Tal Linzen[1,3], Emmanuel Dupoux[1] and Yoav Goldberg[2]

[1] *LSCP & IJN, ENS Paris*
[2] *Bar Ilan University*
[3] *Johns Hopkins University*

Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies

The Unreasonable Effectiveness of Recurrent Neural Networks

May 21, 2015

There's something magical about Recurrent Neural Networks (RNNs). I still remember when I trained my first

- Machine translation (Bahdanau et al., 2015)
- Language modeling (Mikolov et al., 2010; Sundermeyer et al., 2012)
- Parsing (Vinyals et al., 2015)
- ...

What do we know about RNN LMs?

- Low perplexity: 30.6 compared to 67 for the best n-gram model (Jozefowicz et al., 2016)

What do we know about RNN LMs?

- Low perplexity: 30.6 compared to 67 for the best n-gram model (Jozefowicz et al., 2016)
- Samples:

  About 800 people gathered at Hever Castle on Long Beach from noon to 2pm , three to four times that of the funeral cortège .

What do we know about RNN LMs?

- Low perplexity: 30.6 compared to 67 for the best n-gram model (Jozefowicz et al., 2016)
- Samples:

  About 800 people gathered at Hever Castle on Long Beach from noon to 2pm , three to four times that of the funeral cortège .

What do we know about RNN LMs?

- Low perplexity: 30.6 compared to 67 for the best n-gram model (Jozefowicz et al., 2016)
- Samples:

  About 800 people gathered at Hever Castle on Long Beach from noon to 2pm , three to four times that of the funeral cortège .

- What did the model learn?

What do we know about RNN LMs?

- Low perplexity: 30.6 compared to 67 for the best n-gram model (Jozefowicz et al., 2016)
- Samples:

  About 800 people gathered at Hever Castle on Long Beach from noon to 2pm , three to four times that of the funeral cortège .

- What did the model learn?
- Linguistics and psycholinguistics can help characterize the model's strengths and weaknesses

Subject-verb agreement

Subject-verb agreement

1. The station has a single platform.
2. *The station have a single platform.

Subject-verb agreement

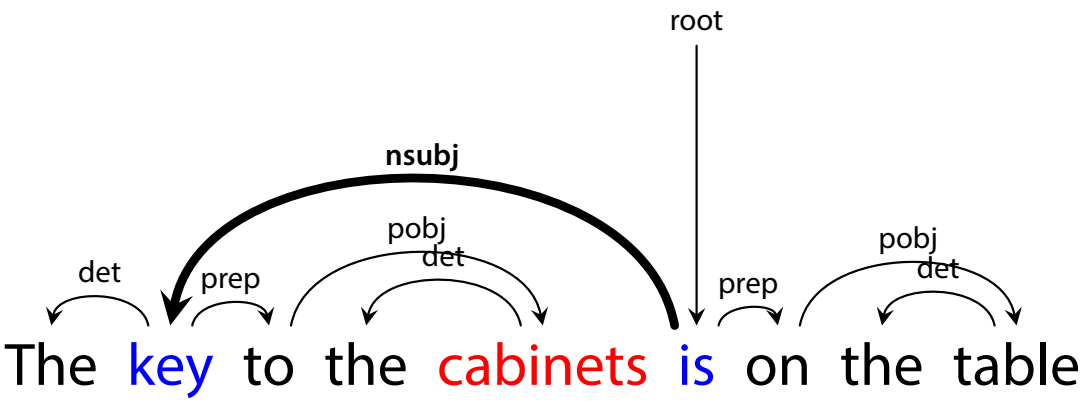
1. The station has a single platform.
2. *The station have a single platform.
3. The ratio of men who survive to the women and children who survive is not clear in this story.

Subject-verb agreement

1. The station has a single platform.
2. *The station have a single platform.
3. The ratio of men who survive to the women and children who survive is not clear in this story.
4. The professor said that the soils carried in the floodwaters add nutrients.

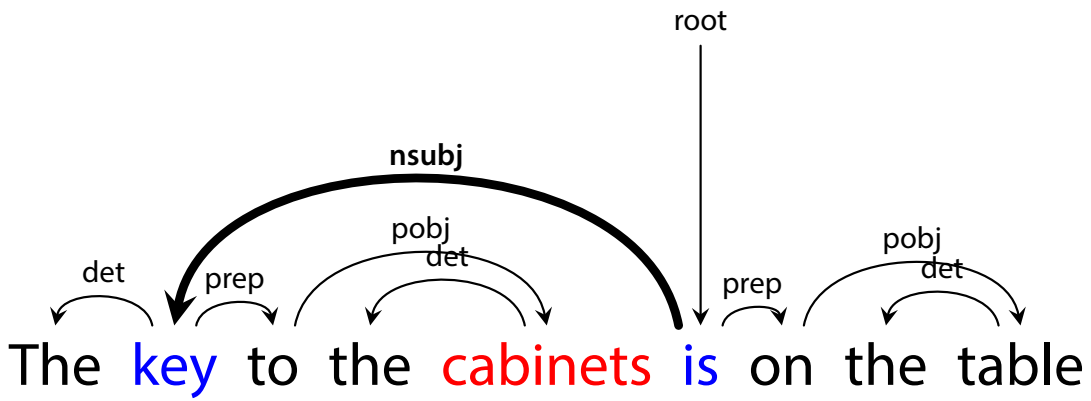
Why is agreement interesting?

- Straightforward with a structured representation of the sentence:



Why is agreement interesting?

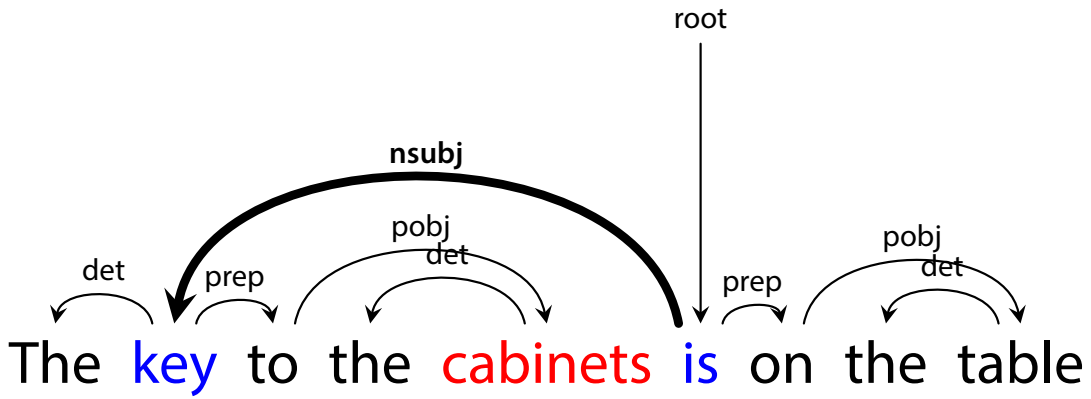
- Straightforward with a structured representation of the sentence:



- RNNs are sequence models

Why is agreement interesting?

- Straightforward with a structured representation of the sentence:



- RNNs are sequence models
- **Syntactic knowledge without explicit structure?** (Elman, 1991)

Outline

- The number prediction task

Outline

- The number prediction task
- An RNN trained specifically to perform the task

Outline

- The number prediction task
- An RNN trained specifically to perform the task
- **A focus on structurally interesting sentences**

Outline

- The number prediction task
- An RNN trained specifically to perform the task
- **A focus on structurally interesting sentences**
- RNN trained on a language modeling objective

Outline

- The number prediction task
- An RNN trained specifically to perform the task
- **A focus on structurally interesting sentences**
- RNN trained on a language modeling objective
- Preview: LSTMs are pretty good, but only with specific supervision, and should be improved even in that setup

Outline

- The number prediction task
- An RNN trained specifically to perform the task
- A focus on structurally interesting sentences
- RNN trained on a language modeling objective

Number prediction

The length of the forewings...

Number prediction

The length of the forewings... **SINGULAR**

Number prediction

The length of the forewings... **SINGULAR**

The keys to the cabinets...

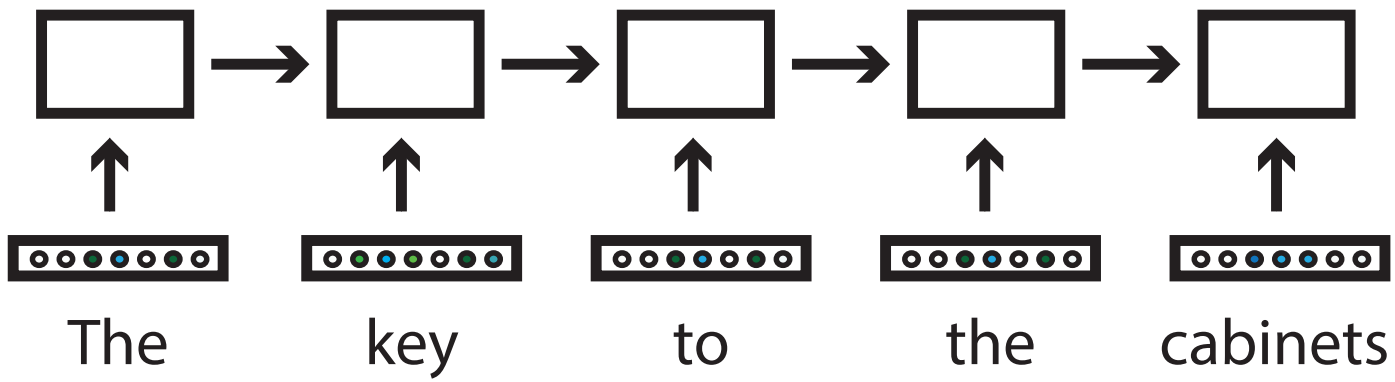
Number prediction

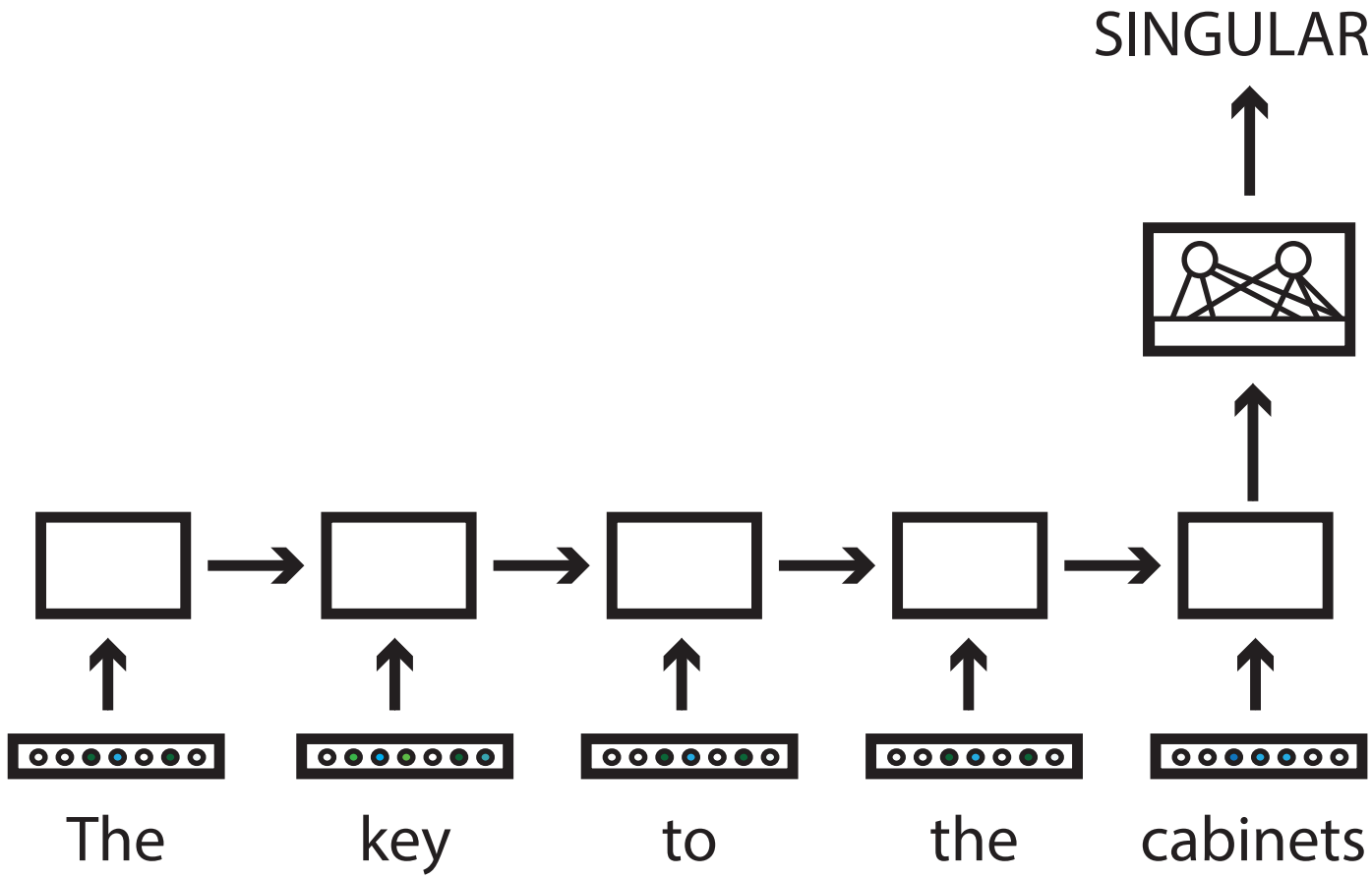
The length of the forewings... **SINGULAR**

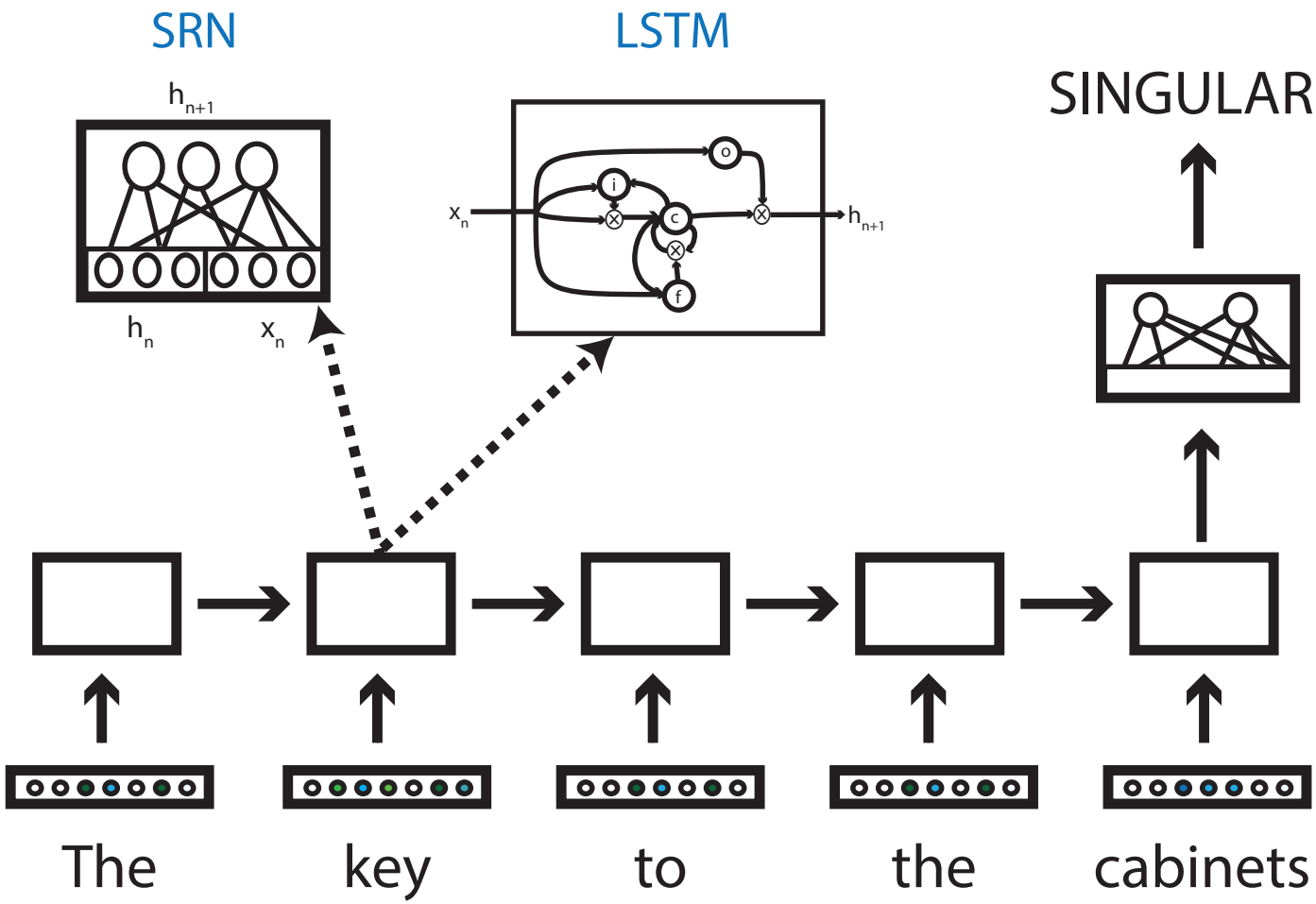
The keys to the cabinets... **PLURAL**

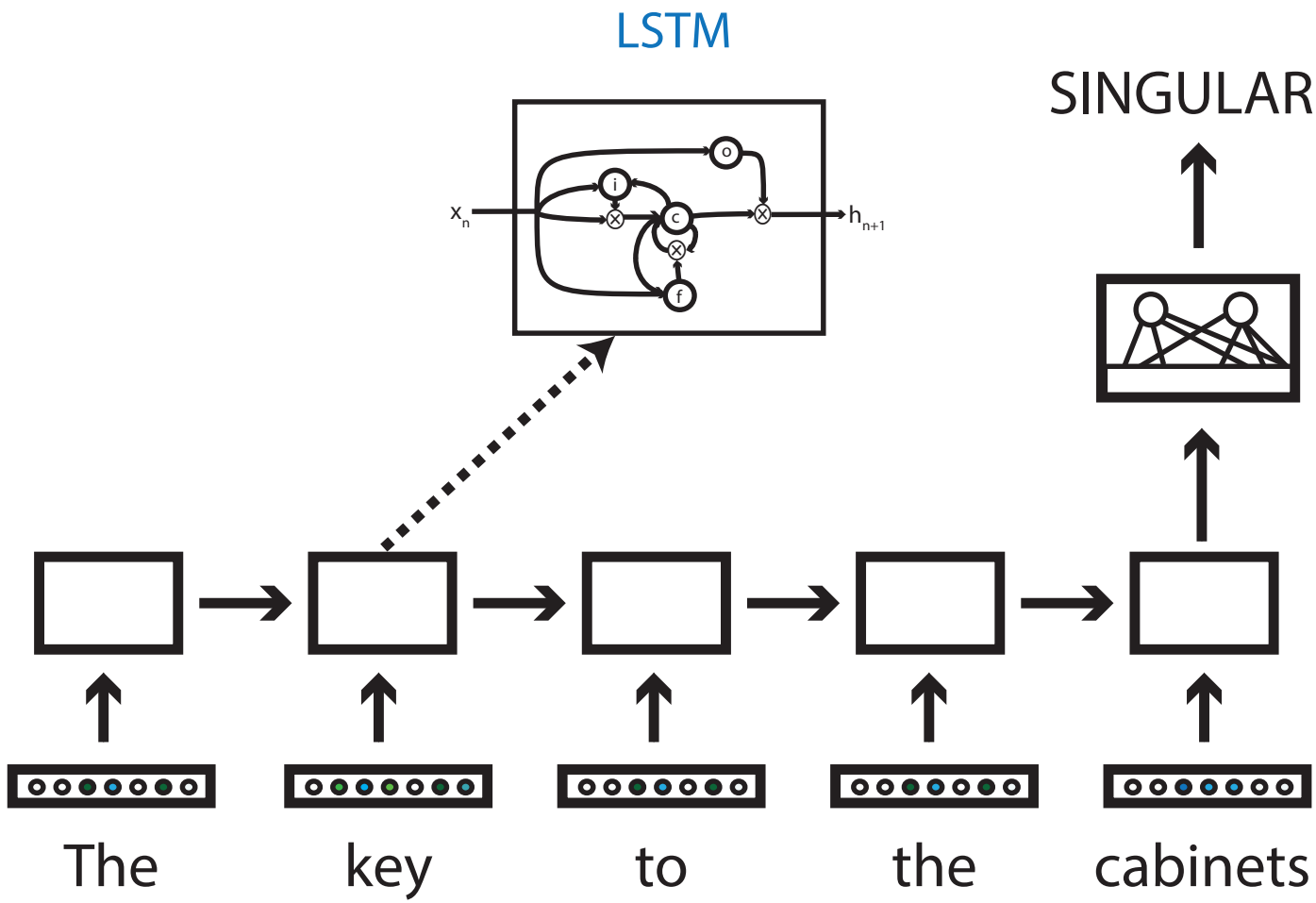
Outline

- The number prediction task
- An RNN trained specifically to perform the task
- A focus on structurally interesting sentences
- RNN trained on a language modeling objective


