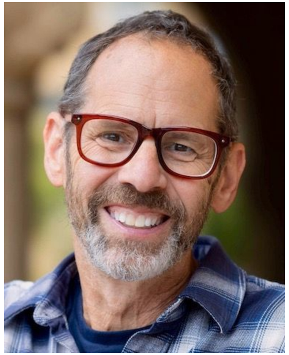
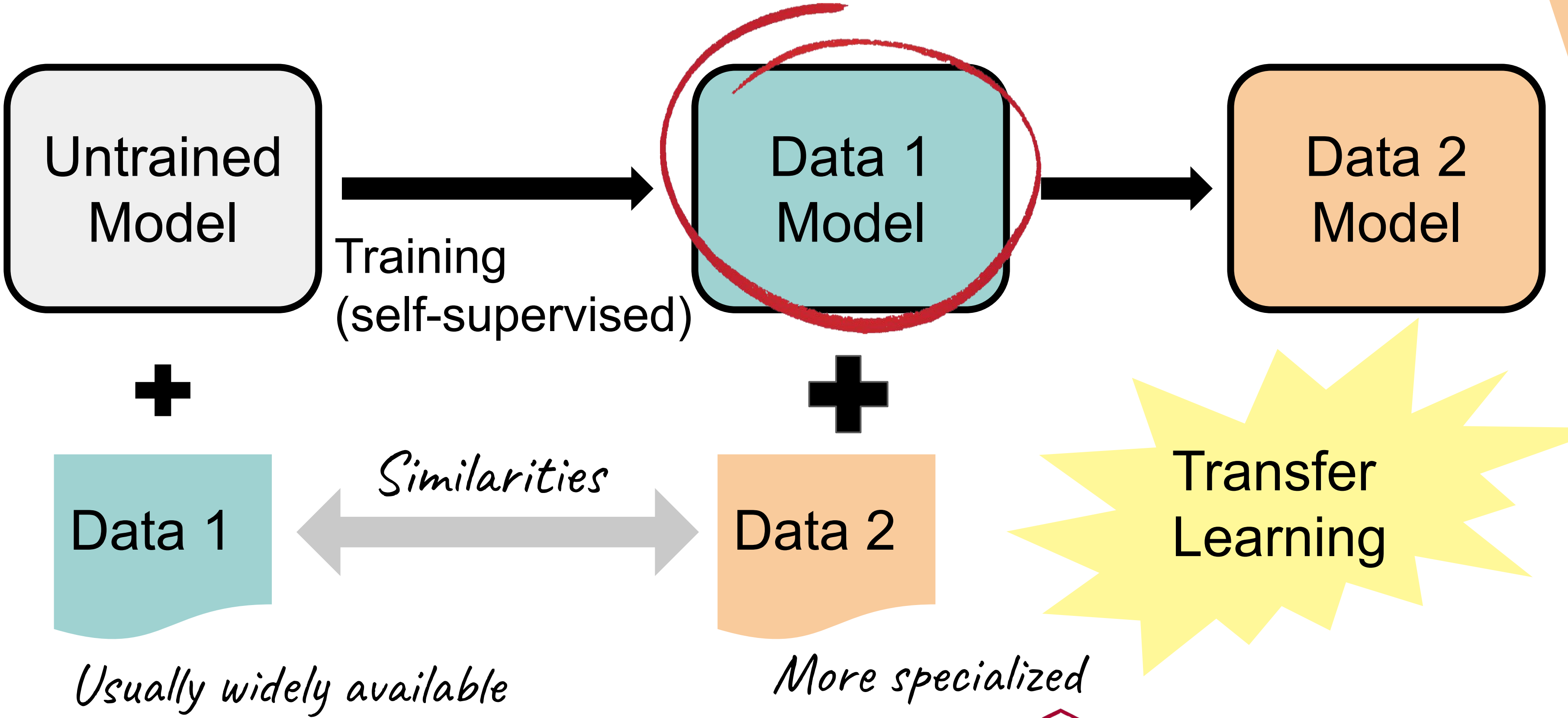
Structural Transfer Learning: Exploring Neural Models and Language Structure Through Understanding Transfer

