim P(\cdot \mid g, s)$ **(Policy)**

State = the

Picture of Language Production



Goal $=$ 

Word sat $\sim P(\cdot \mid g, s)$ **(Policy)**

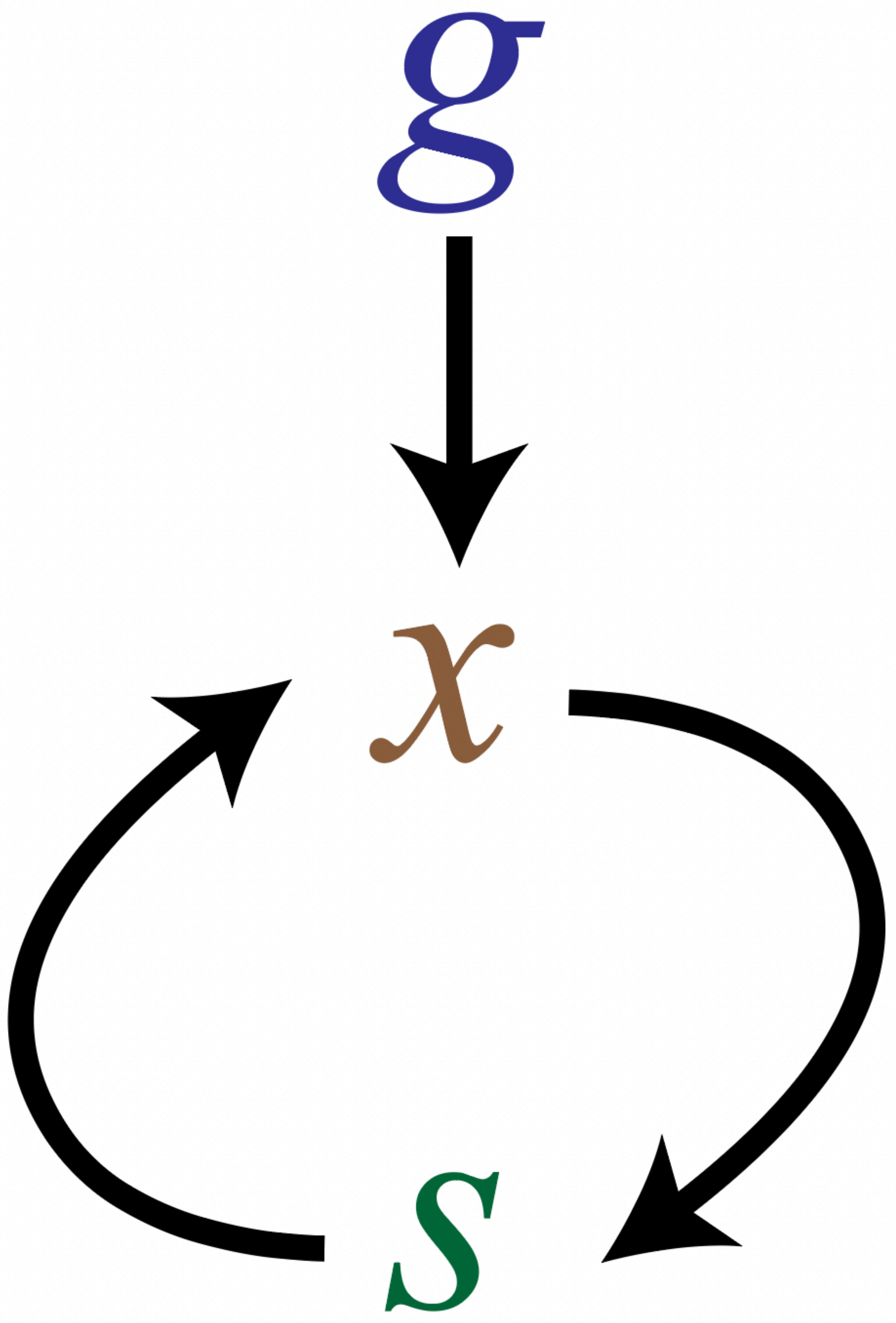
State $=$ the cat

Picture of Language Production

Goal

Word

State



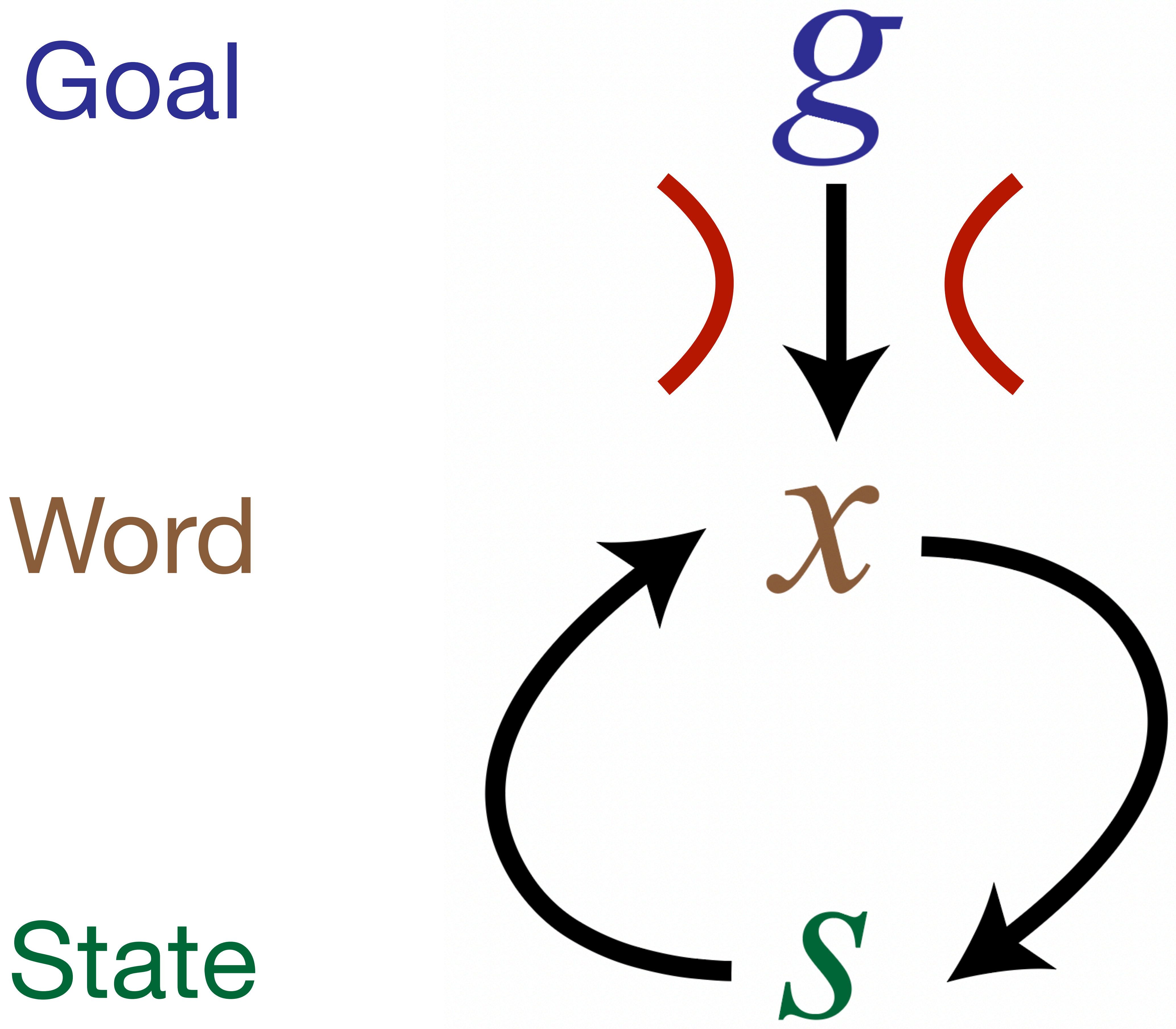