The
key
to
the
cabinets


SINGULAR
The
key
to
the
cabinets


SRN
LSTM
SINGULAR
h n+1
h n
x n
x n
i
o
c
f
h n+1
The
key
to
the
cabinets


LSTM
SINGULAR
o
i
c
f
x_n
h_{n+1}
The
key
to
the
cabinets

Experimental setup

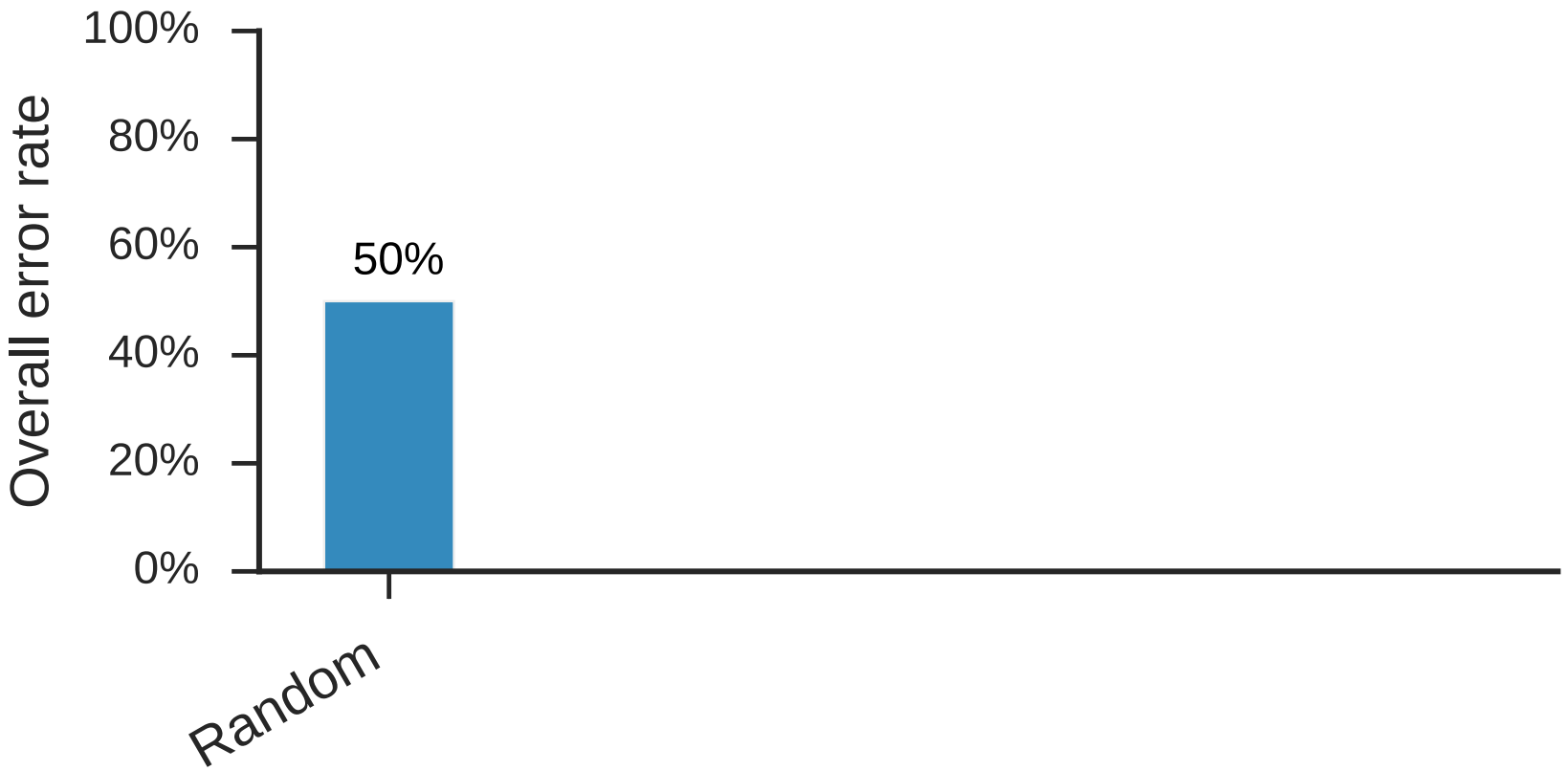
- Sentences from the English Wikipedia that have a present-tense verb
- 121K in training set, 1.21M in test set
- 50 hidden units
- 50-dimensional word embeddings
- Results averaged over 20 runs

Experimental setup

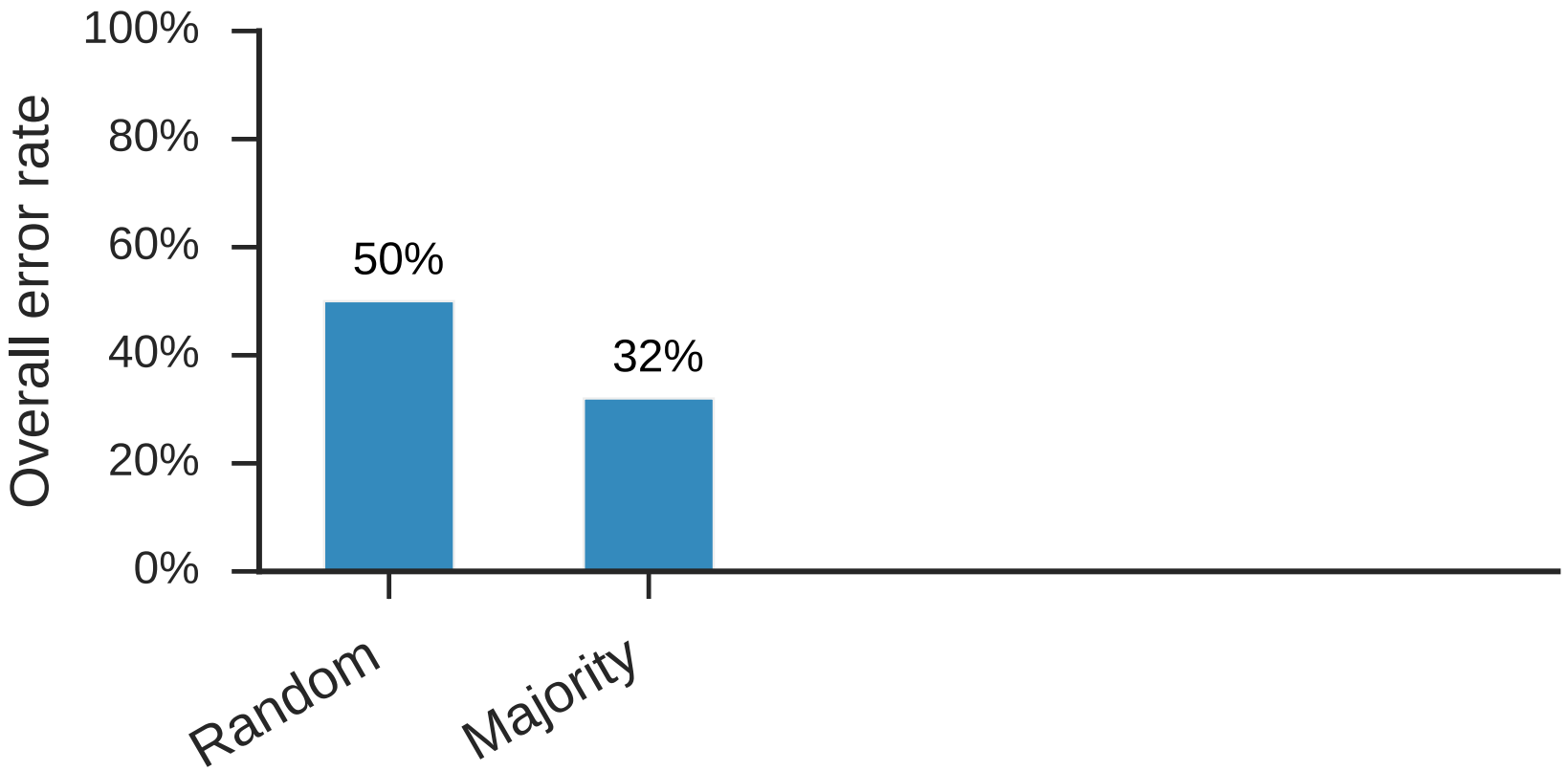
- Sentences from the English Wikipedia that have a present-tense verb
- 121K in training set, 1.21M in test set
- 50 hidden units
- 50-dimensional word embeddings
- Results averaged over 20 runs
- **(Only) prerequisite: automatic identification of verb number**