Stanford | NLP

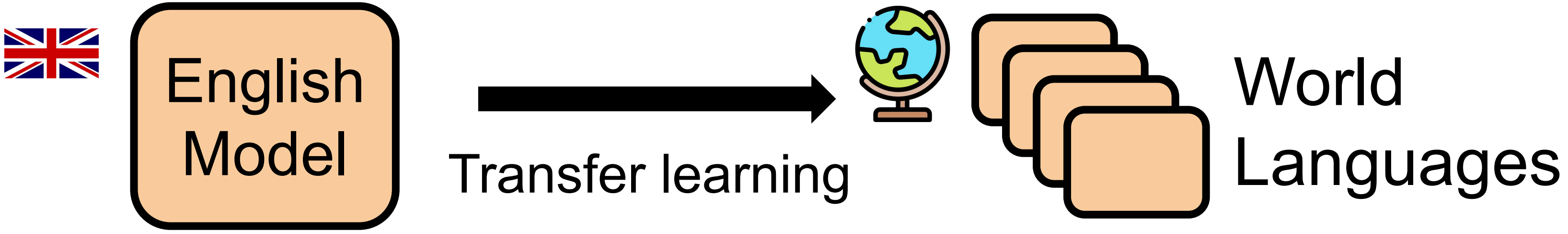
Isabel Papadimitriou



Transfer Learning



Transfer learning has practical applications



But also an **analysis methodology** for understanding data and learning

- Power machine learning models let us explore questions about language in new ways