=

on $\sim P(\cdot \mid g, s)$ **(Policy)**

=

the cat sat

Picture of Language Production

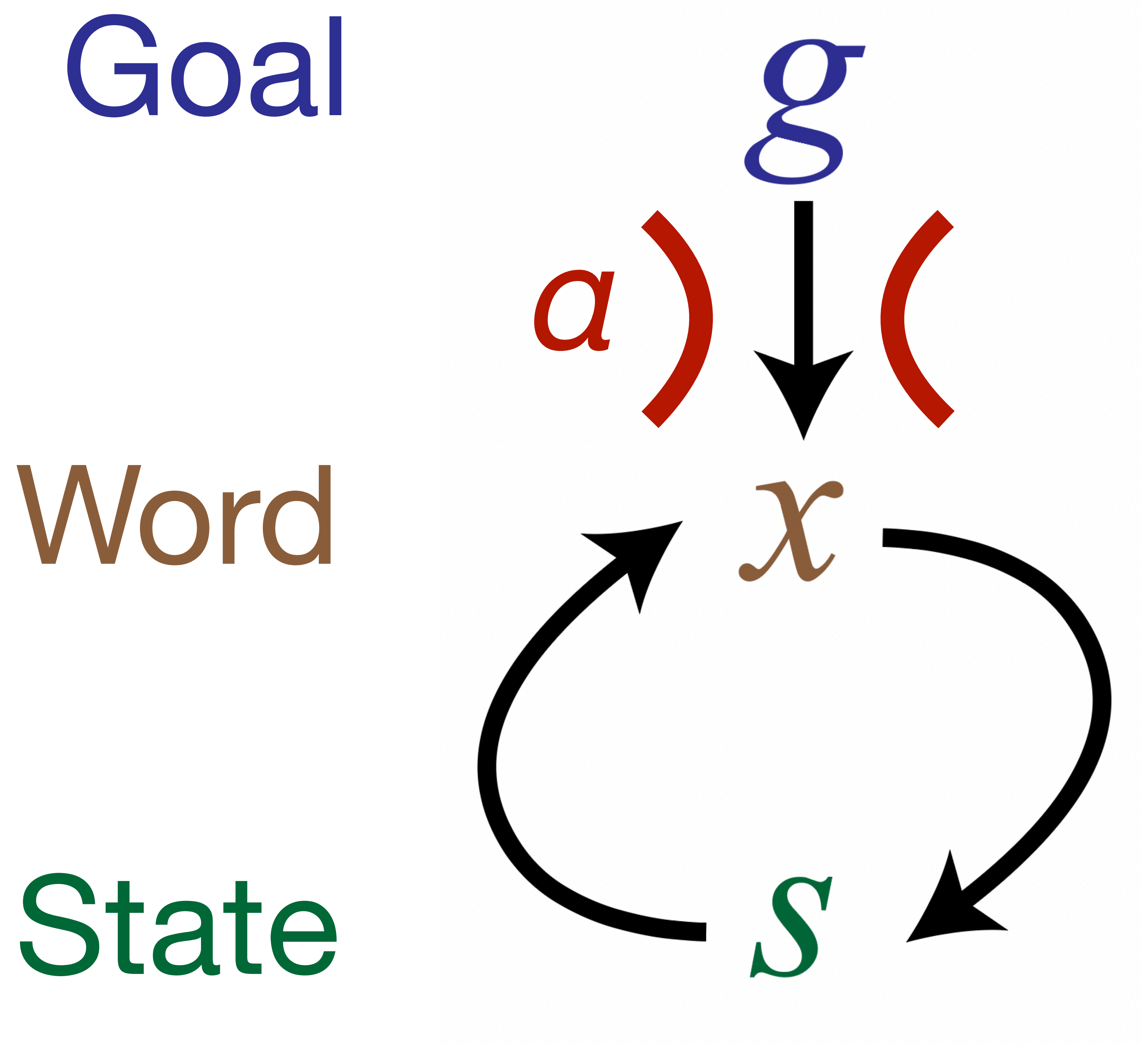


=

$\sim P(\cdot \mid g, s)$ **(Policy)**

= the cat sat on

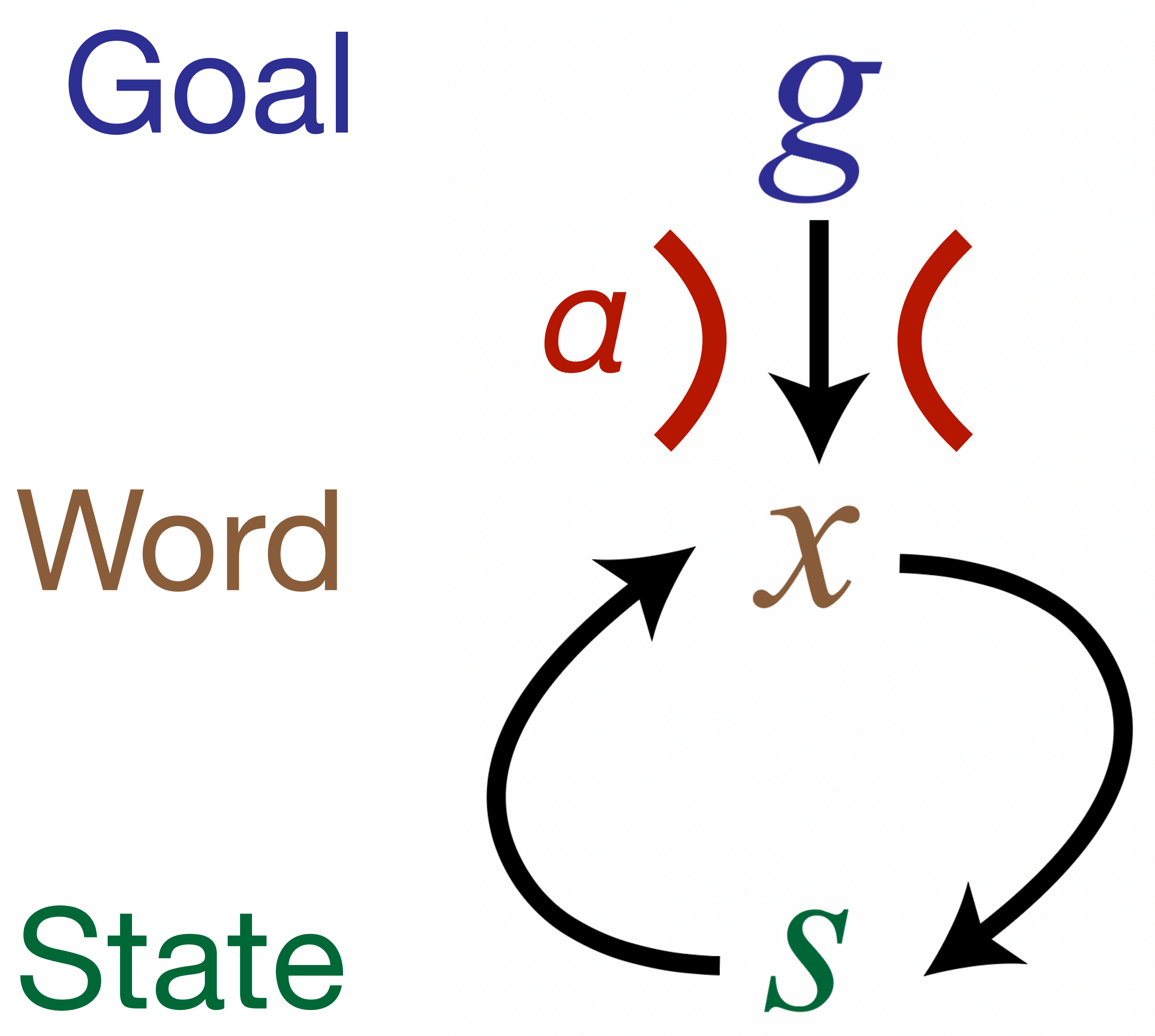
Optimization Problem for Language Production