Overall results

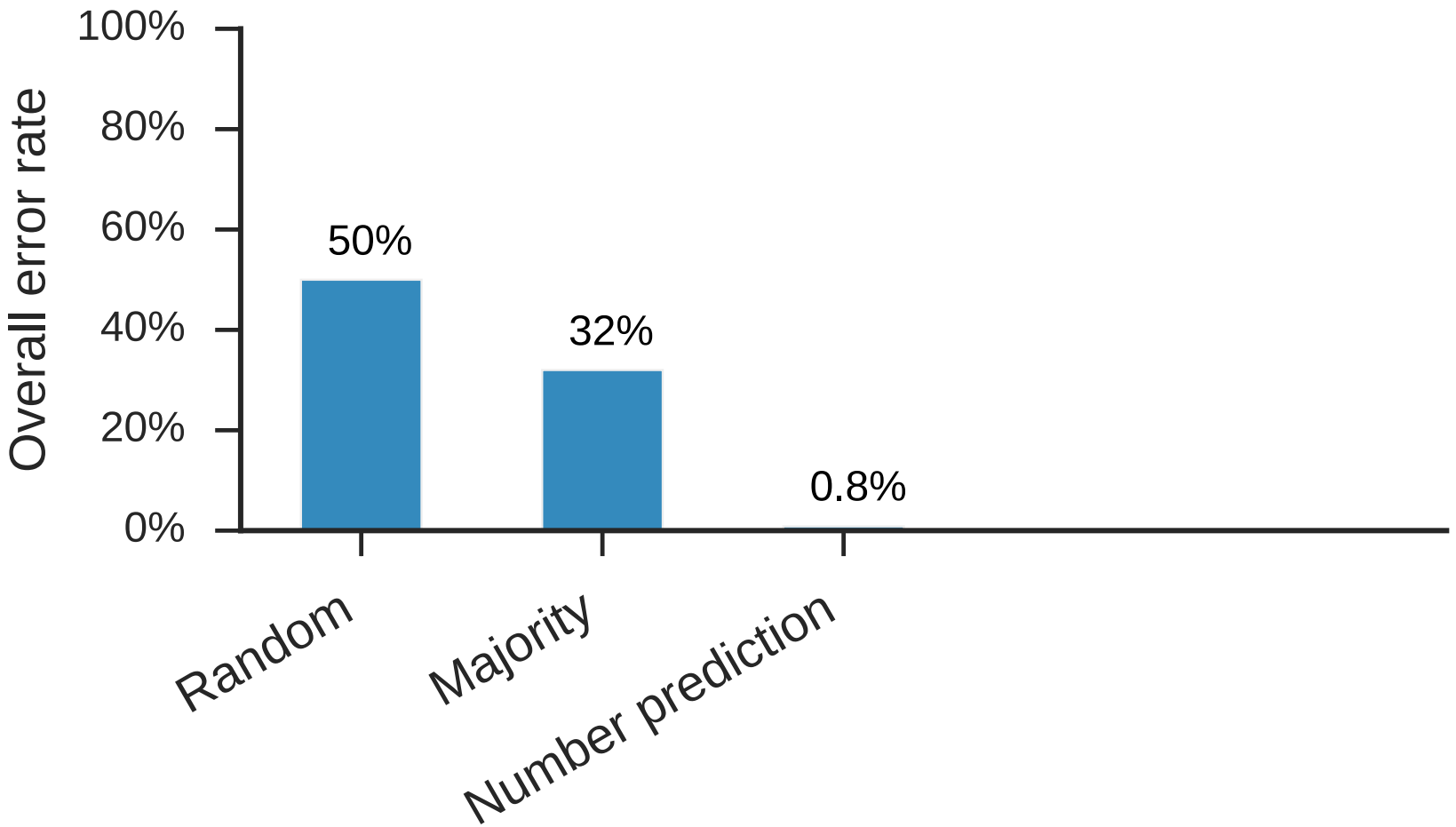
Overall results



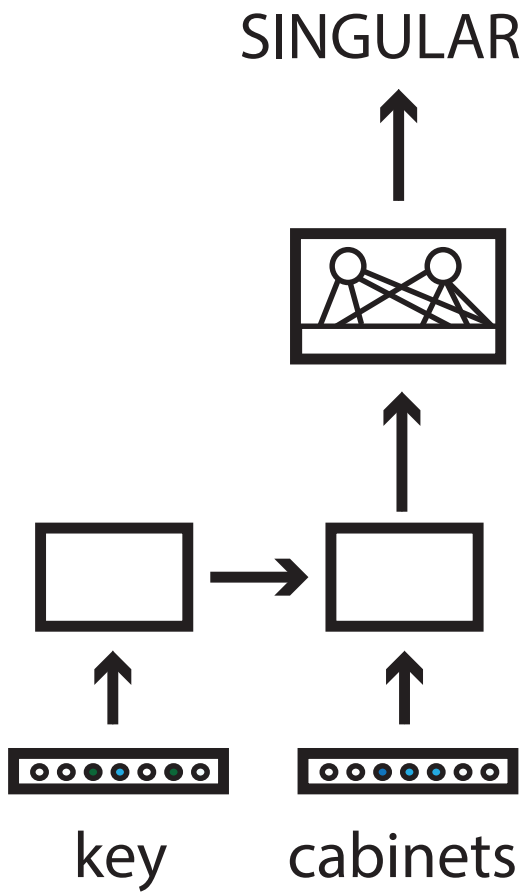
Overall results



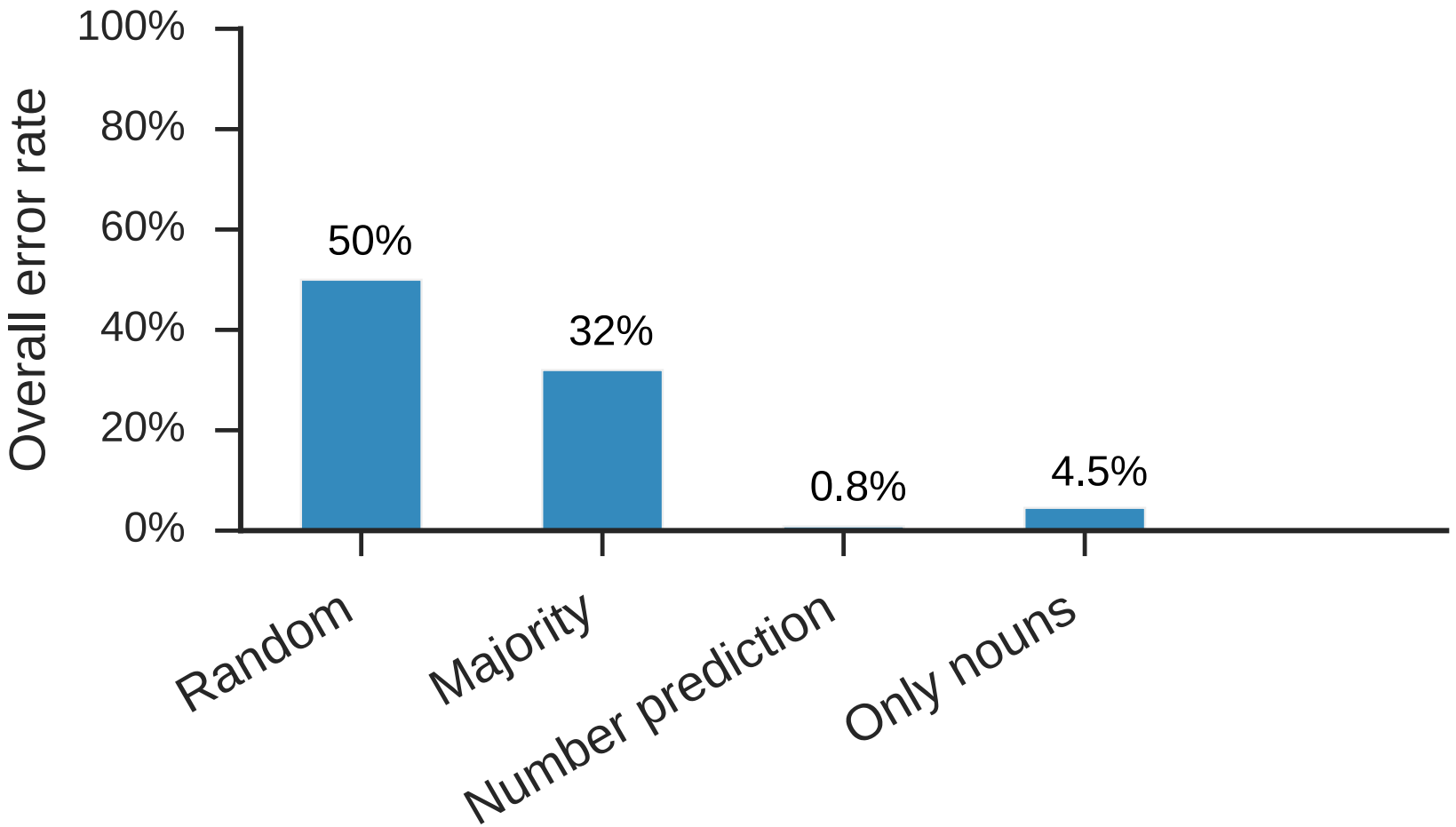
Overall results



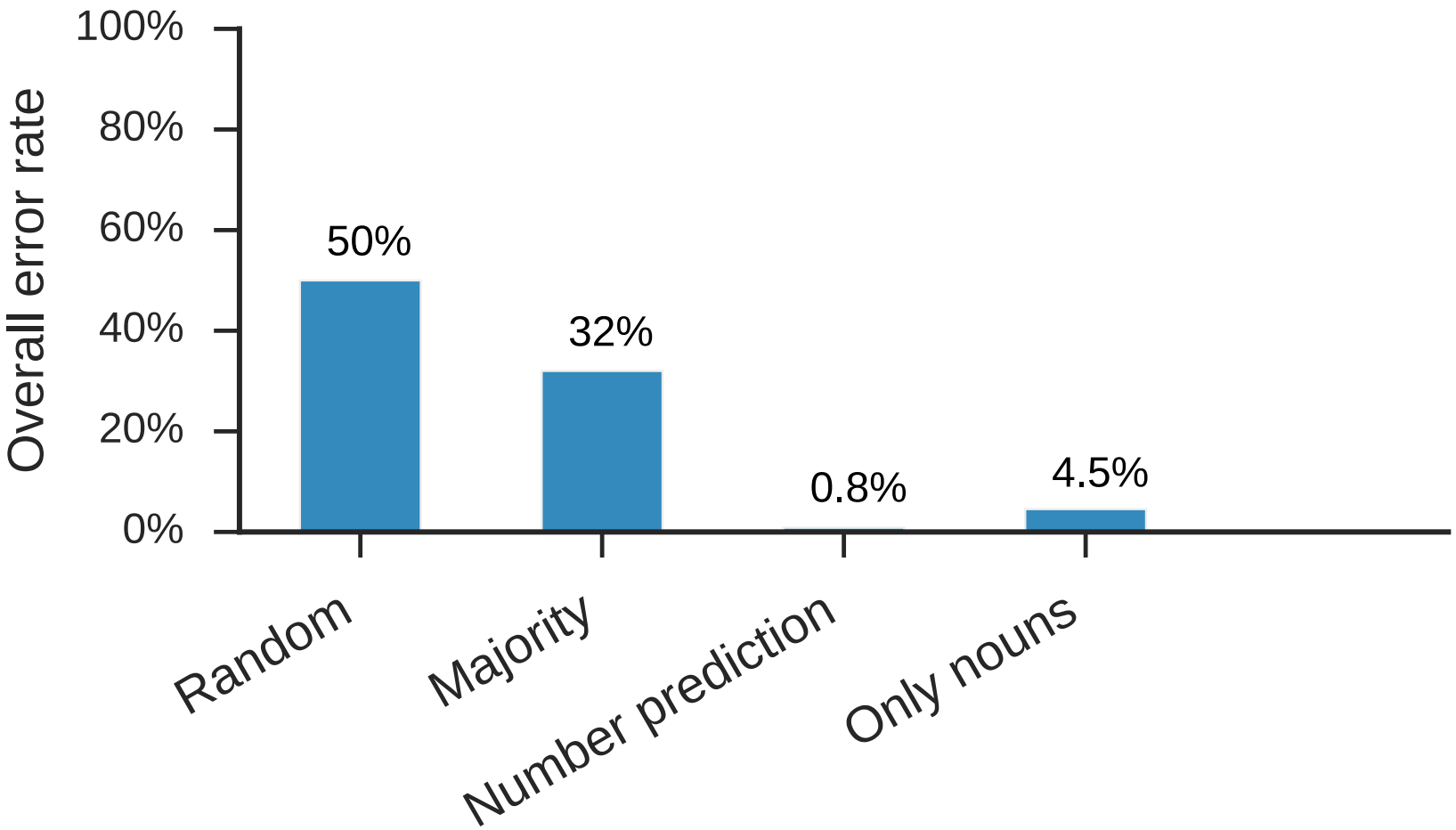
Is it just about the nouns?



Is it just about the nouns?

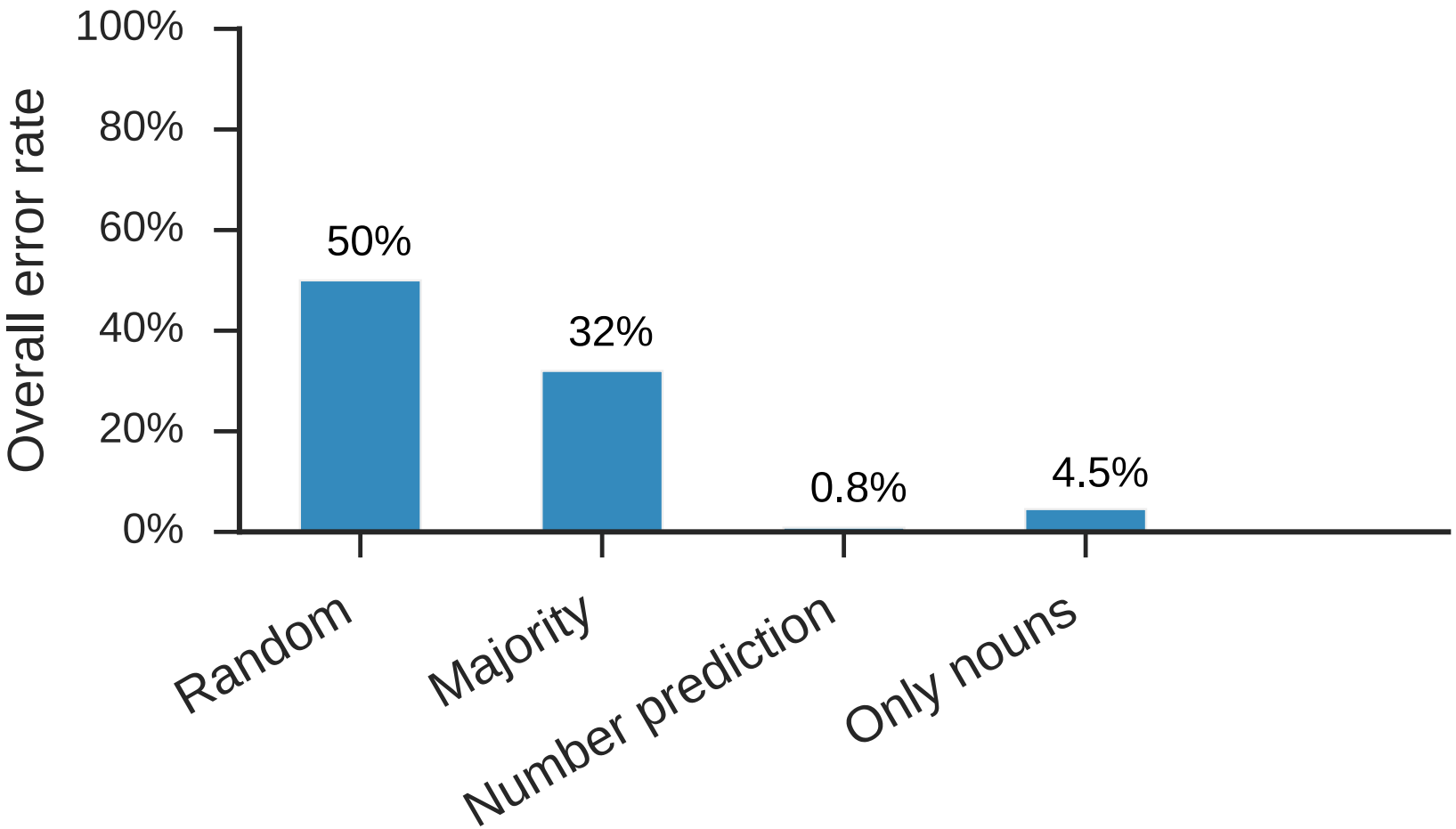


Is it just about the nouns?



- Other parts of speech are essential: more than five times as many errors without them

Is it just about the nouns?



- Other parts of speech are essential: more than five times as many errors without them
- On average the task isn't very hard...

Outline

- The number prediction task
- An RNN trained specifically to perform the task
- A focus on structurally interesting sentences
- RNN trained on a language modeling objective

Attractors
(Bock & Miller, 1991)

1. The keys are on the table.

Attractors

(Bock & Miller, 1991)

1. The keys are on the table.
2. The keys to the cabinet are on the table.

Attractors

(Bock & Miller, 1991)

1. The keys are on the table.
2. The keys to the cabinet are on the table.
3. The keys to the cabinets are on the table.

Attractors

(Bock & Miller, 1991)

1. The keys are on the table.
2. The keys to the cabinet are on the table.
3. ~~The keys to the cabinets are on the table.~~

Attractors

(Bock & Miller, 1991)

1. The keys are on the table.
2. The keys to the cabinet are on the table.
3. ~~The keys to the cabinets are on the table.~~
4. The ratio of men to women is not clear.

Attractors

(Bock & Miller, 1991)

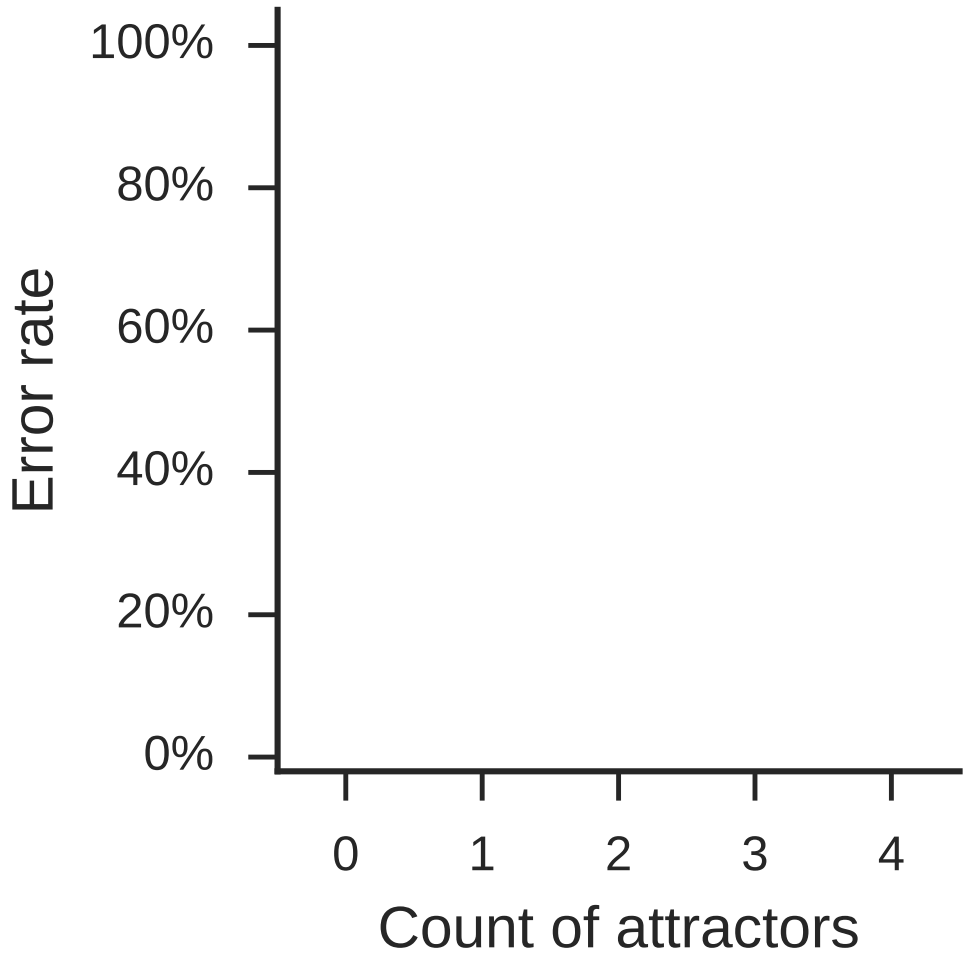
1. The keys are on the table.
2. The keys to the cabinet are on the table.
3. ~~The keys to the cabinets are on the table.~~
4. The ratio of men to women is not clear.
5. The ratio of men to women and children is not clear.

Attractors
(Bock & Miller, 1991)

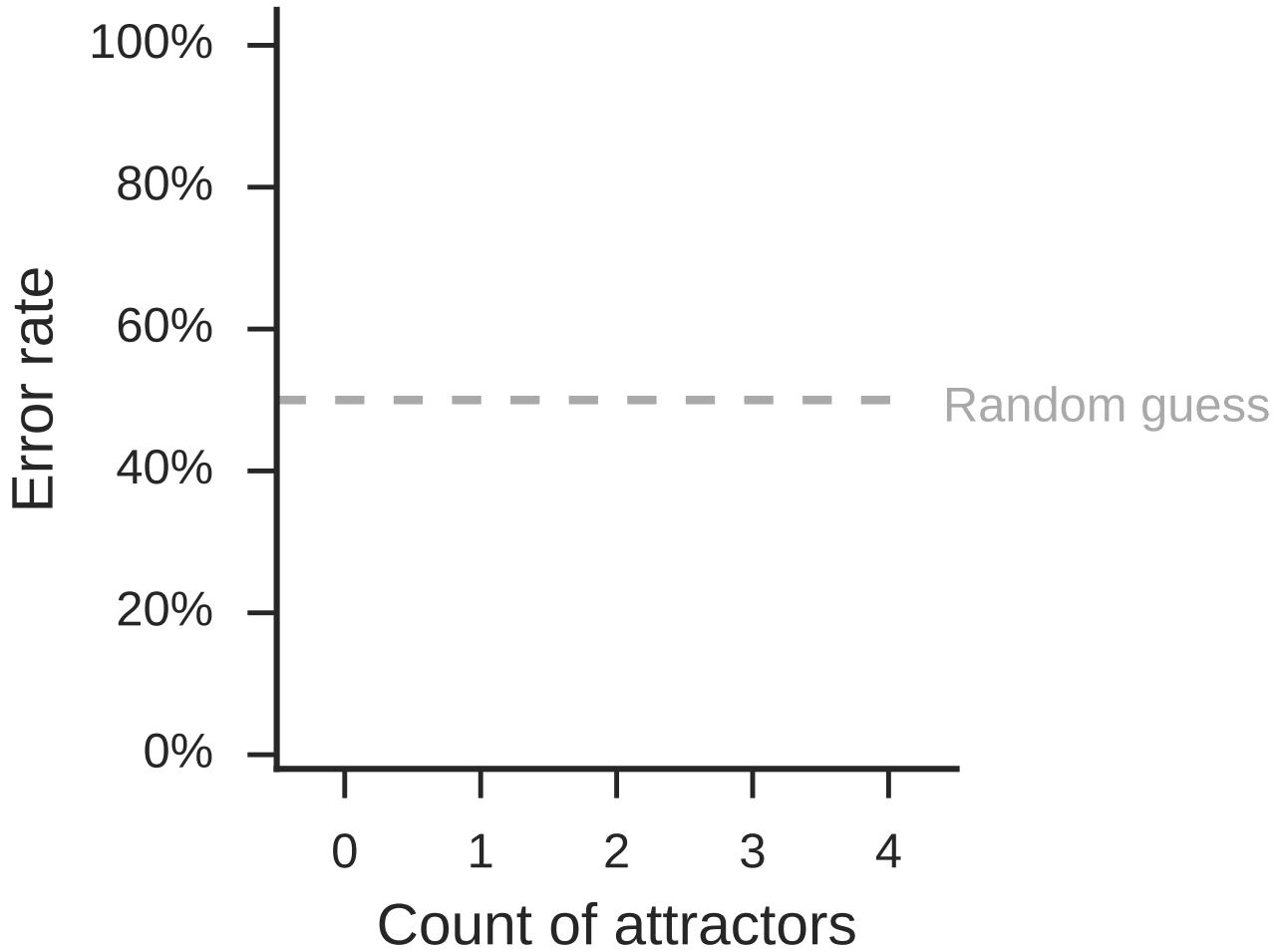
1. The keys are on the table.
2. The keys to the cabinet are on the table.
3. ~~The keys to the cabinets are on the table.~~
4. The ratio of men to women is not clear.
5. The ratio of men to women and children is not clear.

- **Prerequisite: a dependency parse of the test sentences**
  (Goldberg & Nivre, 2012)

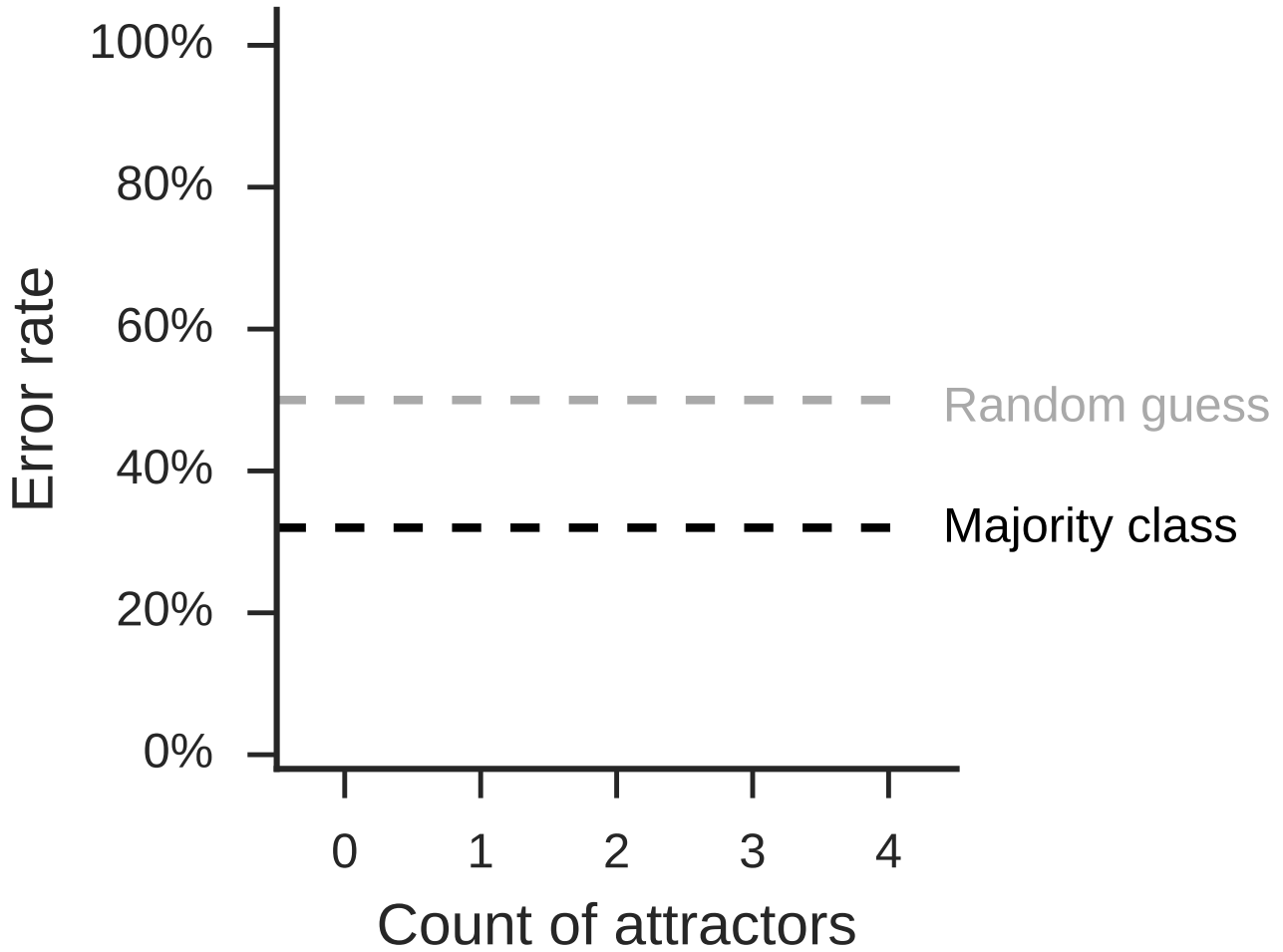
Attractors



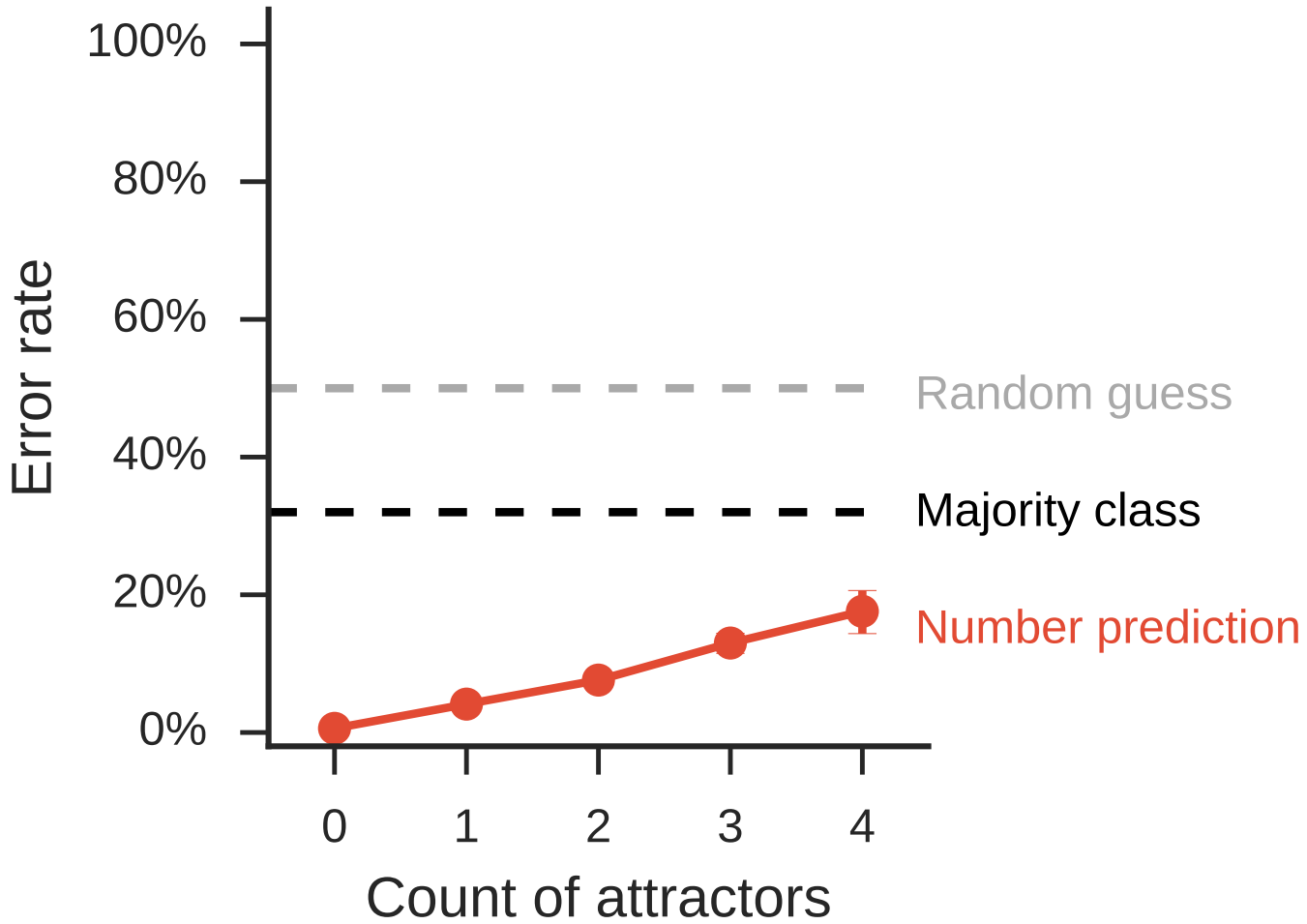
Attractors


100%
80%
60%
40%
20%
0%
Error rate
Random guess
0
1
2
3
4
Count of attractors

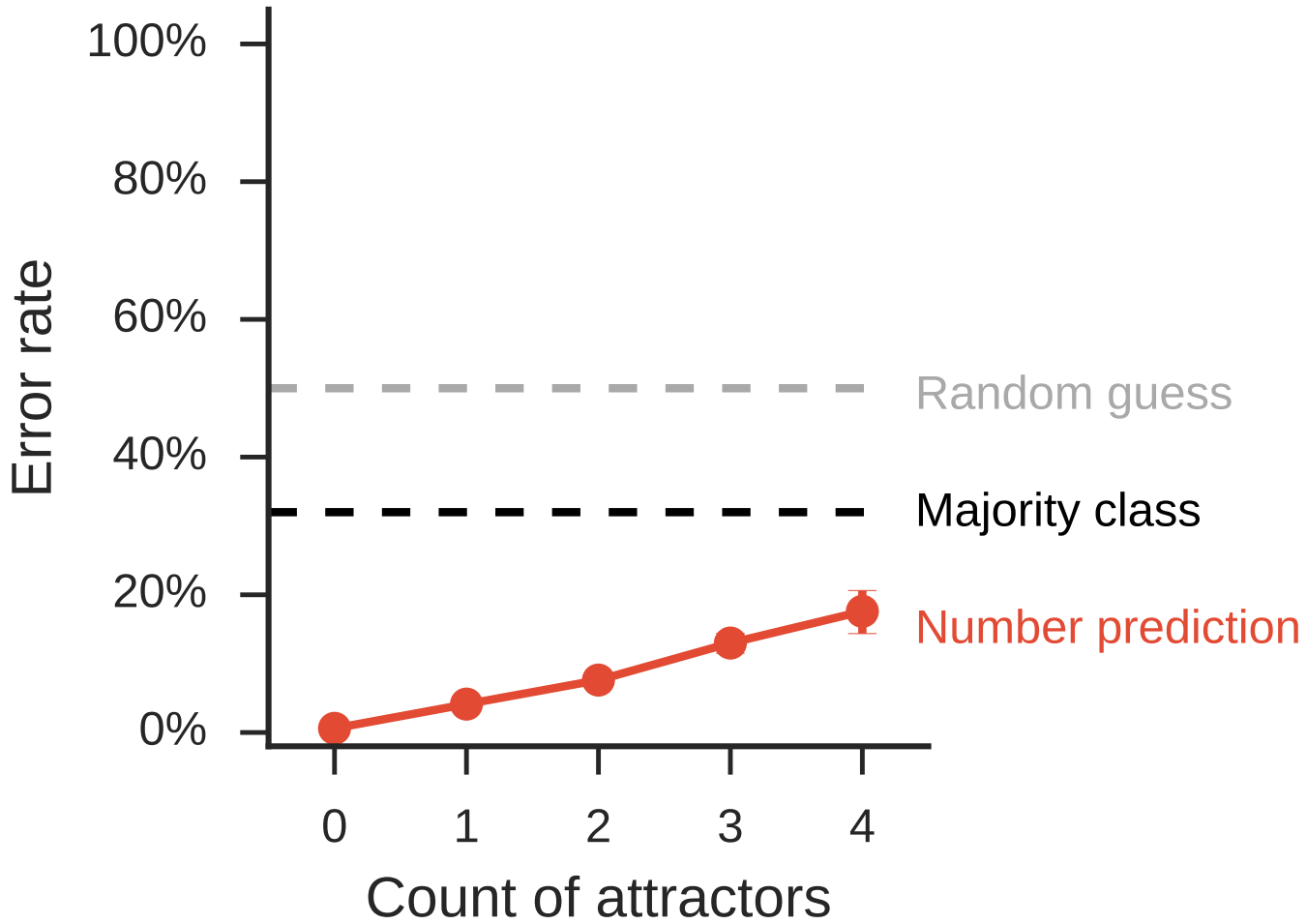
Attractors


100%
80%
60%
40%
20%
0%
Error rate
Random guess
Majority class
0
1
2
3
4
Count of attractors

Attractors

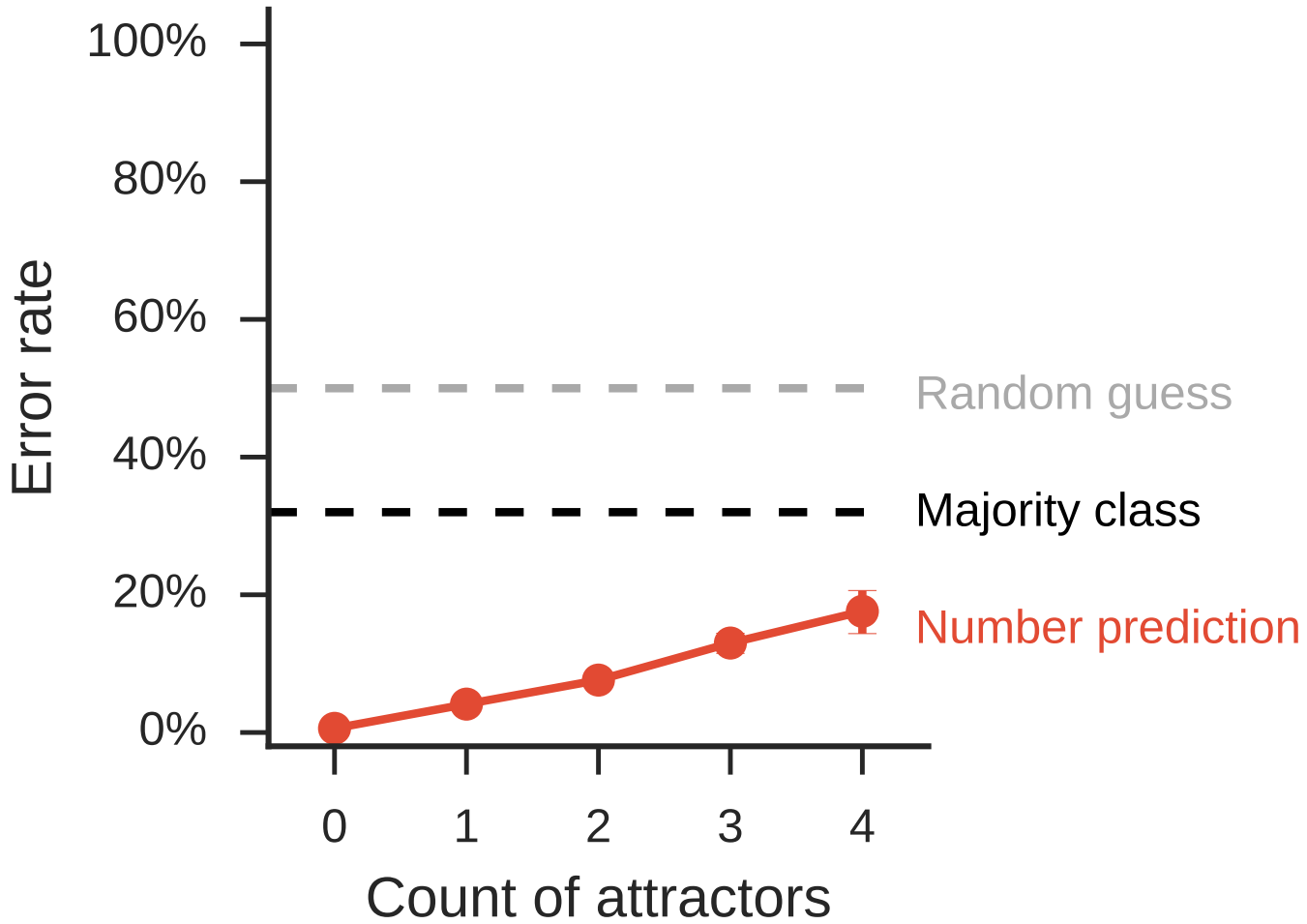

100%
80%
60%
40%
20%
0%
Error rate
0
1
2
3
4
Count of attractors
Random guess
Majority class
Number prediction

Attractors



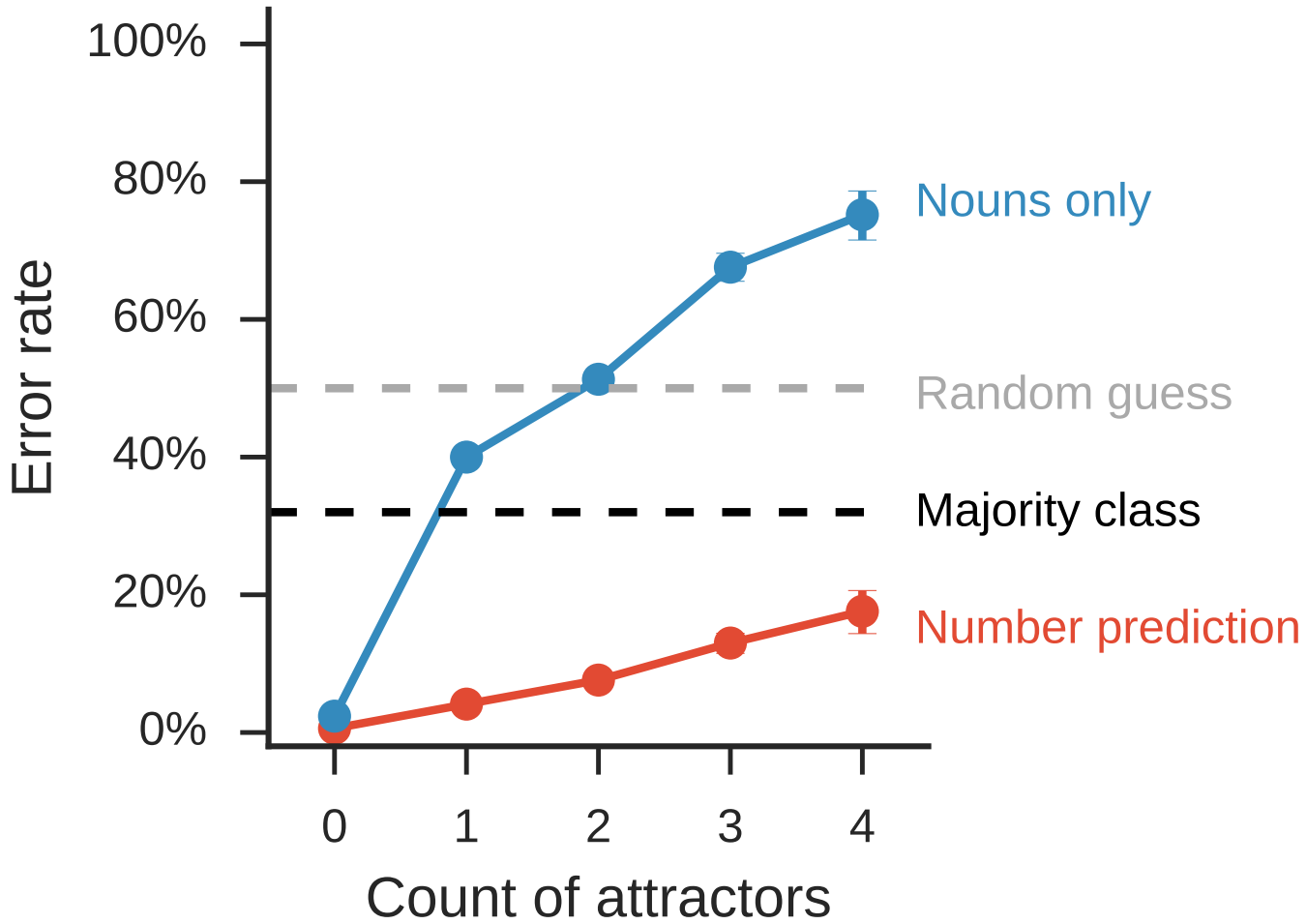
- Still at $< 20\%$ error with four attractors

Attractors



- Still at $< 20\%$ error with four attractors
- Looking at the last noun gets you 100% error!

Attractors



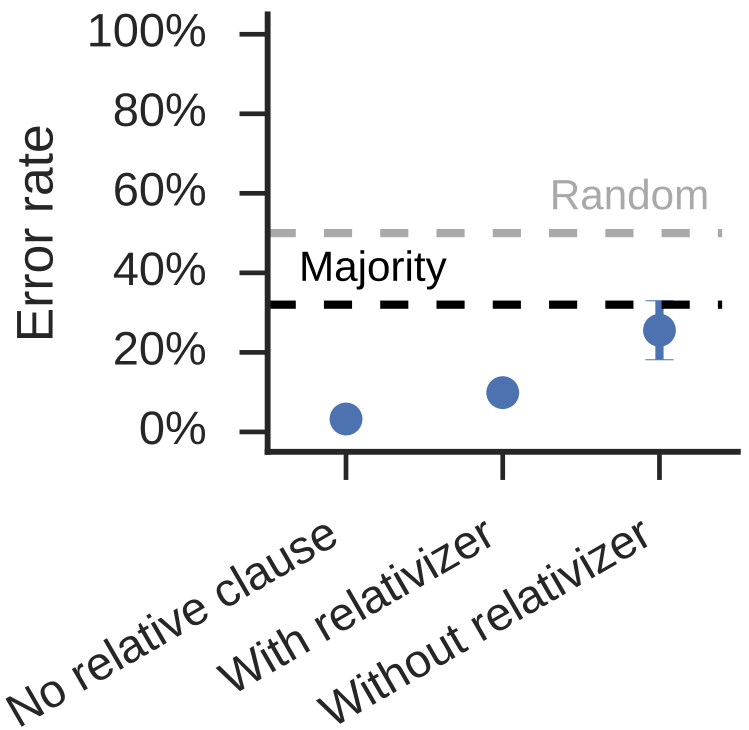
- Still at $< 20\%$ error with four attractors
- Looking at the last noun gets you 100% error!

Relative clauses

1. The landmarks **from** this article are not notable.
2. The landmarks **that** this article lists are not notable.
3. The landmarks this article lists are not notable.

Relative clauses

1. The landmarks **from** this article are not notable.
2. The landmarks **that** this article lists are not notable.
3. The landmarks this article lists are not notable.



Leaky representation of structure

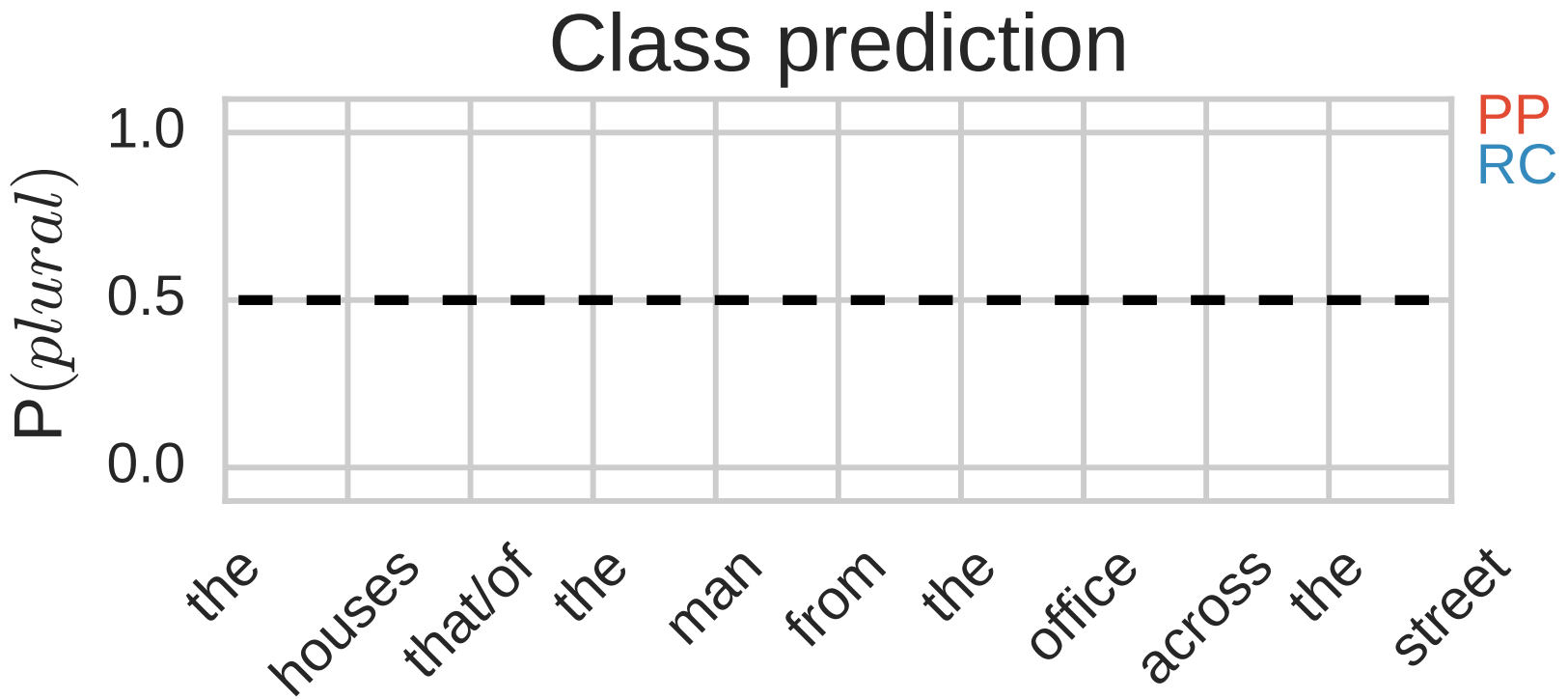
1. **PP:** The houses of the man from the office across the street...