- Idea: You can only use so much information about the goal per word, due to a constraint on **cognitive control**.
  - Cognitive control operates under a **bandwidth constraint**: 50 bits/ms (Fan, 2014; Zénon et al., 2019)
- So, find a policy that
  - Maximizes **communicative accuracy**
  - Subject to a **constraint on** the **mutual information** of g with x in each timestep.

Futrell (2023)

Constrained Optimal Policy



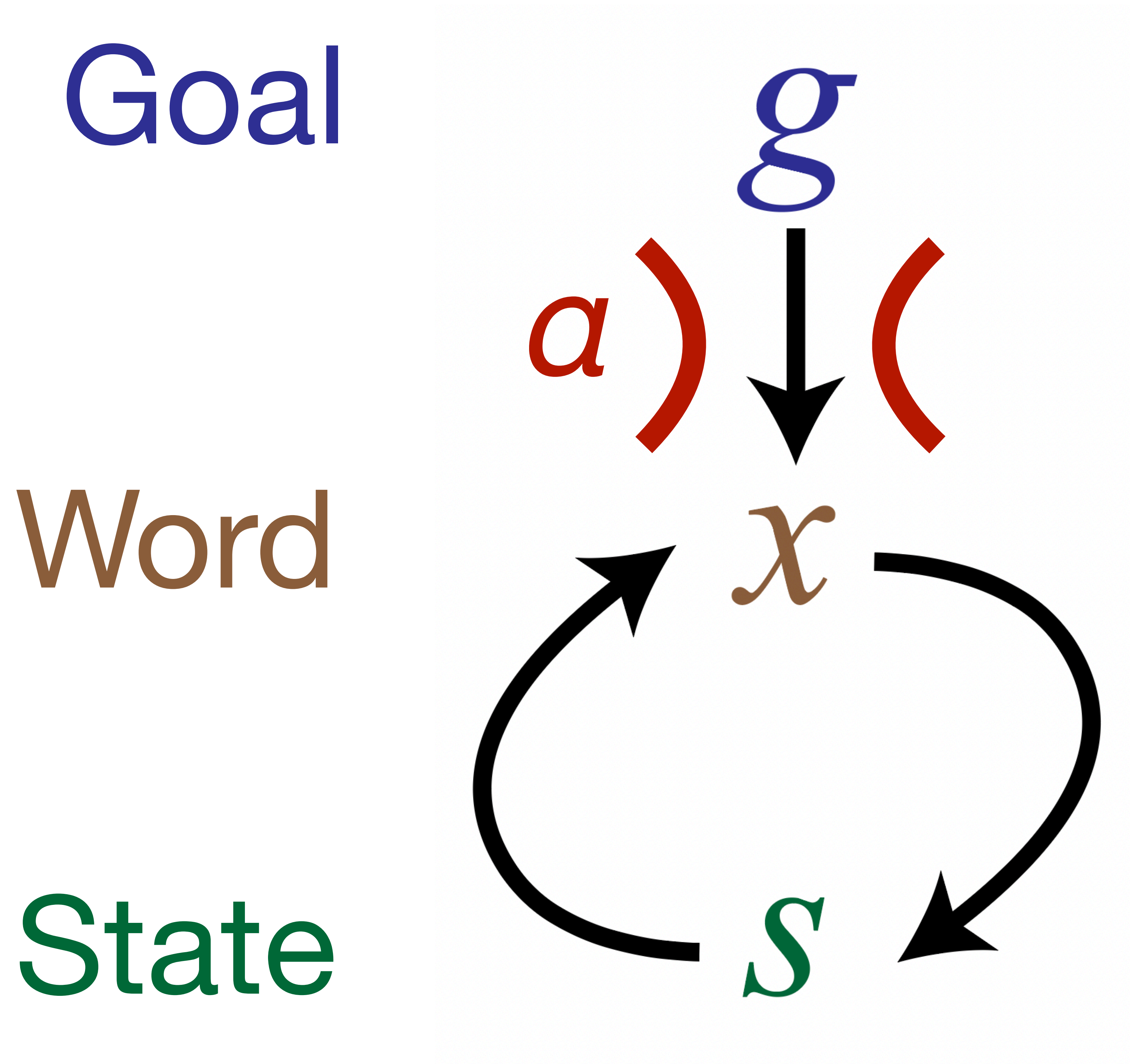
Policy | Surprisal | Control Signal

$$P(\textbf{word} \mid \textbf{goal}, \textbf{state}) \propto \exp\left\{\log P(\textbf{word} \mid \textbf{state}) + \alpha u(\textbf{word} \mid \textbf{goal}, \textbf{state})\right\}$$

- A word is produced if…
  - It is **low surprisal** given the memory state.
  - It is **communicatively accurate**.
- The trade-off of these factors is controlled by the bandwidth of cognitive control, α.

Futrell (2023)

Uses of the Rate-Distortion Theory of Control



- We can use this production model to explain…
  - **Frequency** and **semantic interference** effects in word production (Futrell, 2020; Futrell, 2023, PNAS)
  - **Semantic substitution errors** (Upadhye & Futrell, 2022) and use of **filled pauses** (Futrell, 2023, PNAS)
  - **Accessibility effects** in use of optional complementizers in English (Futrell, 2023, CogSci)
  - **Accessibility effects** in use of Mandarin classifiers (Futrell, 2023, CogSci)

Mandarin Classifiers

- In certain phrases, Mandarin nouns must be preceded by a **classifier** which can be either **specific** or **generic.**

一 台 电脑
one MACHINE computer
'one computer'