Leaky representation of structure

1. **PP:** The houses of the man from the office across the street...
2. **RC:** The houses that the man from the office across the street...

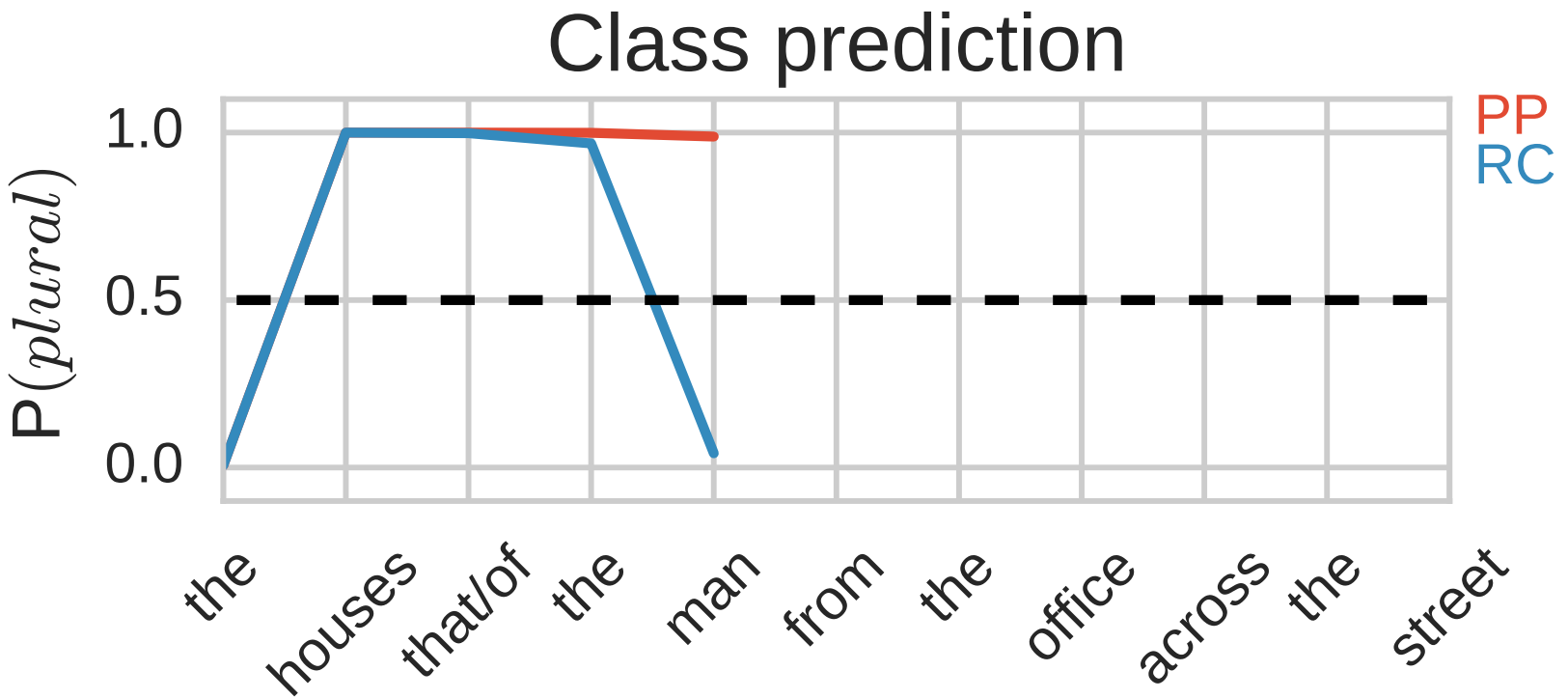
Leaky representation of structure

1. **PP:** The houses of the man from the office across the street...
2. **RC:** The houses that the man from the office across the street...



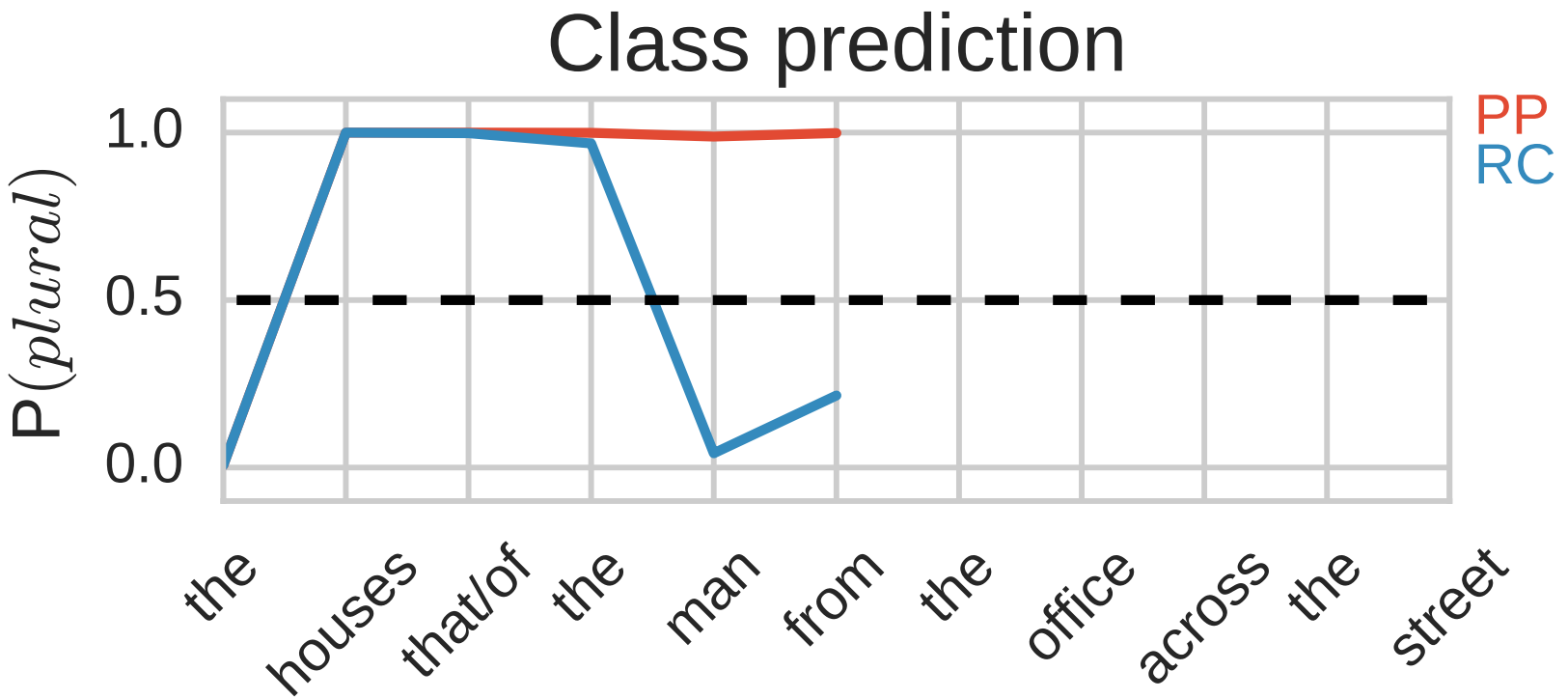
Leaky representation of structure

1. **PP:** The houses of the man from the office across the street...
2. **RC:** The houses that the man from the office across the street...



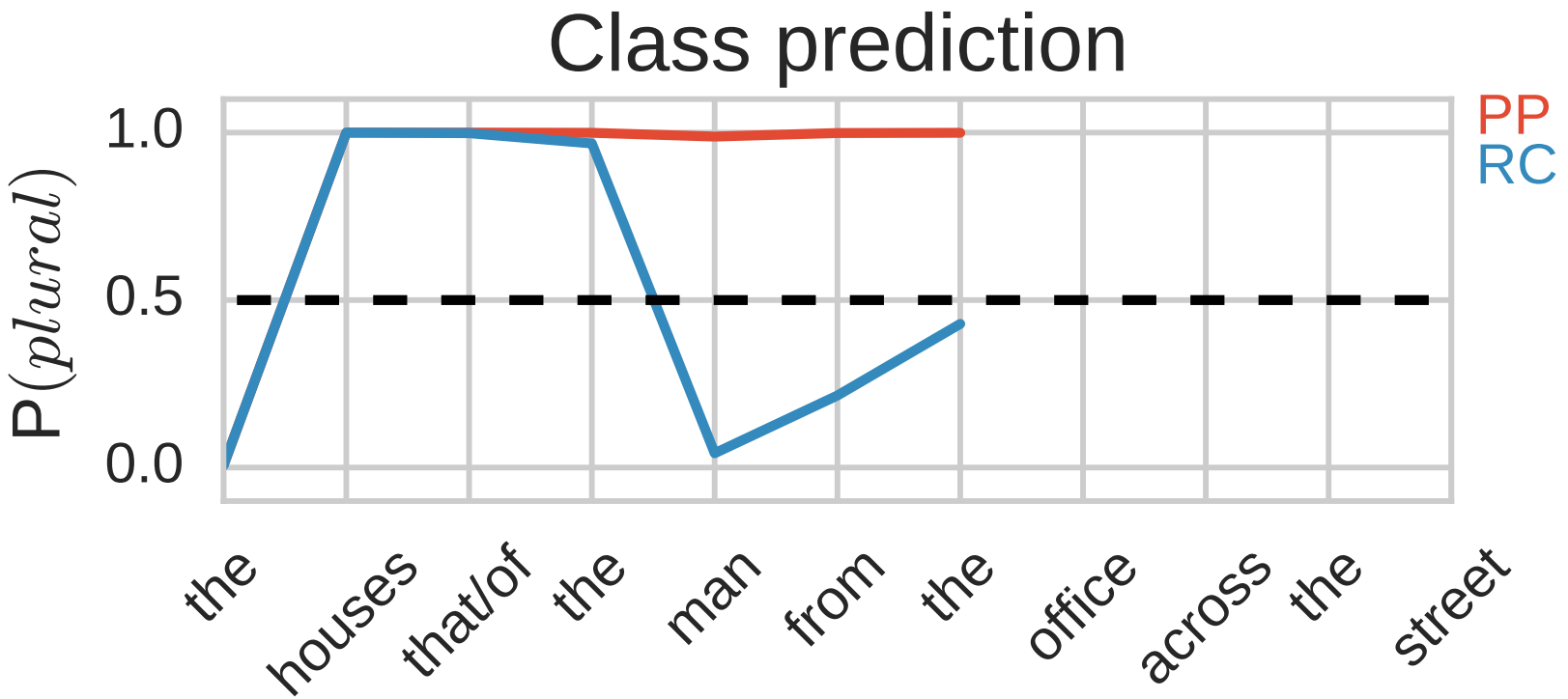
Leaky representation of structure

1. **PP:** The houses of the man from the office across the street...
2. **RC:** The houses that the man from the office across the street...



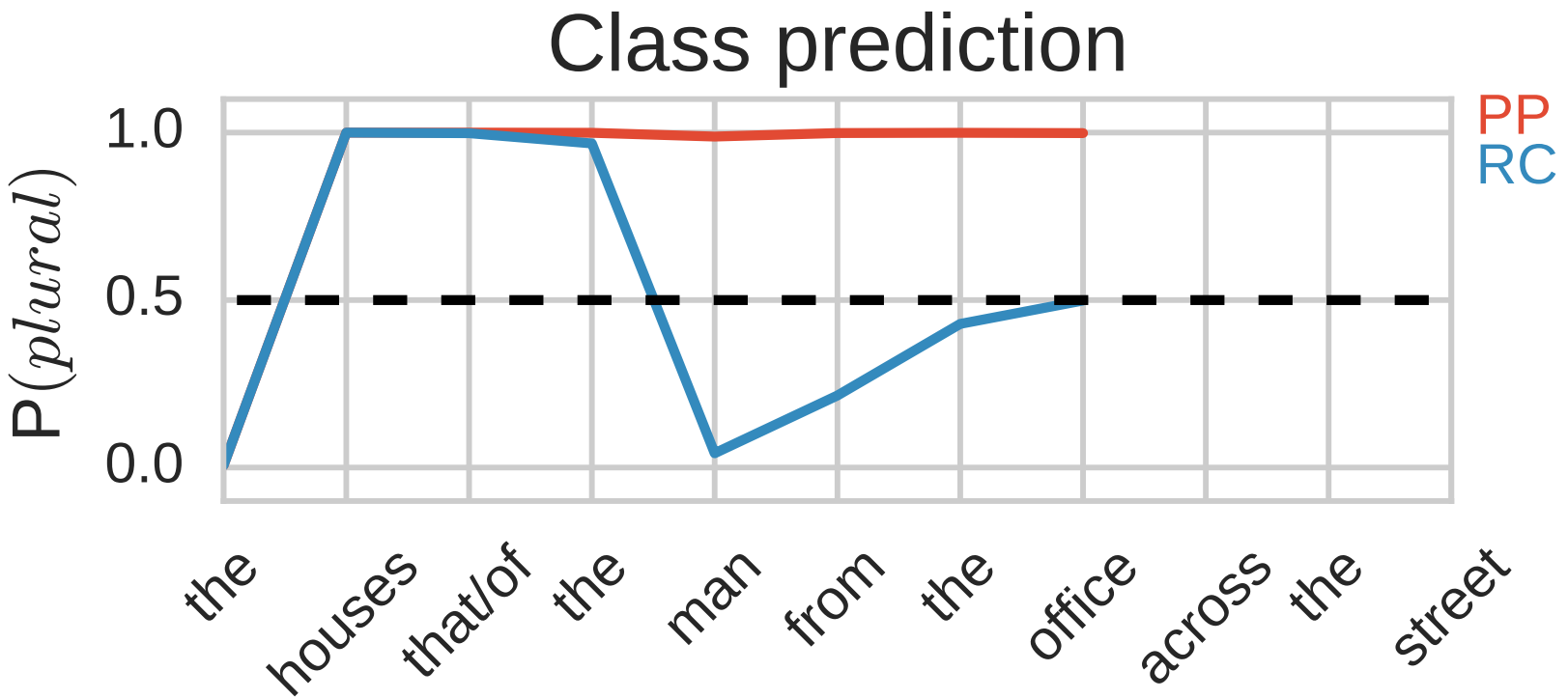
Leaky representation of structure

1. **PP:** The houses of the man from the office across the street...
2. **RC:** The houses that the man from the office across the street...



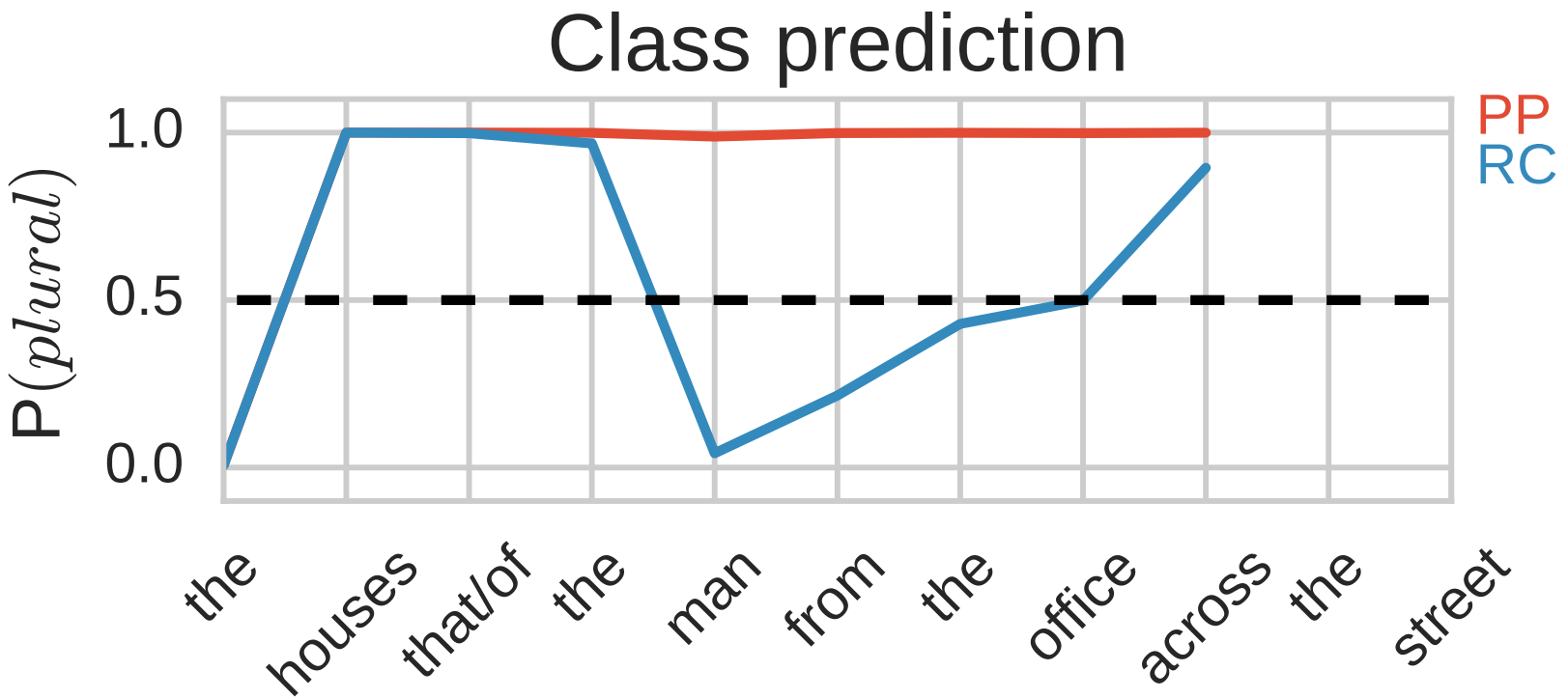
Leaky representation of structure

1. **PP:** The houses of the man from the office across the street...
2. **RC:** The houses that the man from the office across the street...



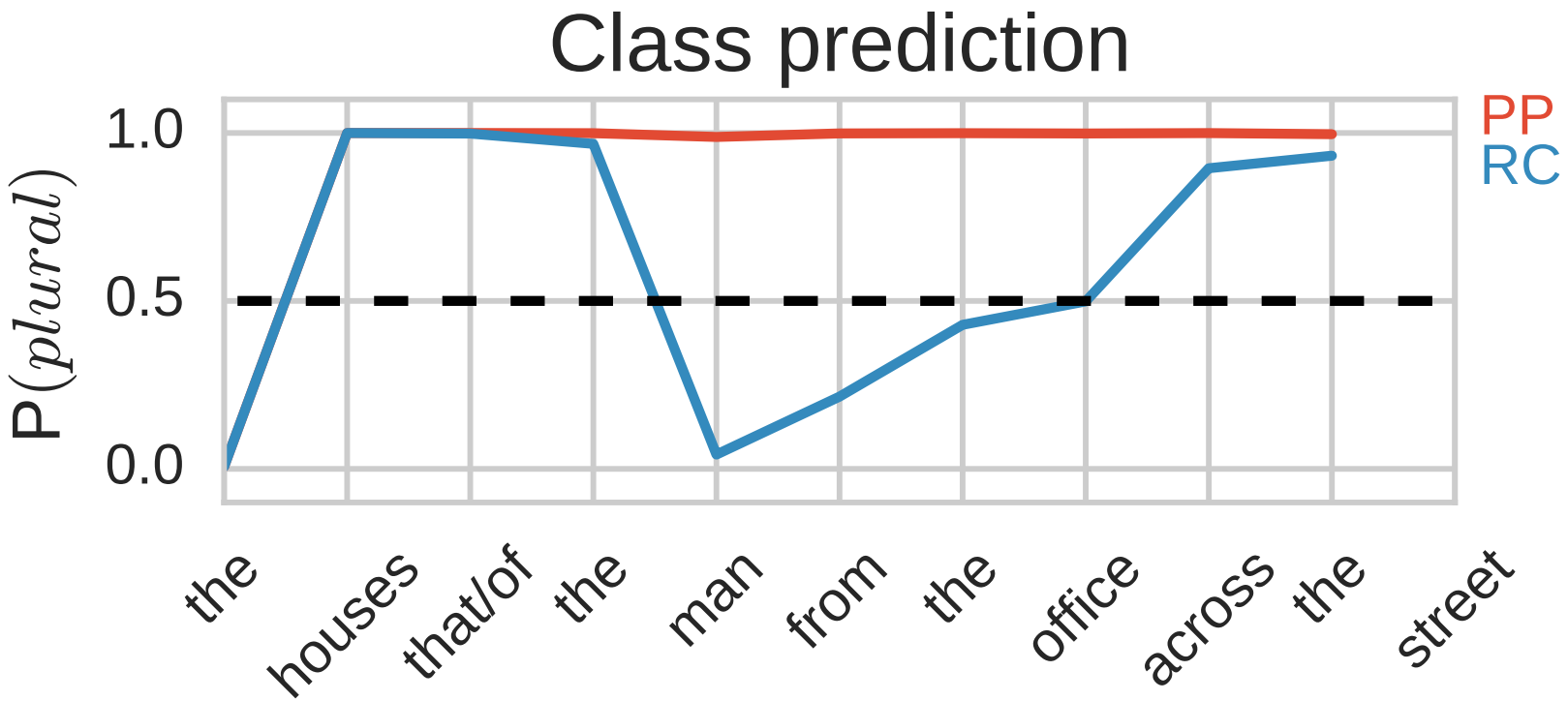
Leaky representation of structure

1. **PP:** The houses of the man from the office across the street...
2. **RC:** The houses that the man from the office across the street...



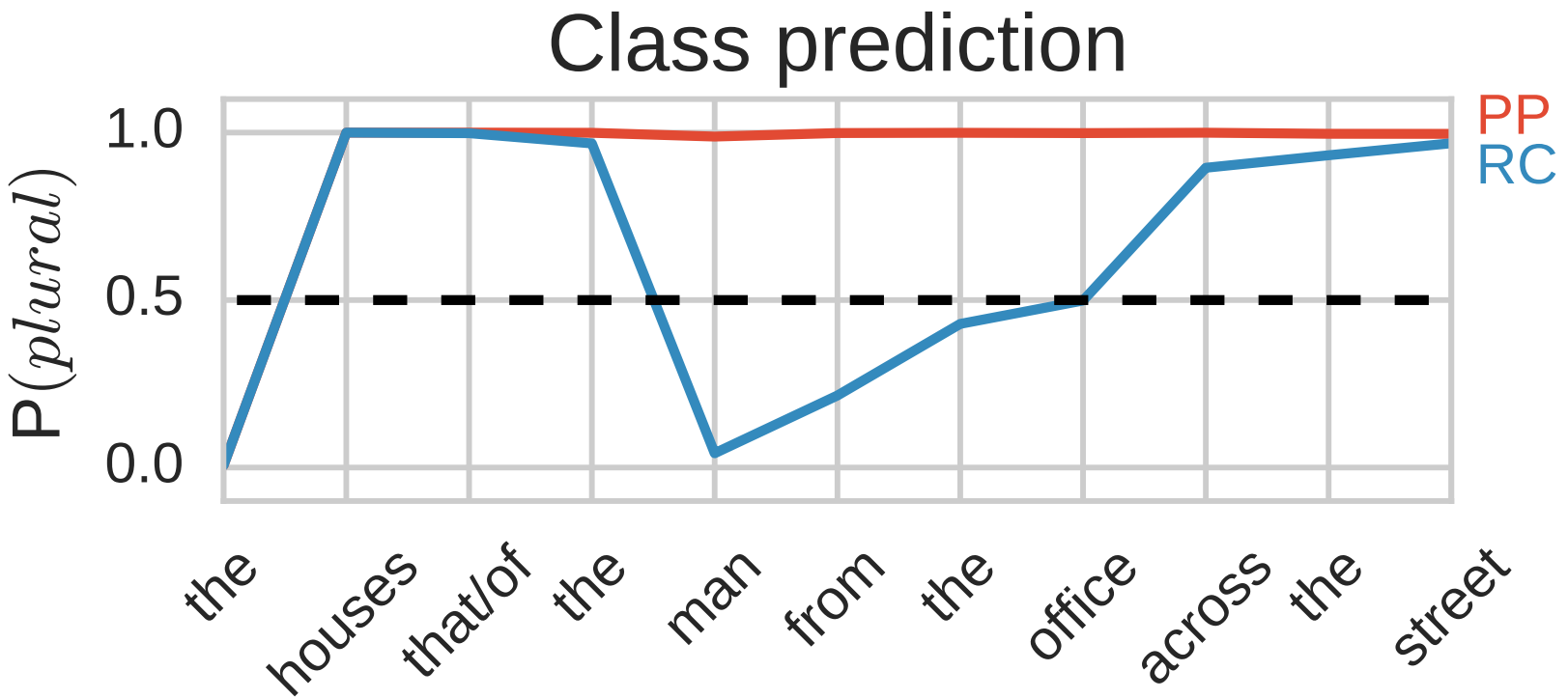
Leaky representation of structure

1. **PP:** The houses of the man from the office across the street...
2. **RC:** The houses that the man from the office across the street...

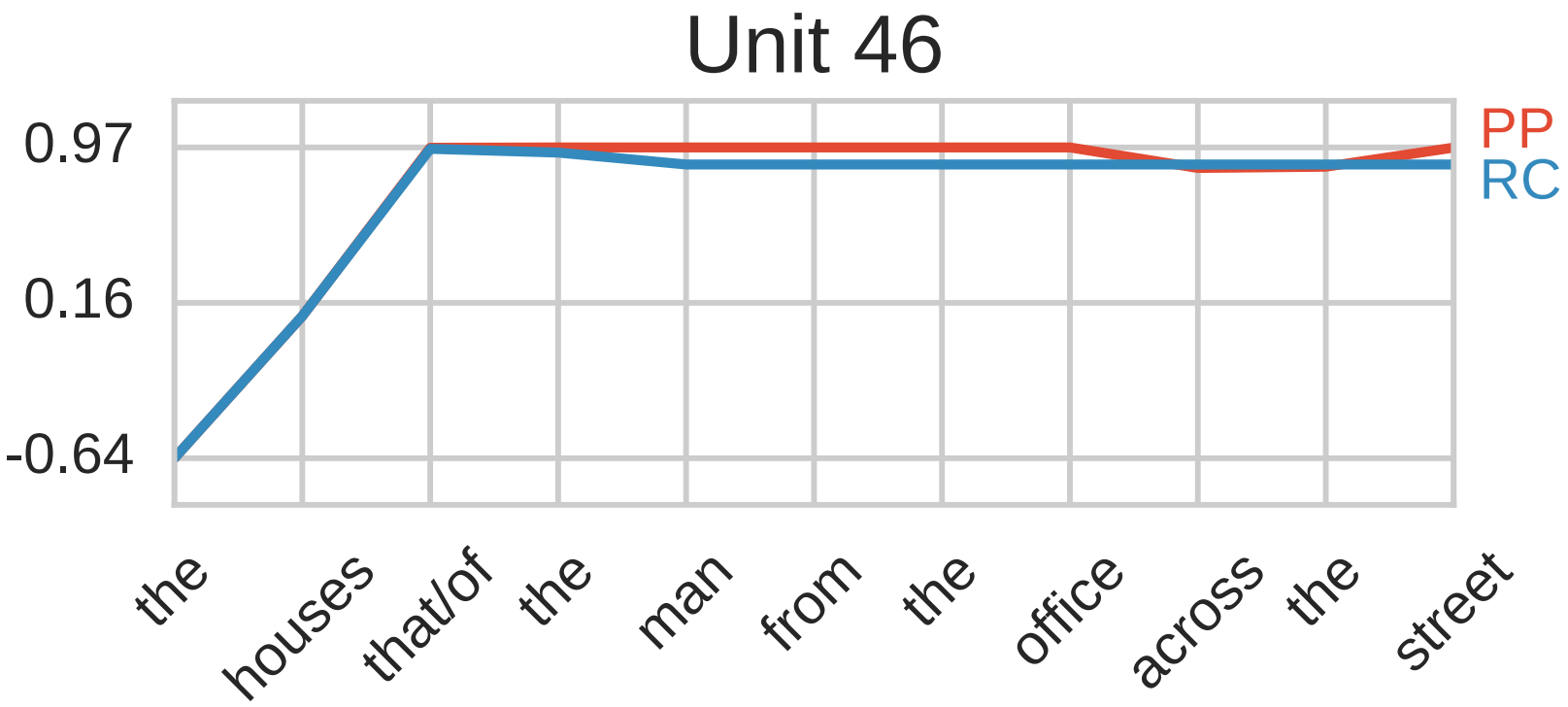


Leaky representation of structure

1. **PP:** The houses of the man from the office across the street...
2. **RC:** The houses that the man from the office across the street...

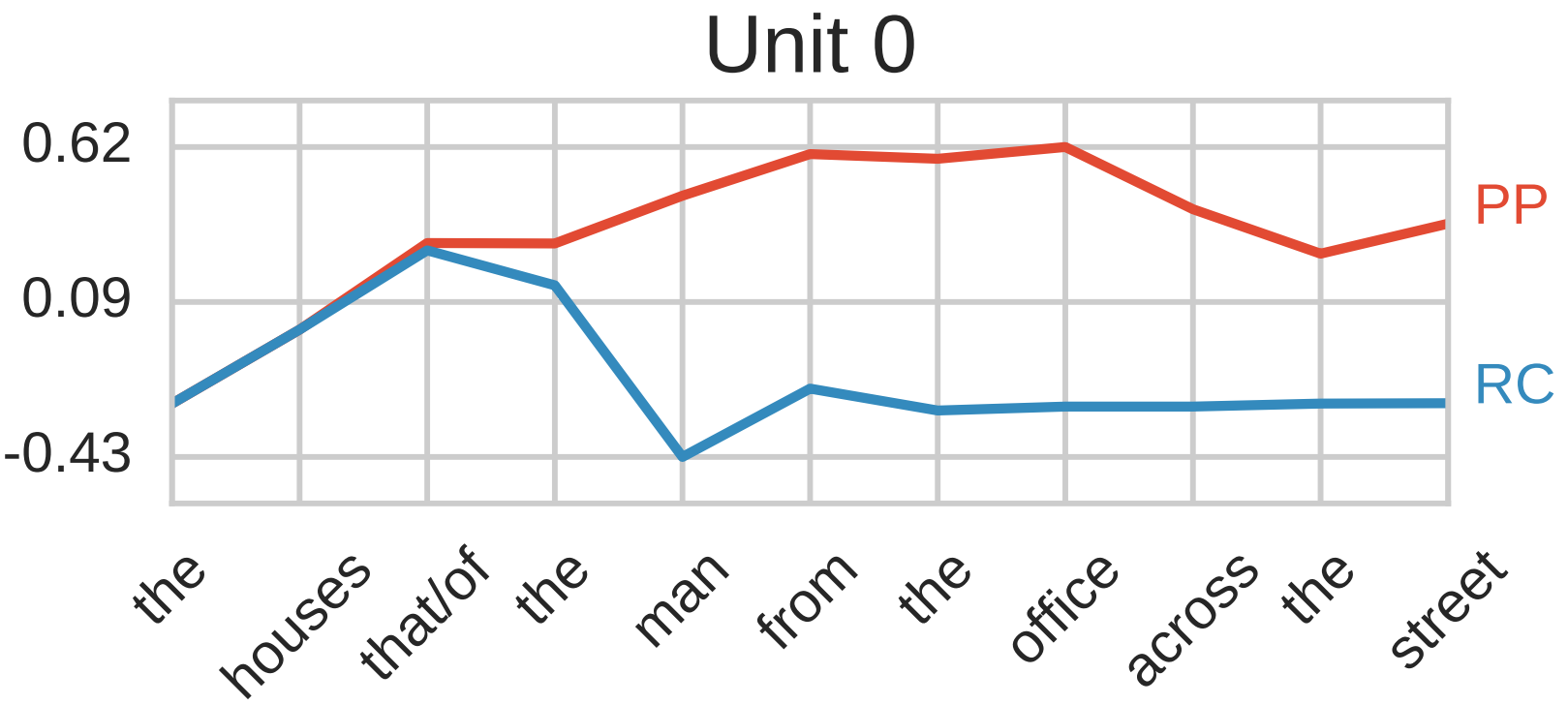


Leaky representation of structure



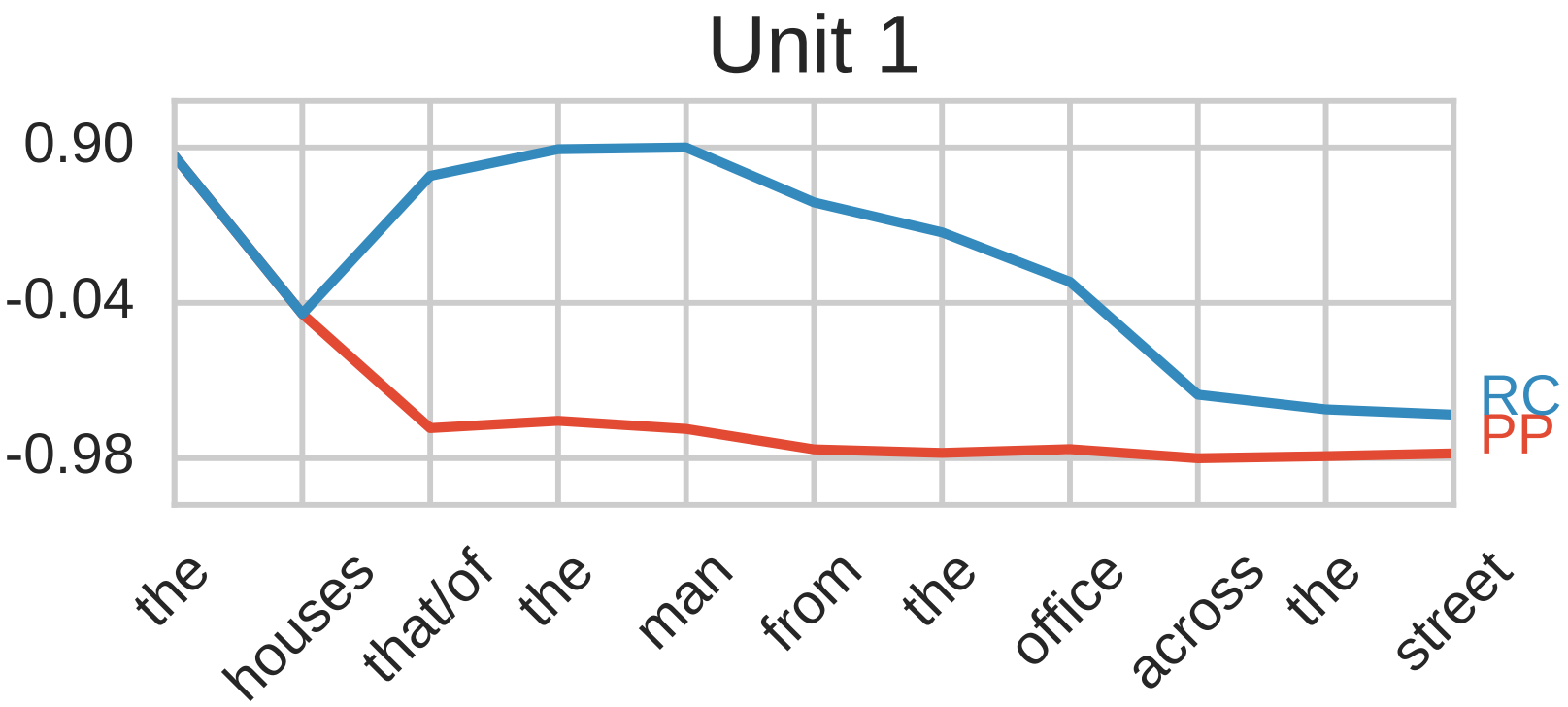
A unit that represents the number of the main clause subject!

Leaky representation of structure



A unit that represents the number of the embedded subject!

Leaky representation of structure



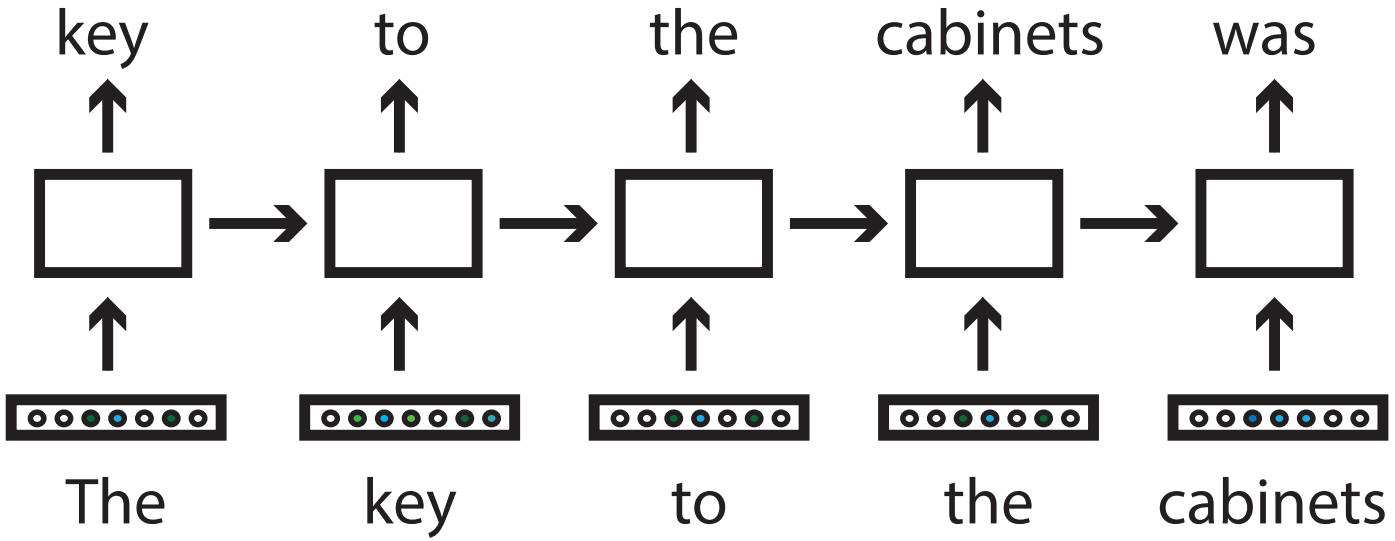
A unit that, uh...