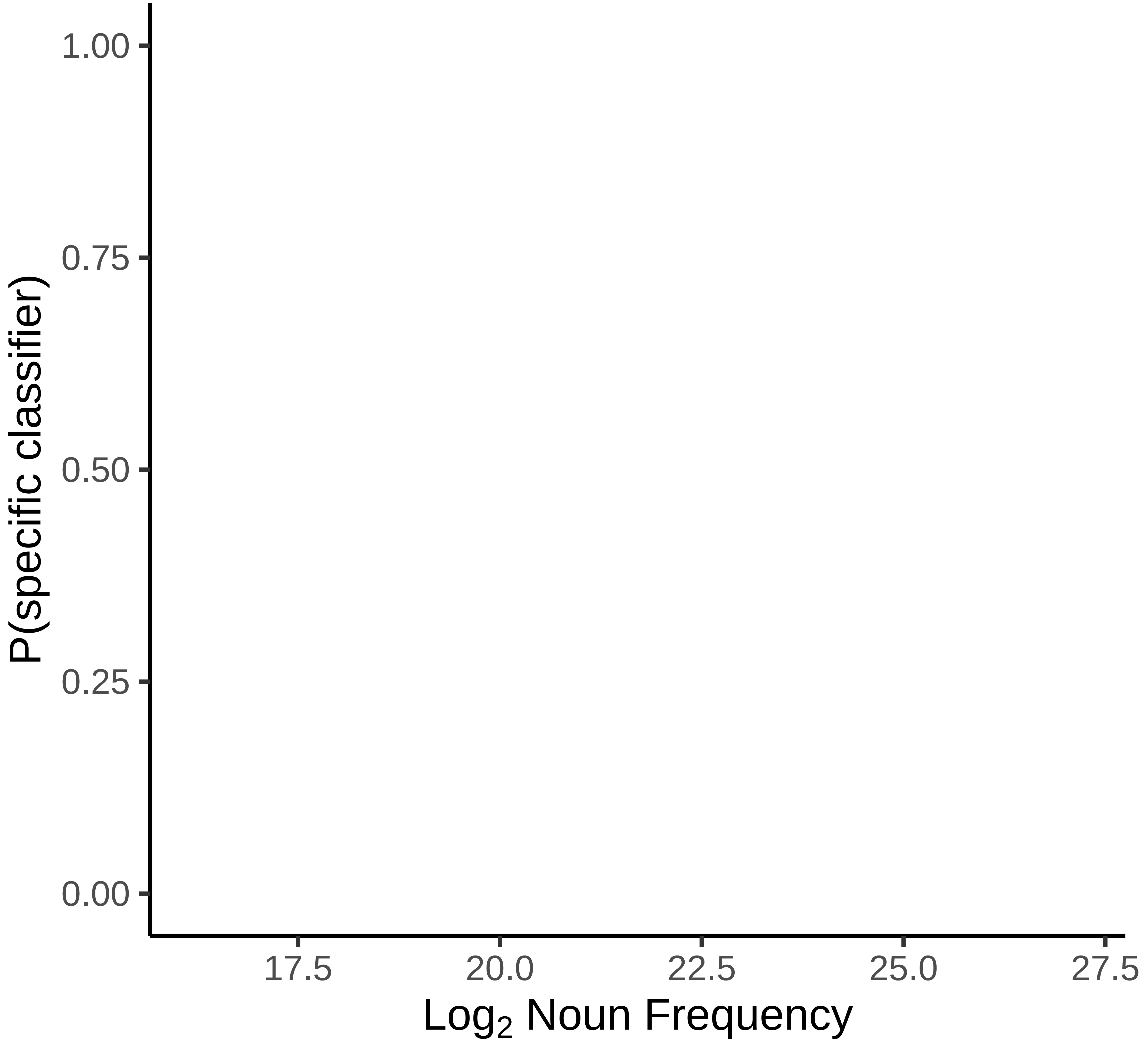
一 个 电脑
one GENERIC computer
'one computer'

一 只 猫
one ANIMAL cat
'one cat'

一 个 猫
one GENERIC cat
'one cat'

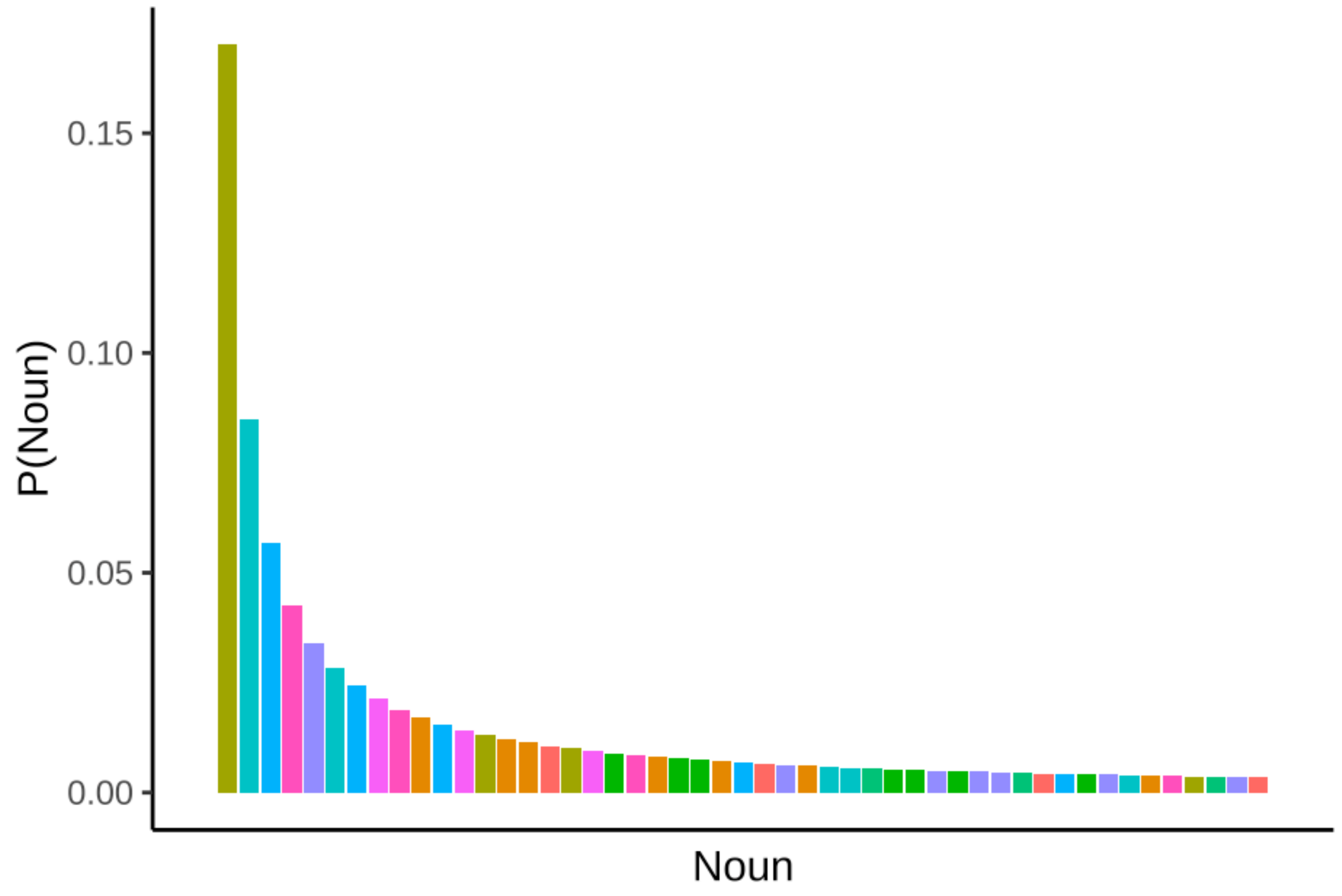
An Accessibility Effect in Mandarin Classifiers

A. Zhan & Levy (2019) Experiment



Mandarin Classifier Simulation

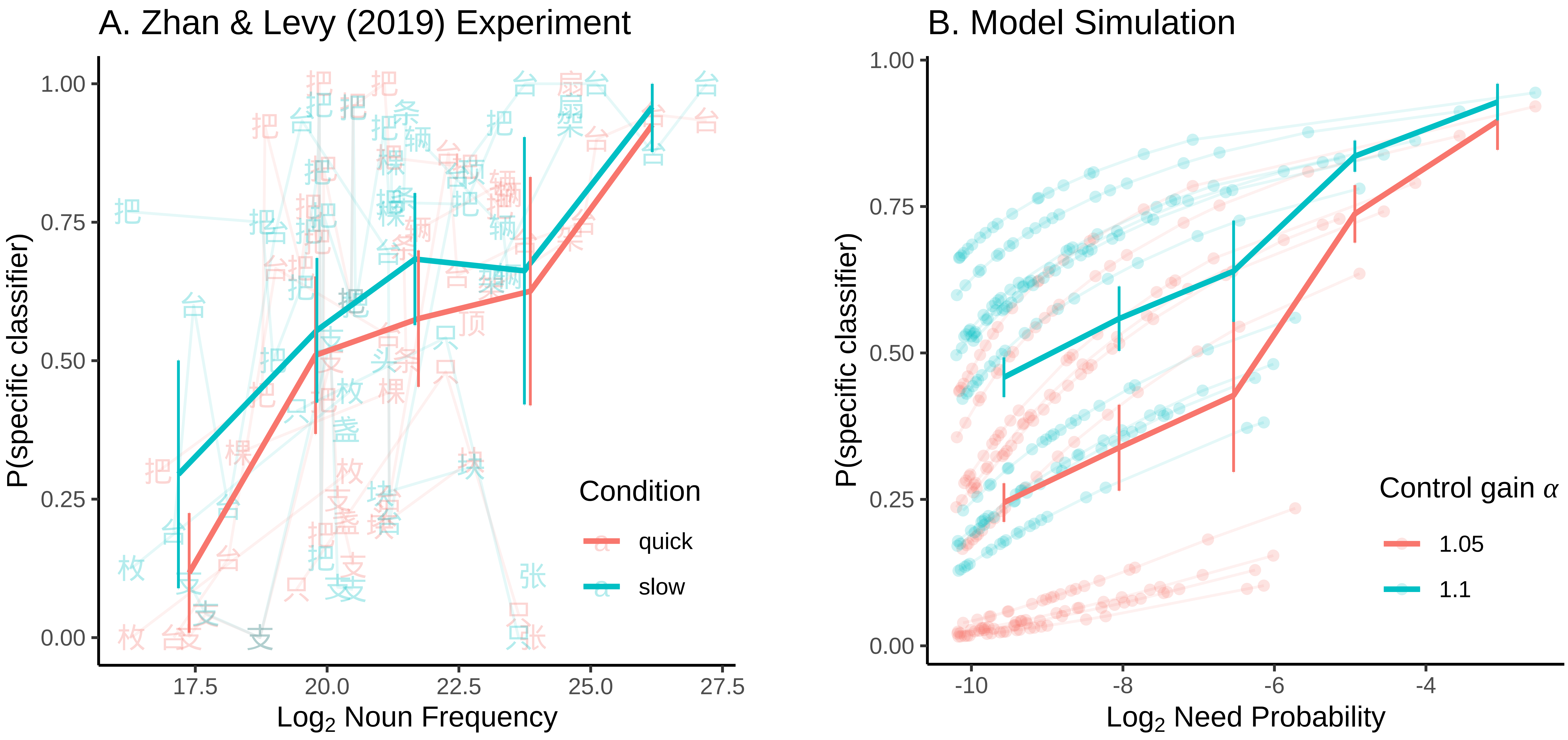
- Set up a toy language where every utterance consists of CLASSIFIER + NOUN, where CLASSIFIER can be generic or specific.
- *N*=200 different nouns, each assigned to one of 10 different specific classifiers.
- Probability distribution on nouns is Zipfian.
- Derive the constrained optimal policy.



$$P(\text{word} \mid \text{goal}, \text{state}) \propto \exp\left\{\log P(\text{word} \mid \text{state}) + \alpha u(\text{word} \mid \text{goal}, \text{state})\right\}$$

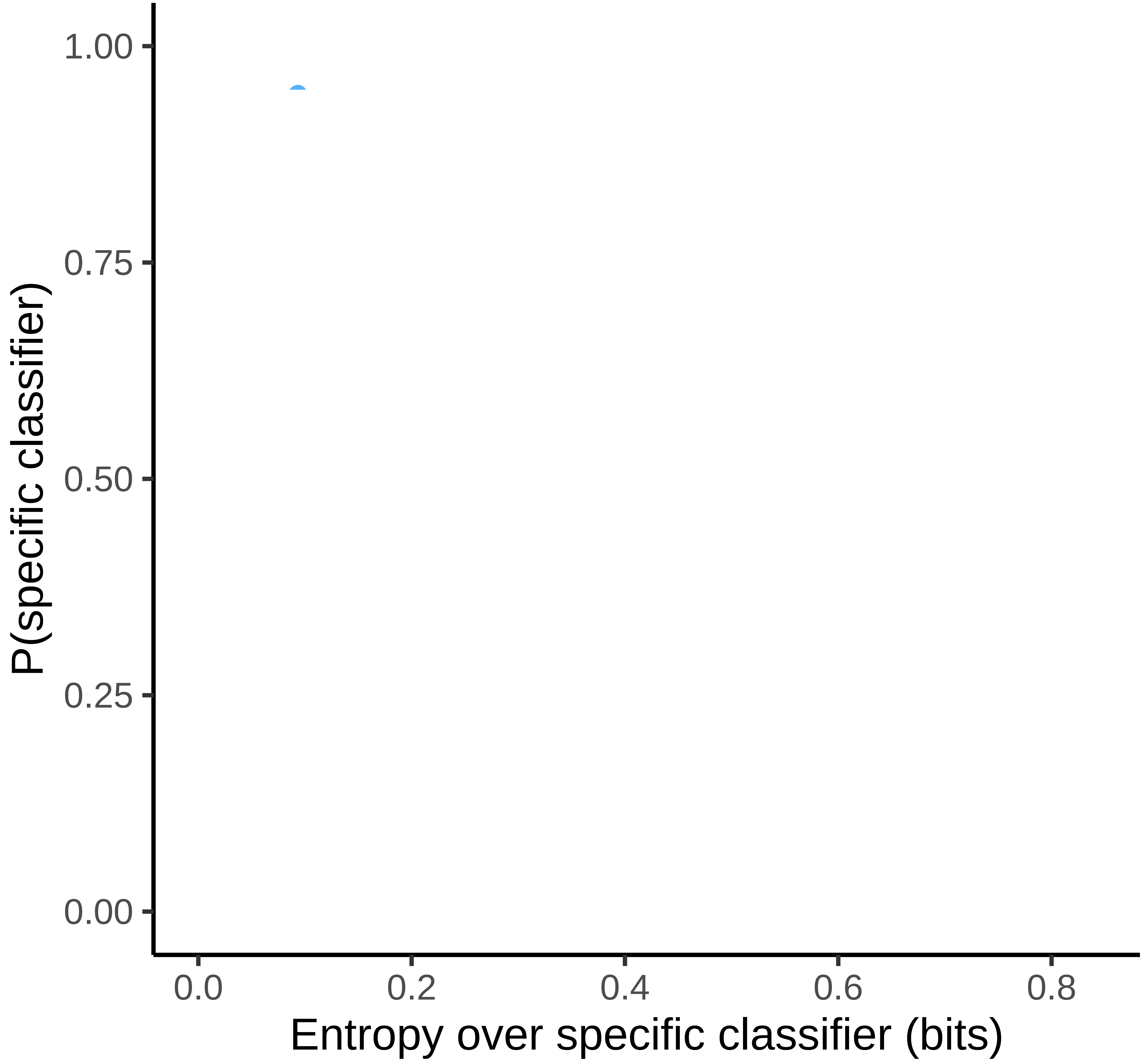
Favors generic classifier **Favors specific classifier**