Outline

- The number prediction task
- An RNN trained specifically to perform the task
- A focus on structurally interesting sentences
- RNN trained on a language modeling objective

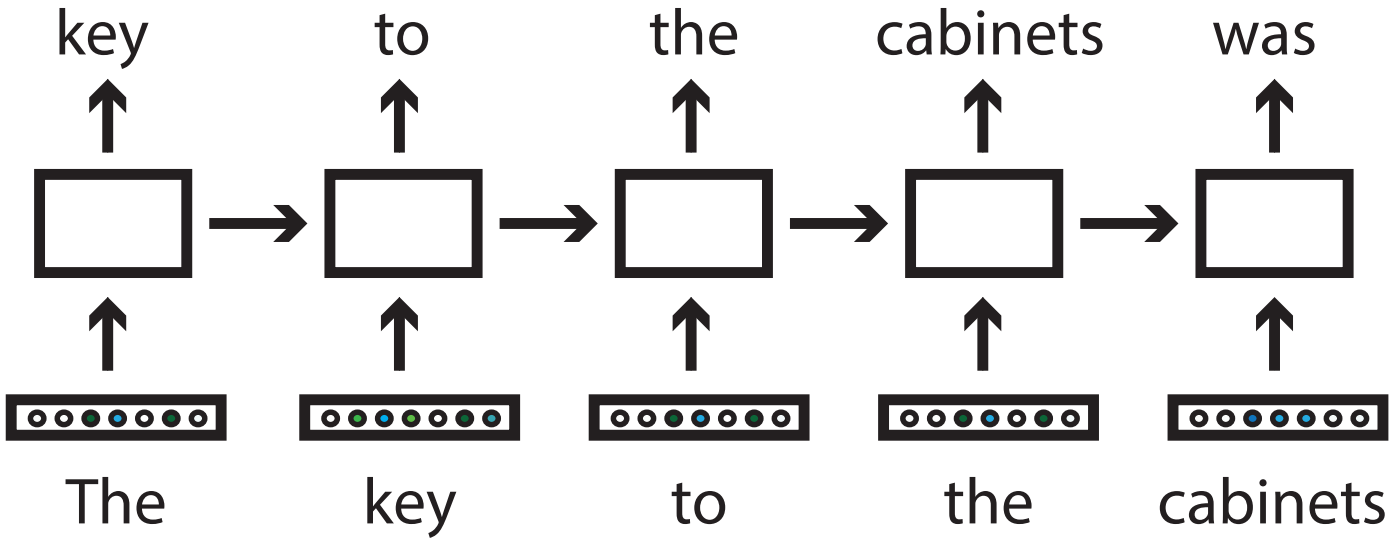
Language model

Training:



Language model

Training:

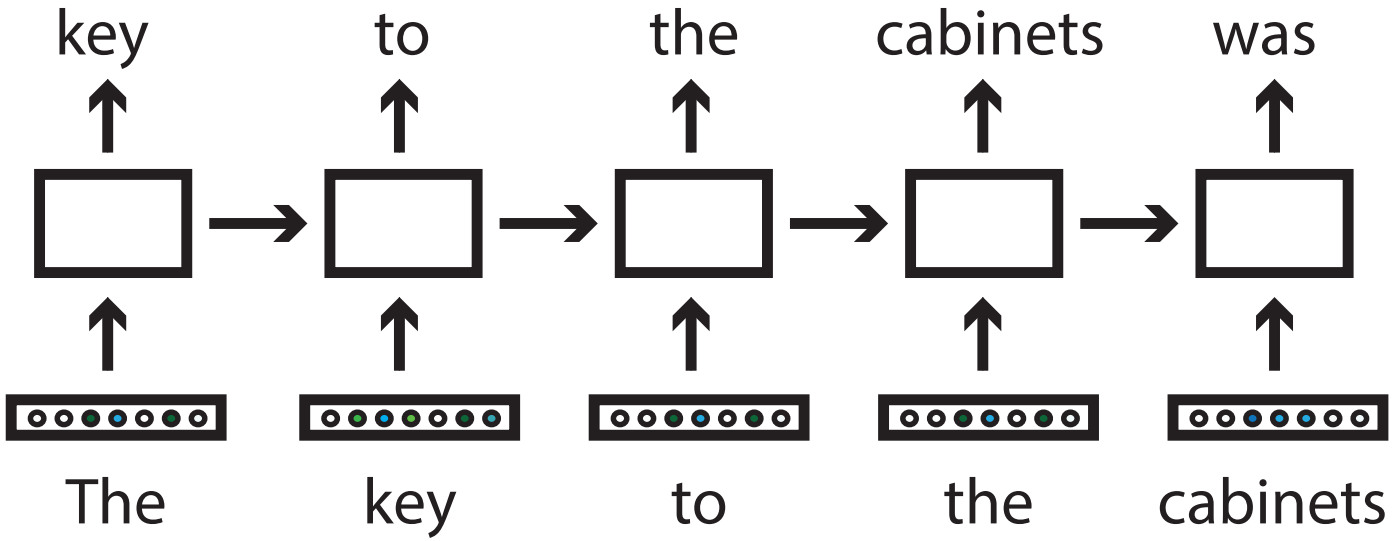


Evaluation:

The length of the forewings...

Language model

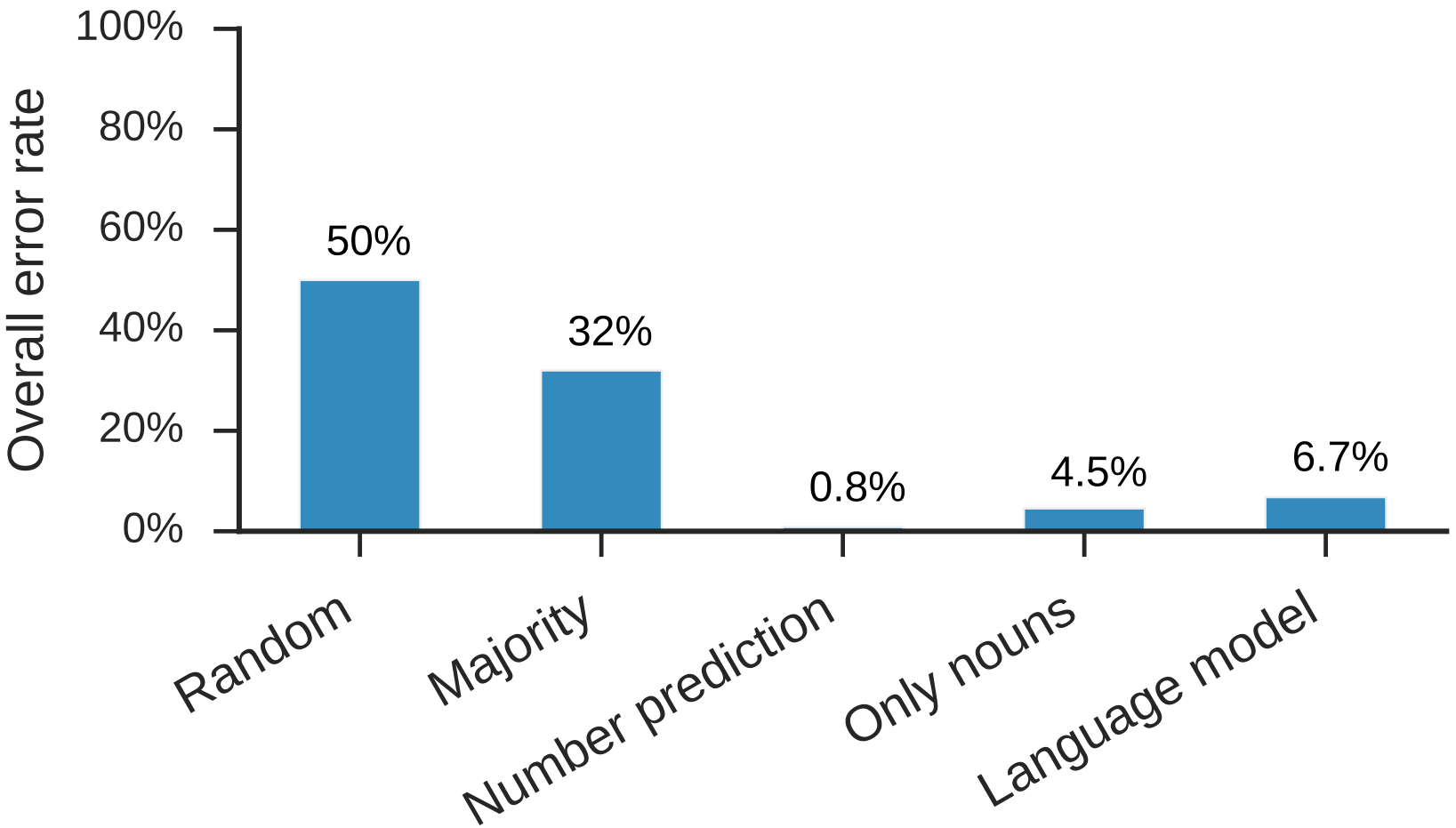
Training:



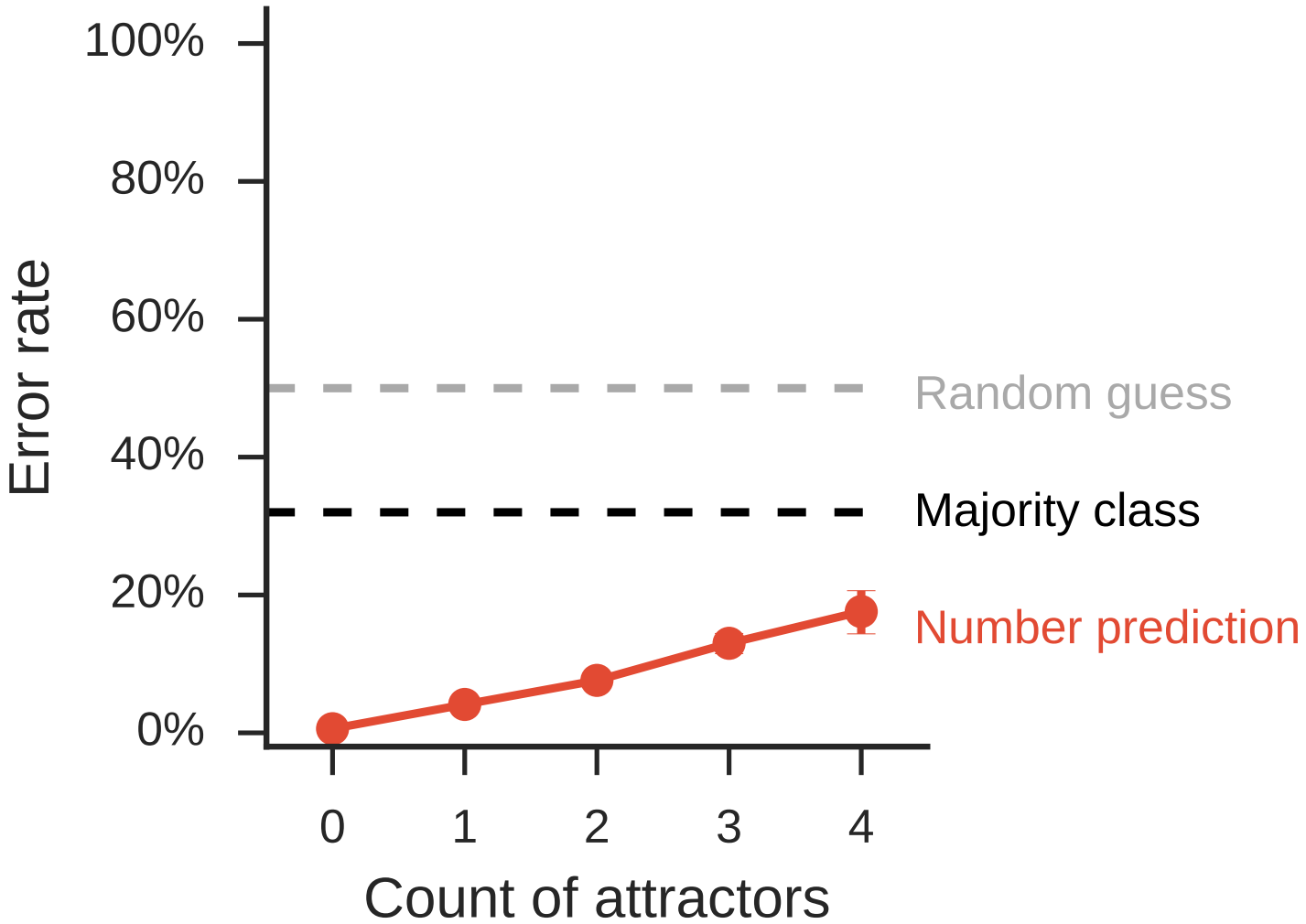
Evaluation:

The length of the forewings... $P(\text{is}) > P(\text{are})$?