Mandarin Classifier Result

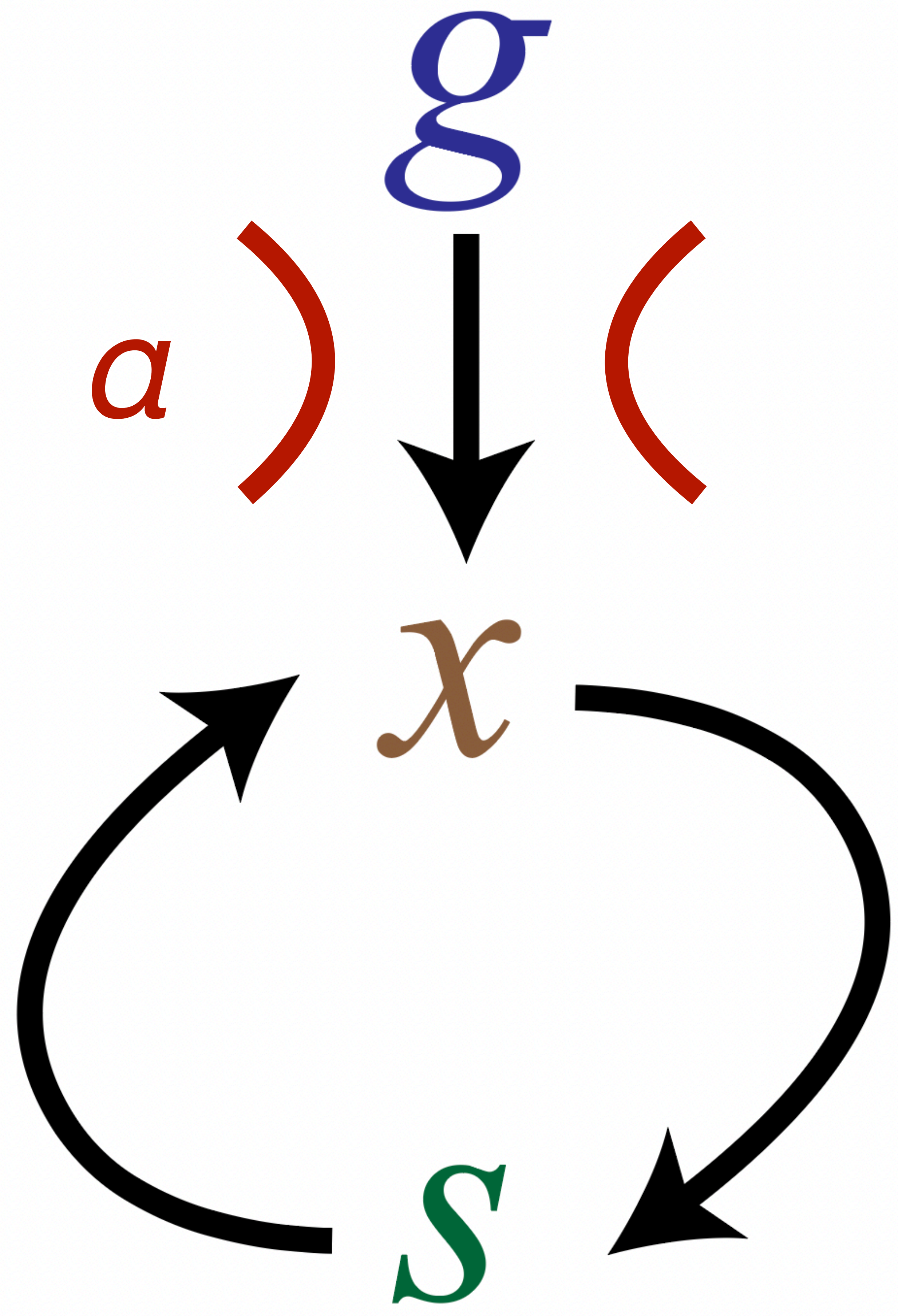


Mandarin Classifier Result

- Production of specific classifier is rare when the model has uncertainty about which specific classifier it should use.
- Matches the intuitive idea of “accessibility.”



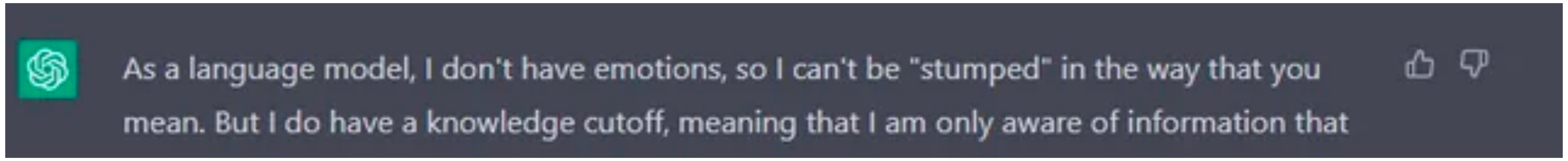
Control Bottleneck in Language Production



- An information-theoretic model captures **accessibility-based production** effects.
- A constrained optimal production policy ends up including a **language model** as a component…

$$P(\textbf{word} \mid \textbf{goal}, \textbf{state}) \propto \exp\left\{\log P(\textbf{word} \mid \textbf{state}) + \alpha u(\textbf{word} \mid \textbf{goal}, \textbf{state})\right\}$$

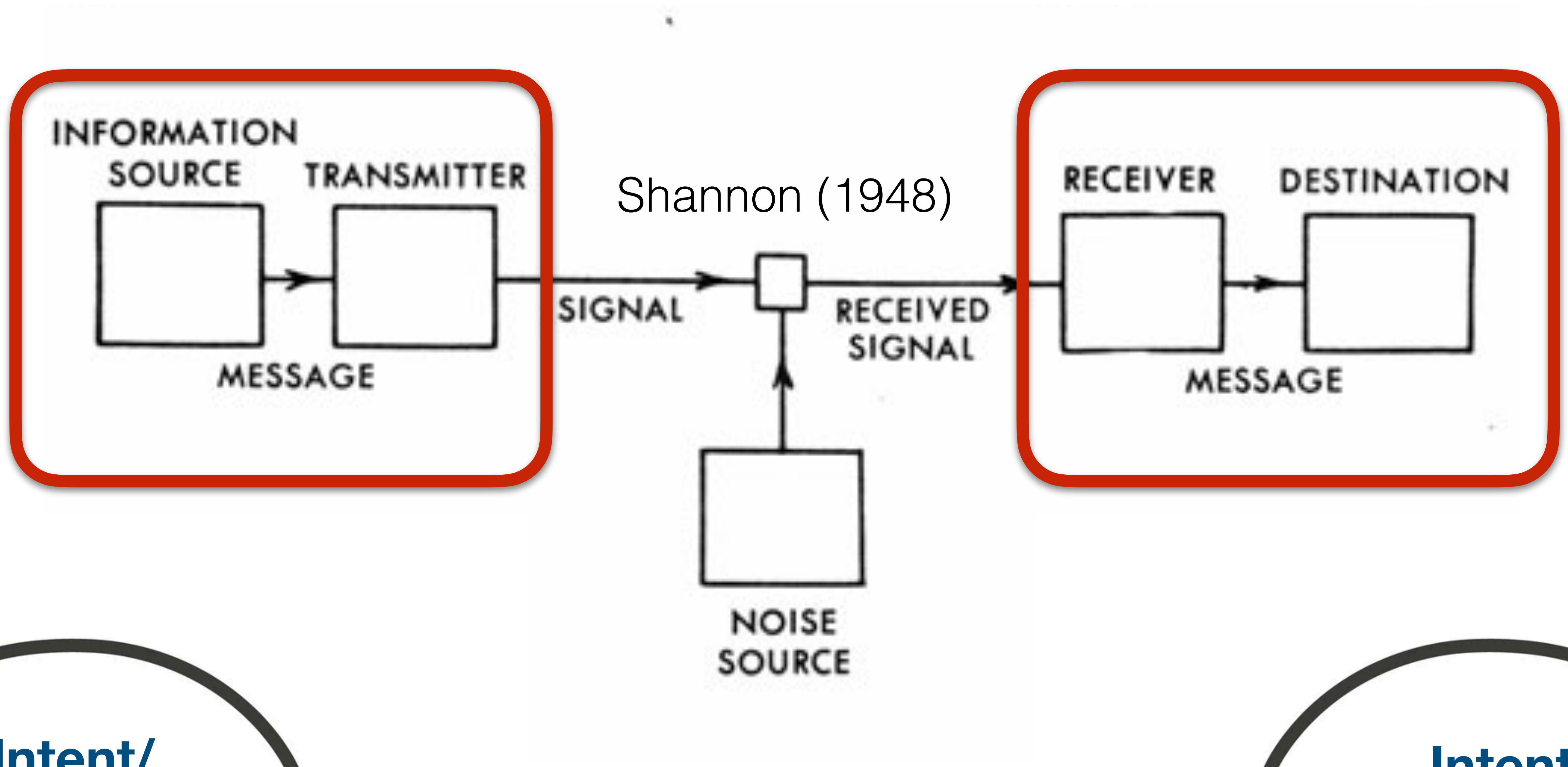
- Really, it's a language model plus a reward model:
  - As in Reinforcement Learning from Human Feedback (RLHF)