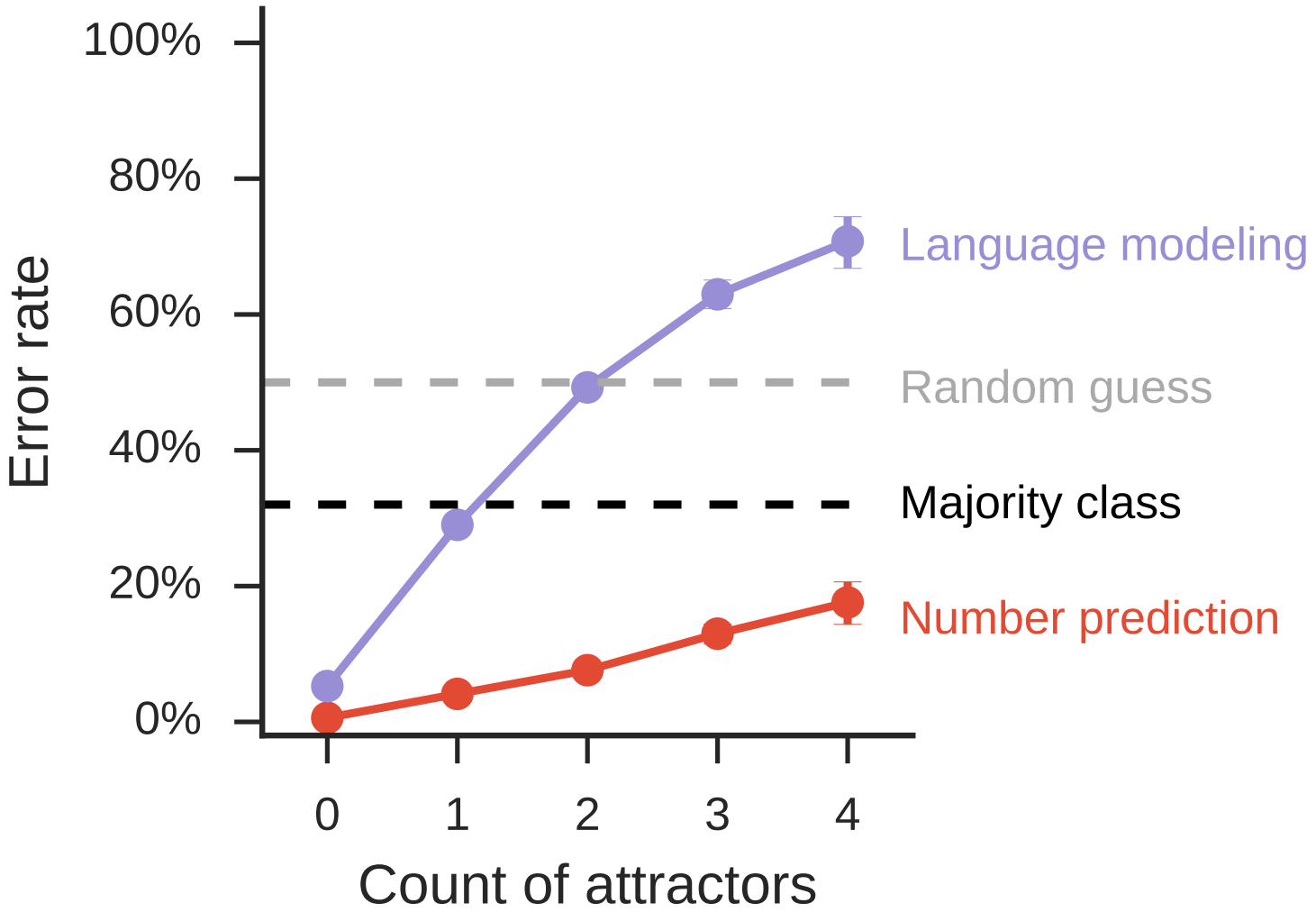
Overall results



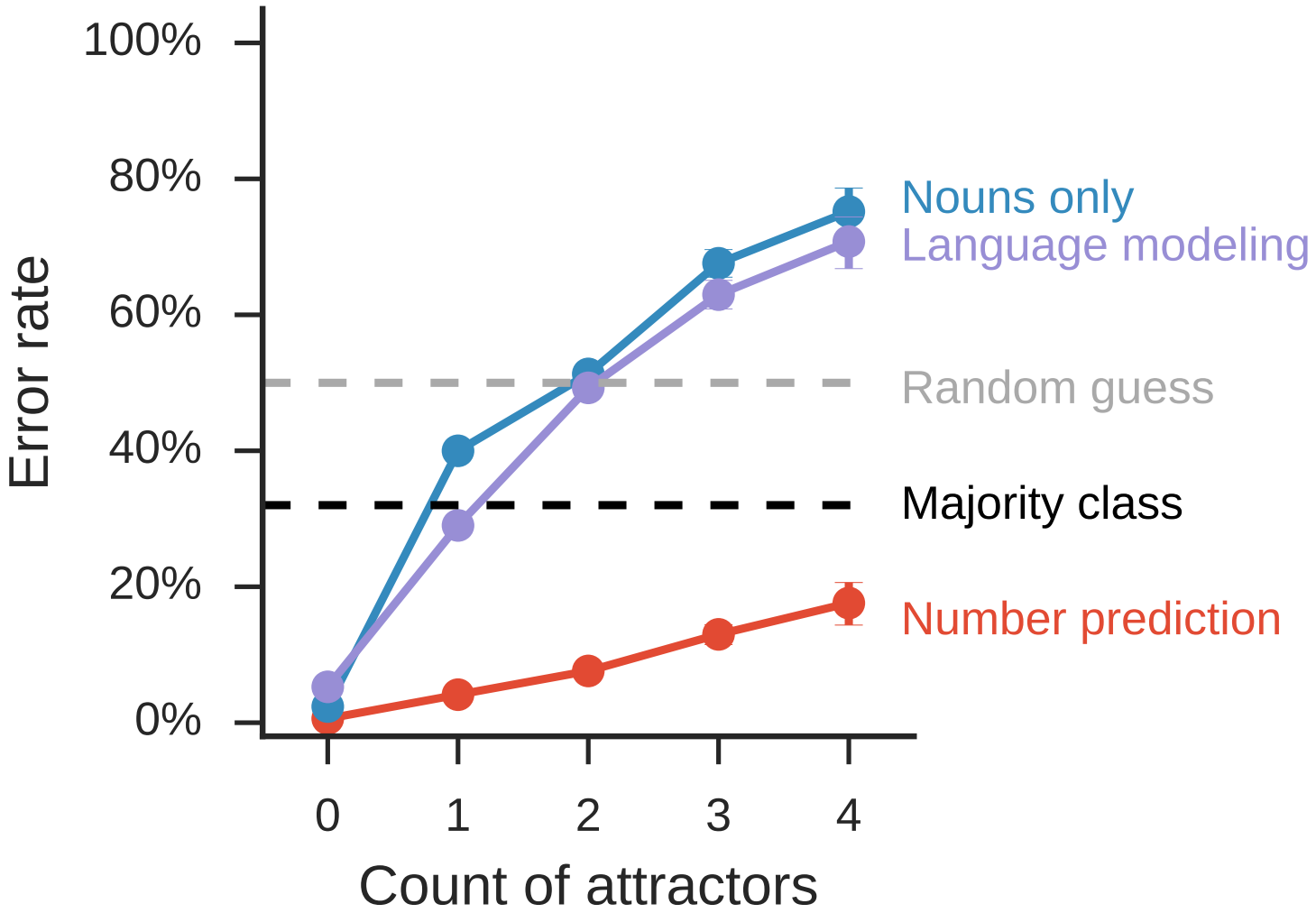
Attraction errors



Attraction errors



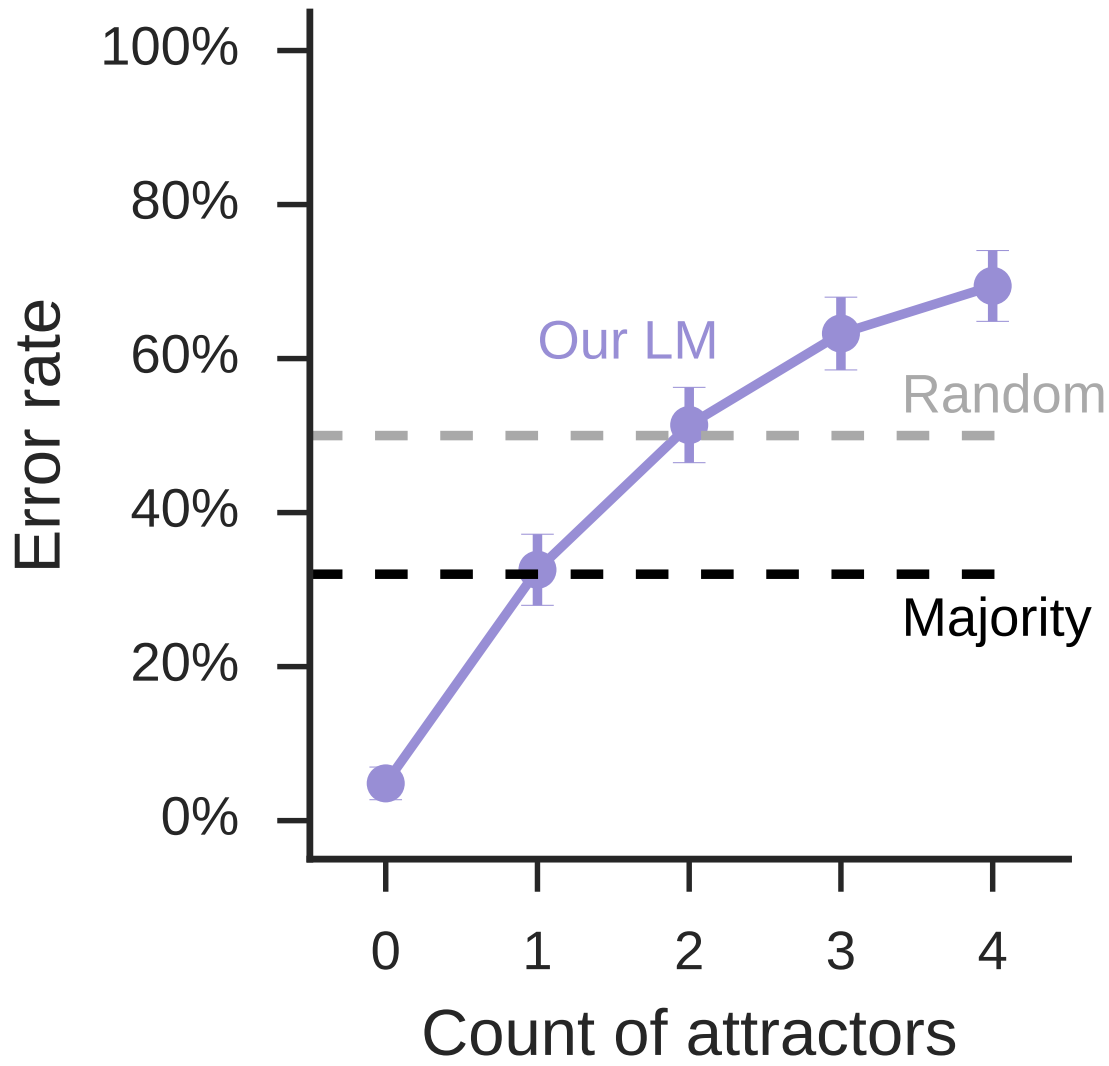
Attraction errors

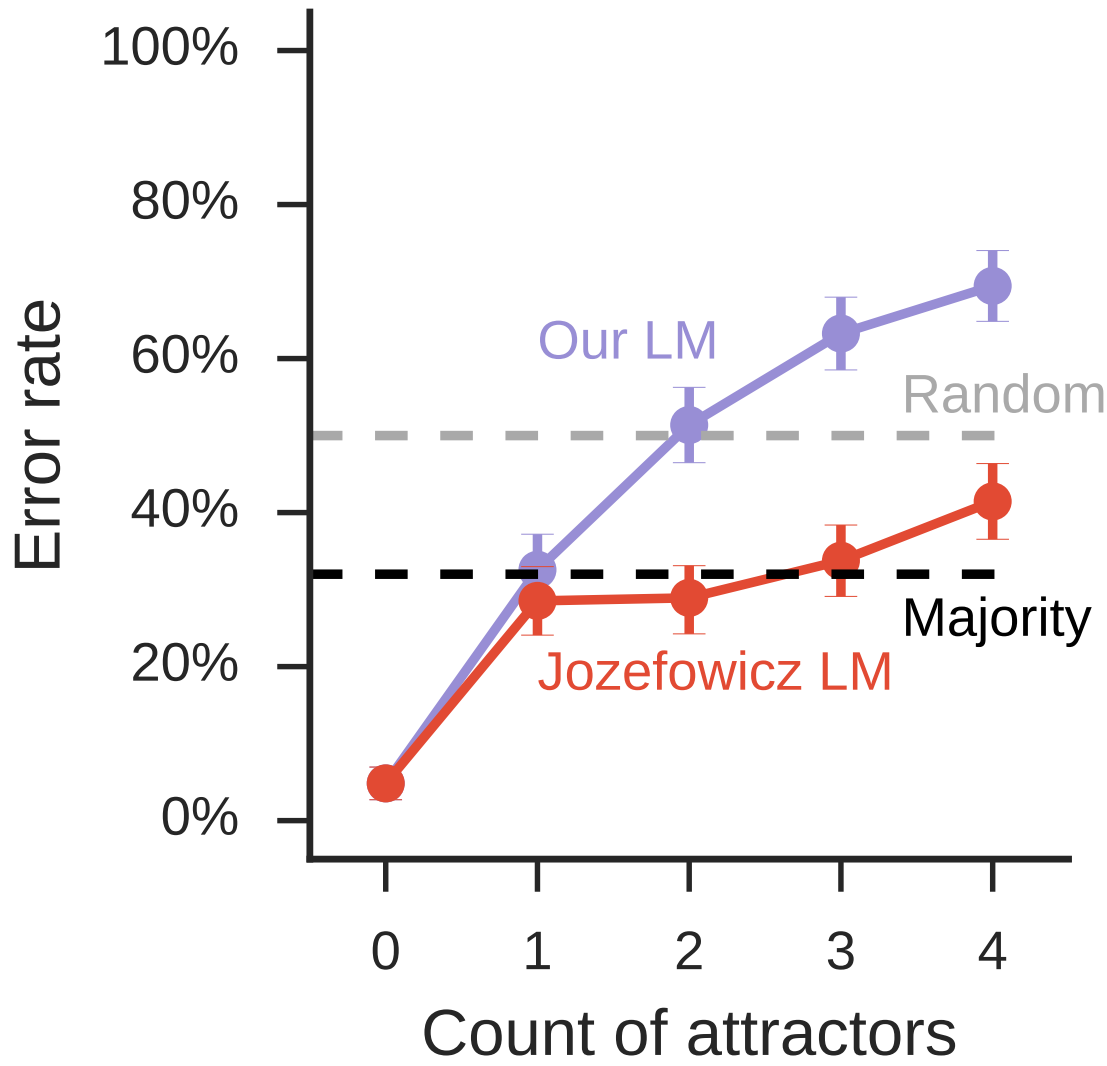


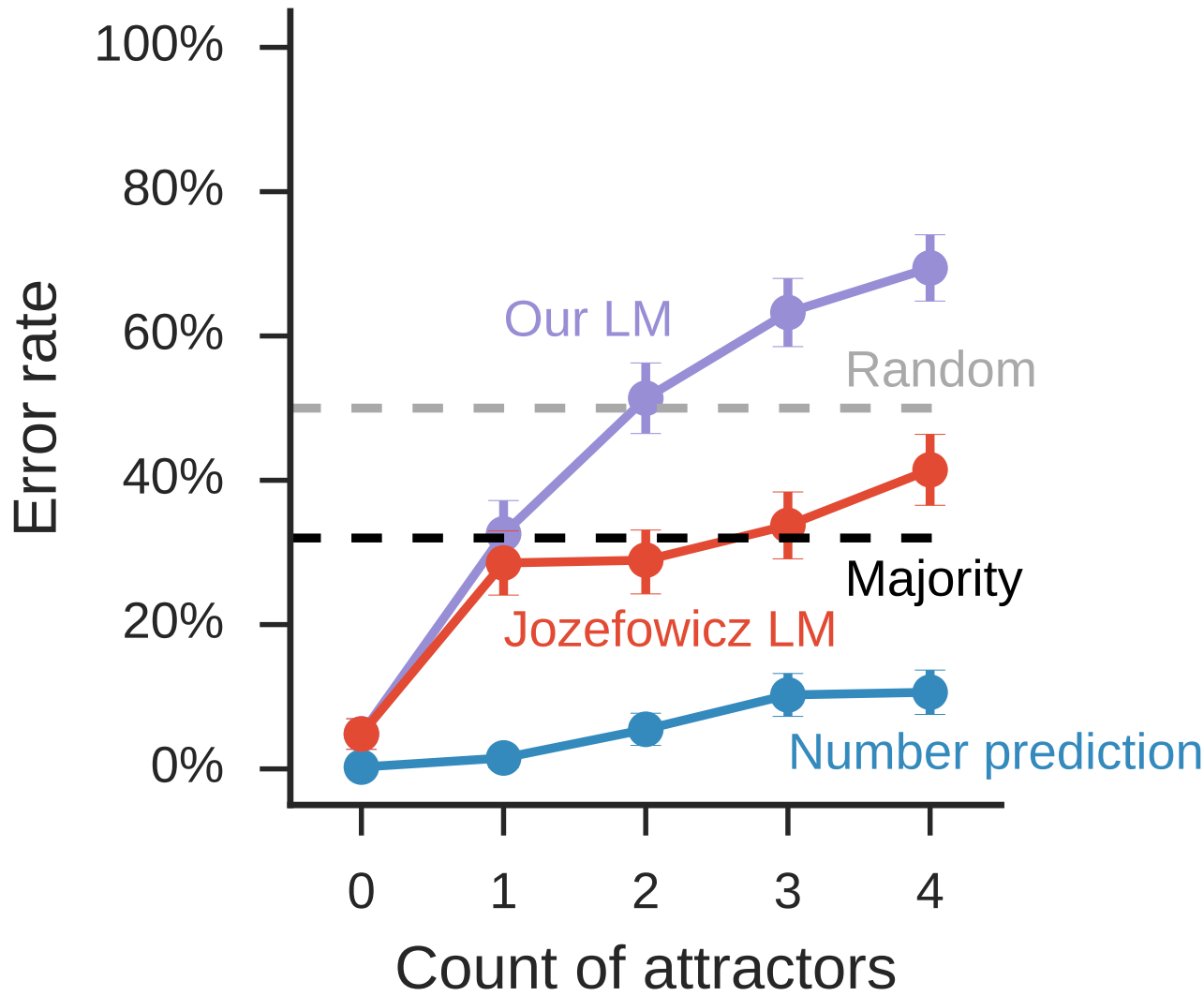
- The language modeling objective leads to dramatically worse results

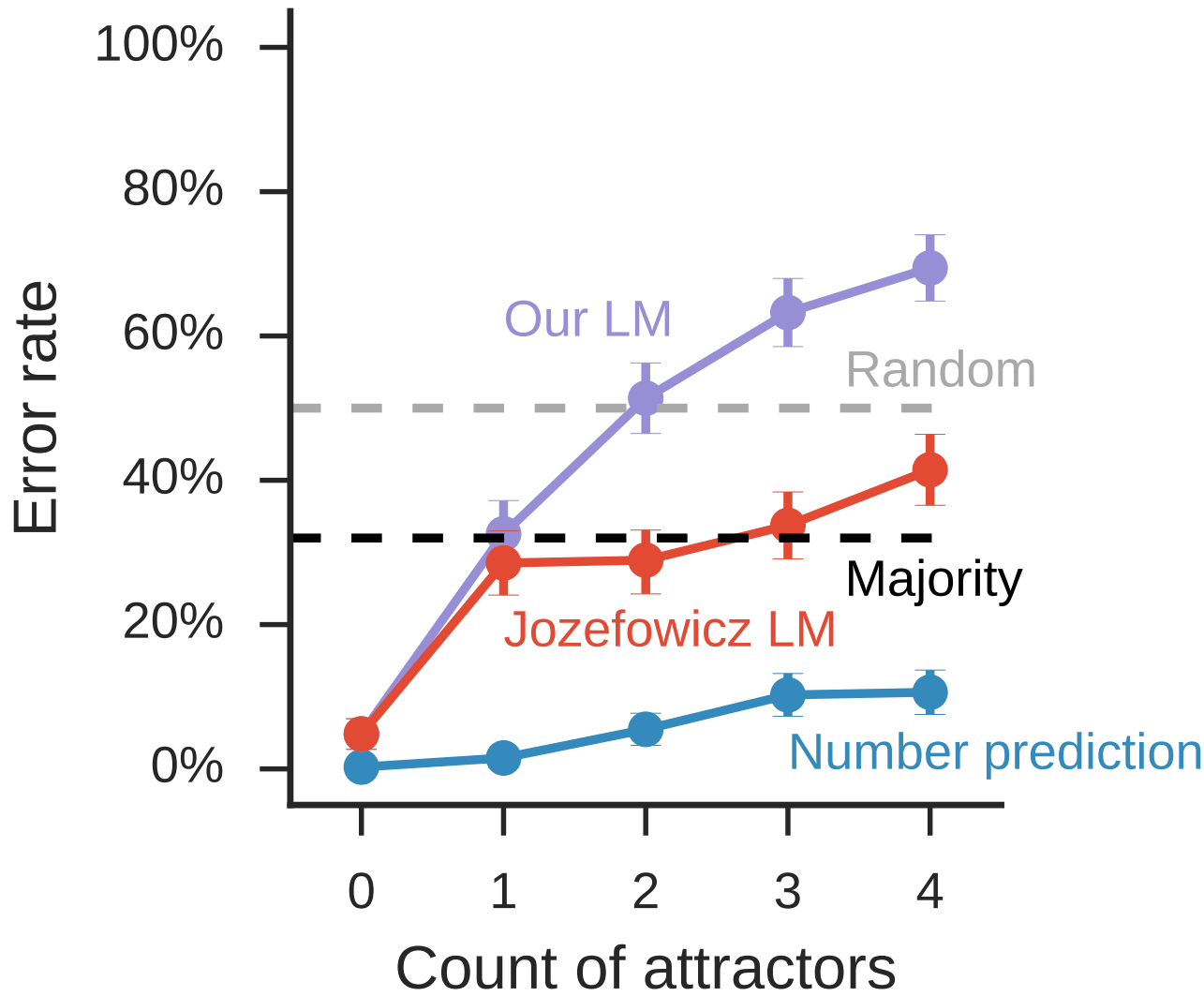
- The language modeling objective leads to dramatically worse results
- Is it because our LM was too small?

- The language modeling objective leads to dramatically worse results
- Is it because our LM was too small?
- Test with the Jozefowicz et al. (2016) language model (2 layers of 8192 LSTM units each)


100%
80%
60%
40%
20%
0%
Error rate
Our LM
Random
Majority
0
1
2
3
4
Count of attractors


100%
80%
60%
40%
20%
0%
Error rate
Our LM
Random
Majority
Jozefowicz LM
0
1
2
3
4
Count of attractors


100%
80%
60%
40%
20%
0%
Error rate
Our LM
Random
Majority
Jozefowicz LM
Number prediction
0
1
2
3
4
Count of attractors



A very large LSTM LM is only somewhat better

Summary of findings

- RNNs performed well overall given grammatical instruction

Summary of findings

- RNNs performed well overall given grammatical instruction
- But were still tripped up by difficult cases (e.g., long relative clauses): representations of structure are leaky

Summary of findings

- RNNs performed well overall given grammatical instruction
- But were still tripped up by difficult cases (e.g., long relative clauses): representations of structure are leaky
- Dramatic decline in performance on difficult cases with language modeling objective (even with SOTA LM)

Summary of findings

- RNNs performed well overall given grammatical instruction
- But were still tripped up by difficult cases (e.g., long relative clauses): representations of structure are leaky
- Dramatic decline in performance on difficult cases with language modeling objective (even with SOTA LM)
- Many more experiments in the paper and in ongoing work

Conclusions

- The agreement task is well-suited to measuring progress in implicit structural learning in neural networks

Conclusions

- The agreement task is well-suited to measuring progress in implicit structural learning in neural networks
- **Evaluate your LMs on this task rather than just perplexity!**

Conclusions

- The agreement task is well-suited to measuring progress in implicit structural learning in neural networks
- **Evaluate your LMs on this task rather than just perplexity!**
- Linguistics and psycholinguistics can help characterize a learner's strengths and weaknesses

Acknowledgements

- Marco Baroni
- Grzegorz Chrupała
- Sol Lago
- Paul Smolensky
- Whitney Tabor
- Roberto Zamparelli

- European Research Council (grant ERC-2011-AdG 295810 BOOTPHON)
- Agence Nationale pour la Recherche (grants ANR-10-IDEX-0001-02 PSL and ANR-10-LABX-0087 IEC)
- Israeli Science Foundation (grant number 1555/15)

Thank you!

Bahdanau, D., Cho, K., & Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International conference for learning representations*.

Bock, K., & Miller, C. A. (1991). Broken agreement. *Cognitive Psychology*, *23*(1), 45–93.

Elman, J. L. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*(2-3), 195–225.

Goldberg, Y., & Nivre, J. (2012, December). A dynamic oracle for arc-eager dependency parsing. In *Proceedings of COLING 2012* (pp. 959–976).

Jozefowicz, R., Vinyals, O., Schuster, M., Shazeer, N., & Wu, Y. (2016). Exploring the limits of language modeling. *arXiv preprint arXiv:1602.02410*.

Mikolov, T., Karafiát, M., Burget, L., Cernockỳ, J., & Khudanpur, S. (2010). Recurrent neural network based language model. In *Proceedings of the 11th Annual Conference of the International Speech Communication Association (INTERSPEECH 2010)* (pp. 1045–1048). Makuhari, Chiba, Japan.

Sundermeyer, M., Schlüter, R., & Ney, H. (2012). LSTM neural networks for language modeling. In *Proceedings of the 13th Annual Conference of the International Speech Communication Association (INTERSPEECH)* (pp. 194–197).

Vinyals, O., Kaiser, Ł., Koo, T., Petrov, S., Sutskever, I., & Hinton, G. (2015). Grammar as a foreign language. In *Advances in neural information processing systems* (pp. 2755–2763).