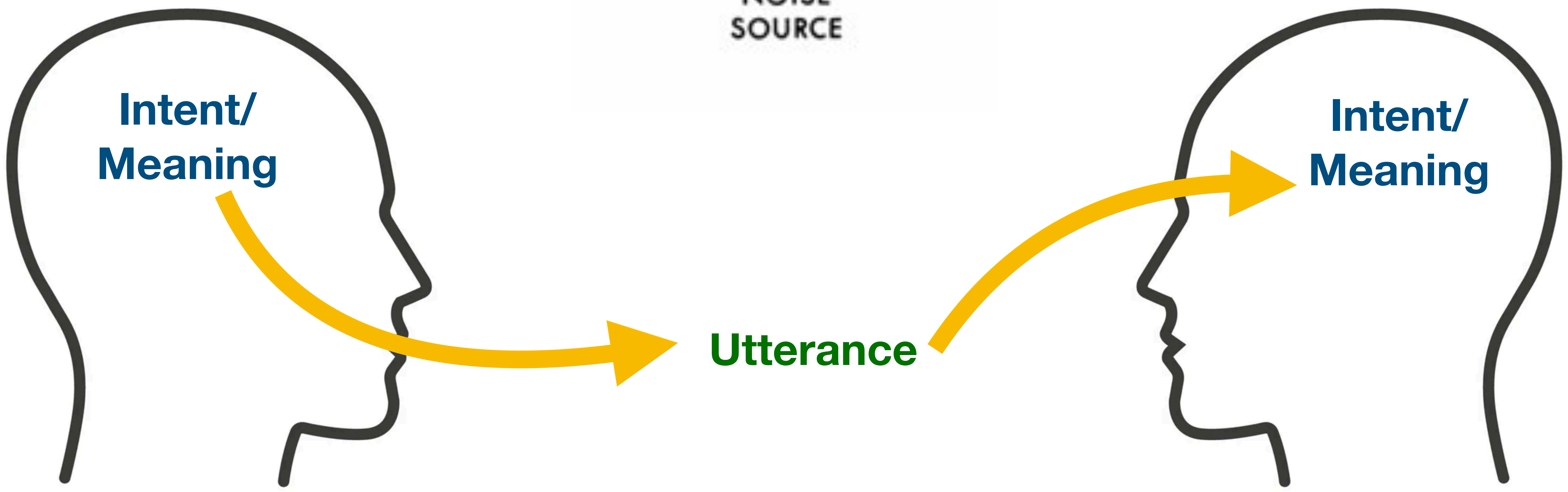


Outline

- Introduction
- Basics of Information-Theoretic Psycholinguistics
- Memory Bottleneck in Language Comprehension
- Control Bottleneck in Language Production
- Conclusion

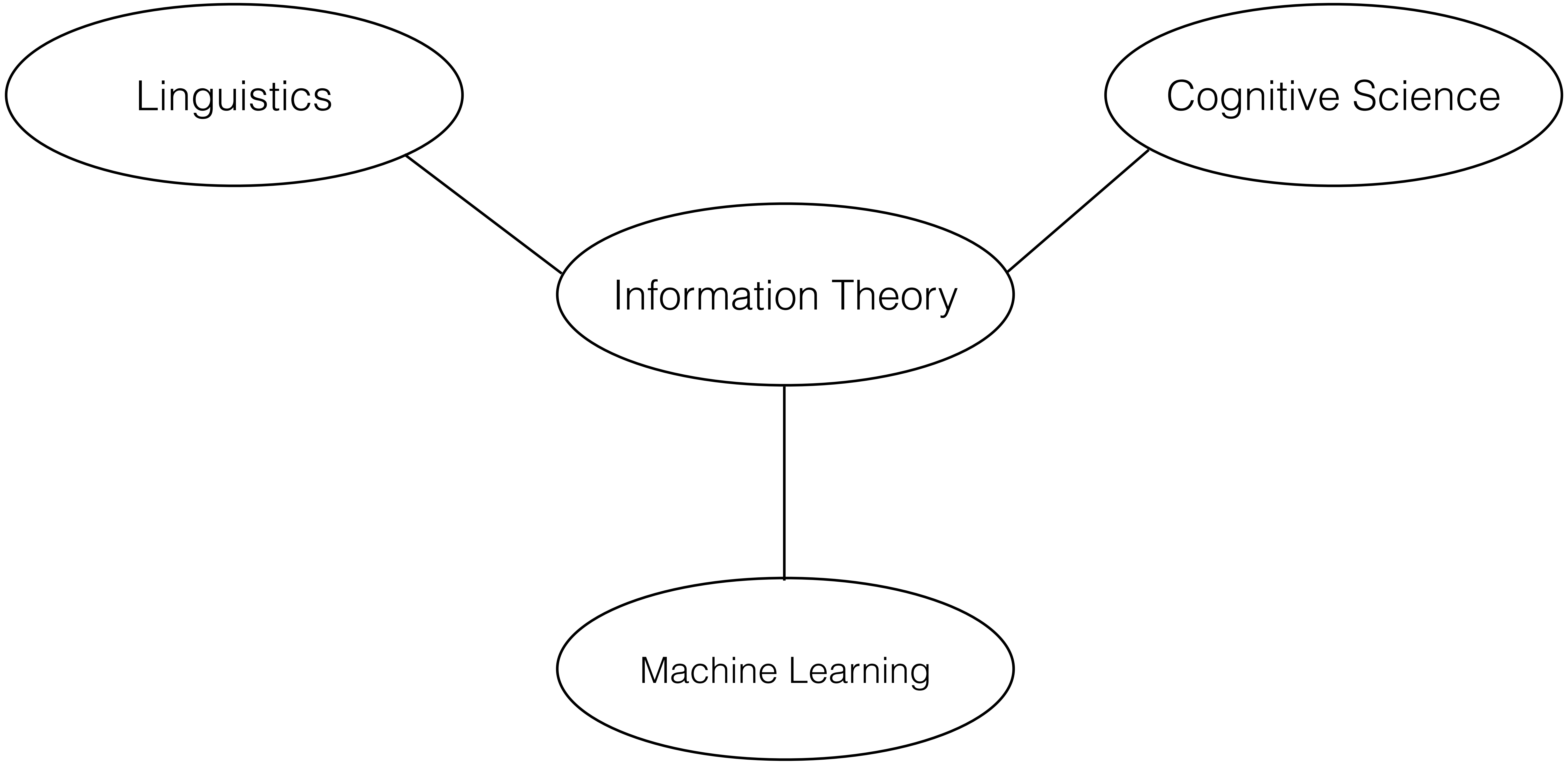
Natural Language as a Code



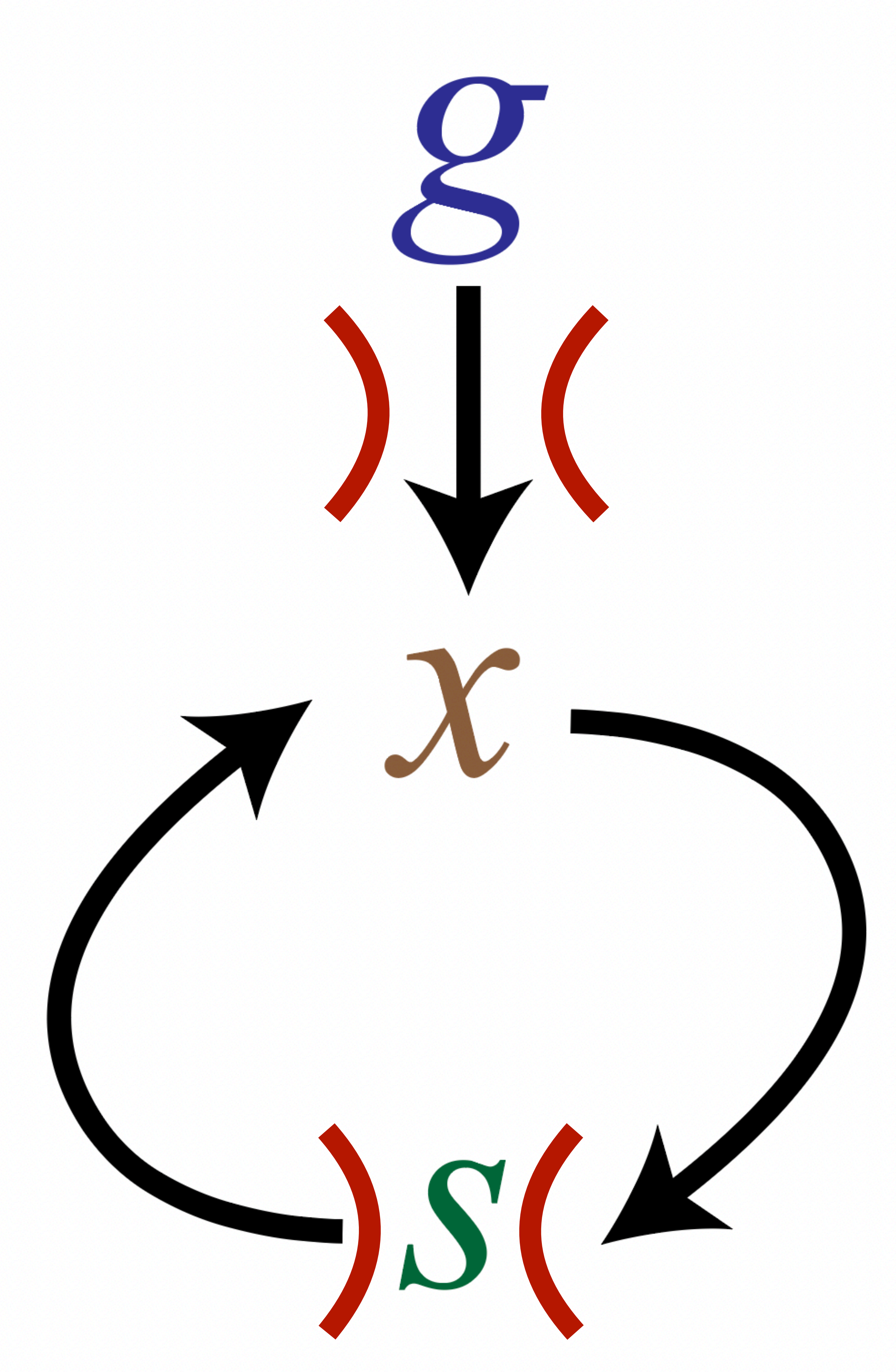


Production **Comprehension**

A Nexus Between Fields



Conclusion



- We can model language processing as optimal *subject to constraints*…
  - On incremental memory.
  - On control.
- **Language models** *P*(*word* | *context*) emerge as a key part of both comprehension and production.
  - Comprehension: They define the **information content** of each word to be processed.
  - Production: They emerge under a **constraint on cognitive control**.
- Information-theoretic psycholinguistics is an open field!

Acknowledgments

- Thanks to Michael Hahn, Kyle Mahowald, Steve Piantadosi, Ted Gibson, Roger Levy, Tal Linzen, Greg Scontras, Xin Xie, Noga Zaslavsky, and the UCI QuantLang Lab, Language Processing Group, and Center for Theoretical Behavioral Sciences for helpful discussion on the production model.
- Thanks to audiences at SyntaxFest 2019, the McQLL lab at McGill University, SigTyp, CUNY 2018 and 2019, the Complexity Sciences Center at UC Davis, the CILVR Lab at NYU, and the Institute for Mathematical Behavioral Science (IMBS) at UCI for previous discussion of the comprehension model.
- Thanks for your attention!

To find out more…

- On **lossy-context surprisal** as a model of human processing difficulty:
  - Richard Futrell, Edward Gibson, and Roger Levy. 2020. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. *Cognitive Science* 44.
  - Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. 2022. A resource-rational model of recursive sentence processing. *PNAS*.
  - Richard Futrell (2019). Information-theoretic locality properties of natural language. In *QuaSy*, pp. 2-15.
  - Richard Futrell, William Dyer, and Gregory Scontras (2020). What determines the order of adjectives in English? Comparing efficiency-based theories using dependency treebanks. In *ACL*.
  - Karthik Sharma, Richard Futrell, and Samar Husain (2021). What determines the order of verbal dependents in Hindi? In *CMCL*.
  - Michael Hahn, Judith Degen, and Richard Futrell. Explaining patterns of word and morpheme order as an efficient tradeoff of memory and surprisal. *Psychological Review*.
- On **RDC production model**

  - Richard Futrell (2021). An information-theoretic account of semantic interference in word production. *Frontiers in Psychology*.

  - Shiva Upadhye & Richard Futrell (2022). An information-theoretic account of semantic substitution errors in speech. In *InfoCog*.

  - Richard Futrell (2023). An information-theoretic account of accessibility effects in incremental language production. In *CogSci*.

  - Richard Futrell (2023). Information-theoretic principles in incremental language production. *PNAS*.