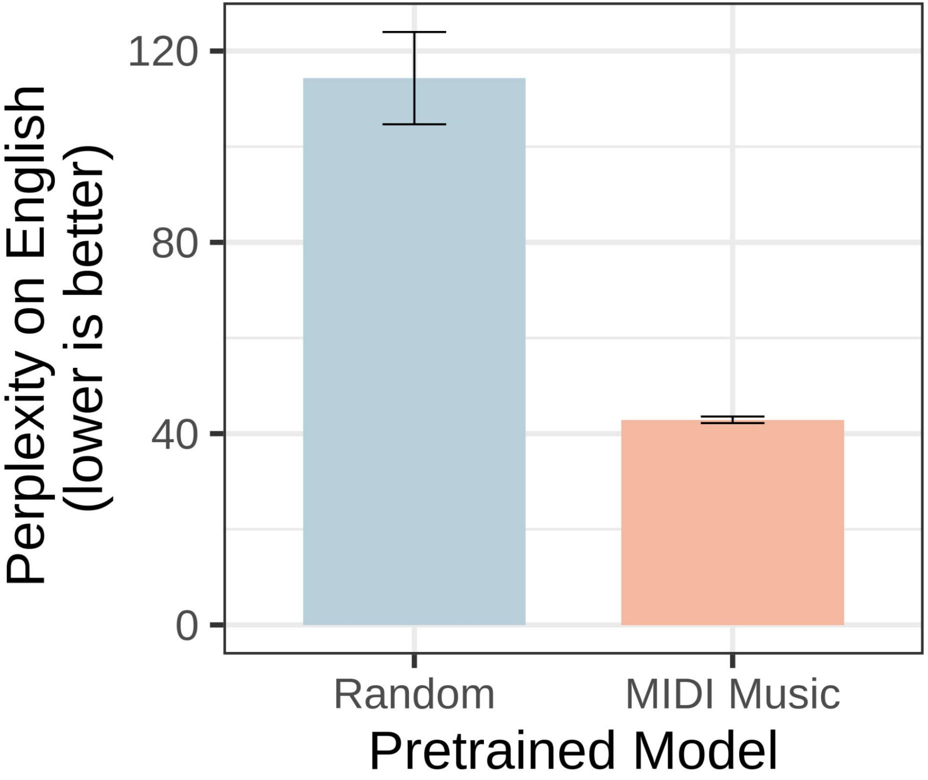
Shared structures between modalities

Data 1

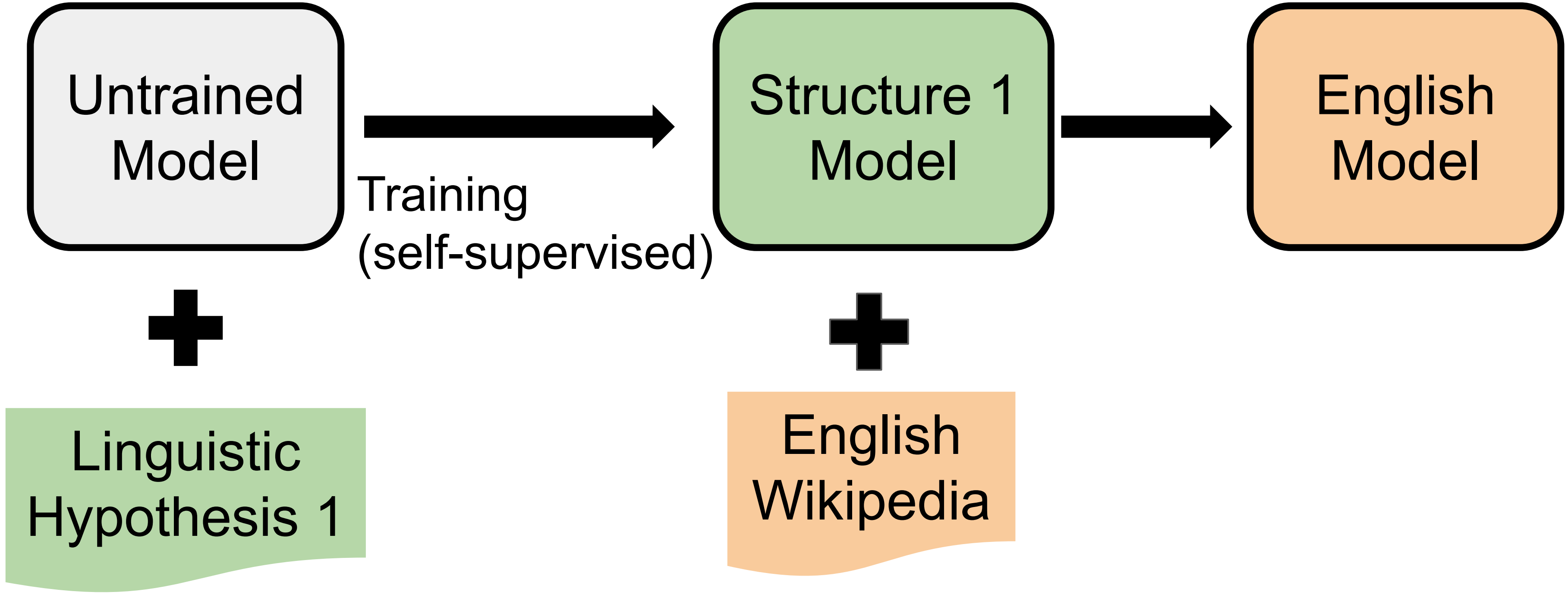
Random numbers
or
MIDI Music

Data 2

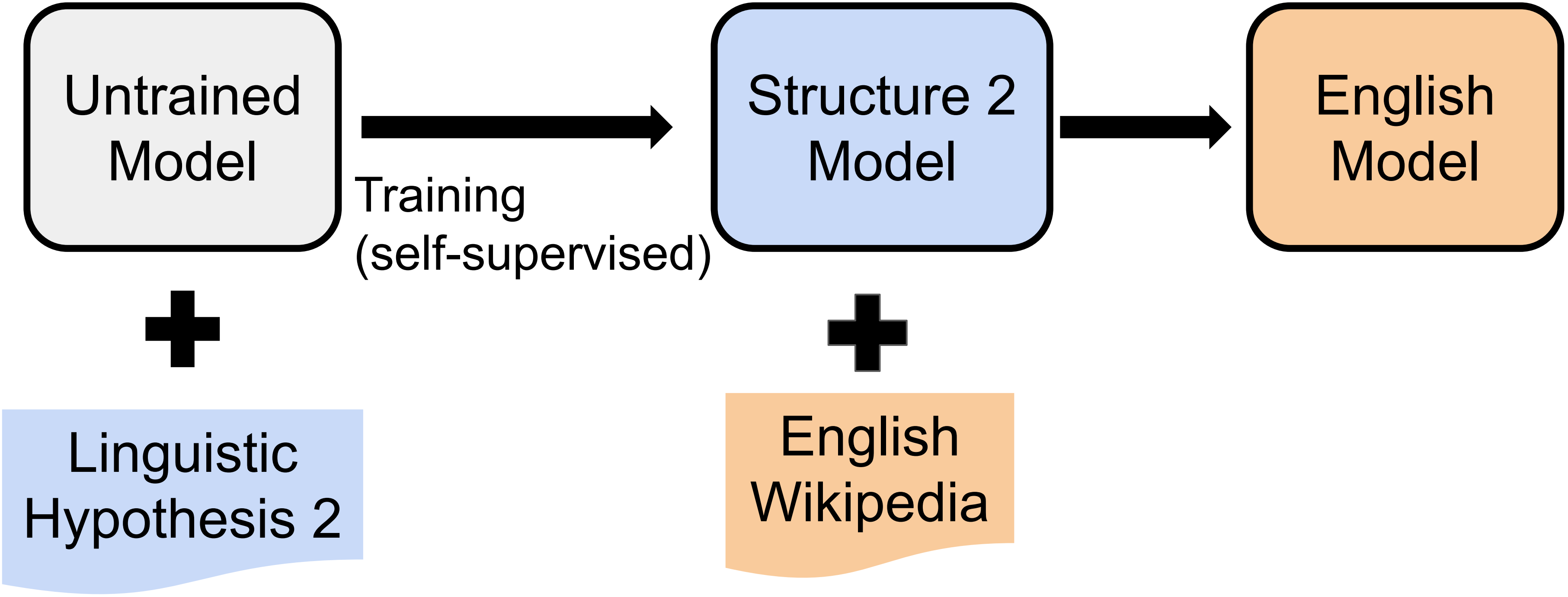
English wikipedia (~100 mil token subset)



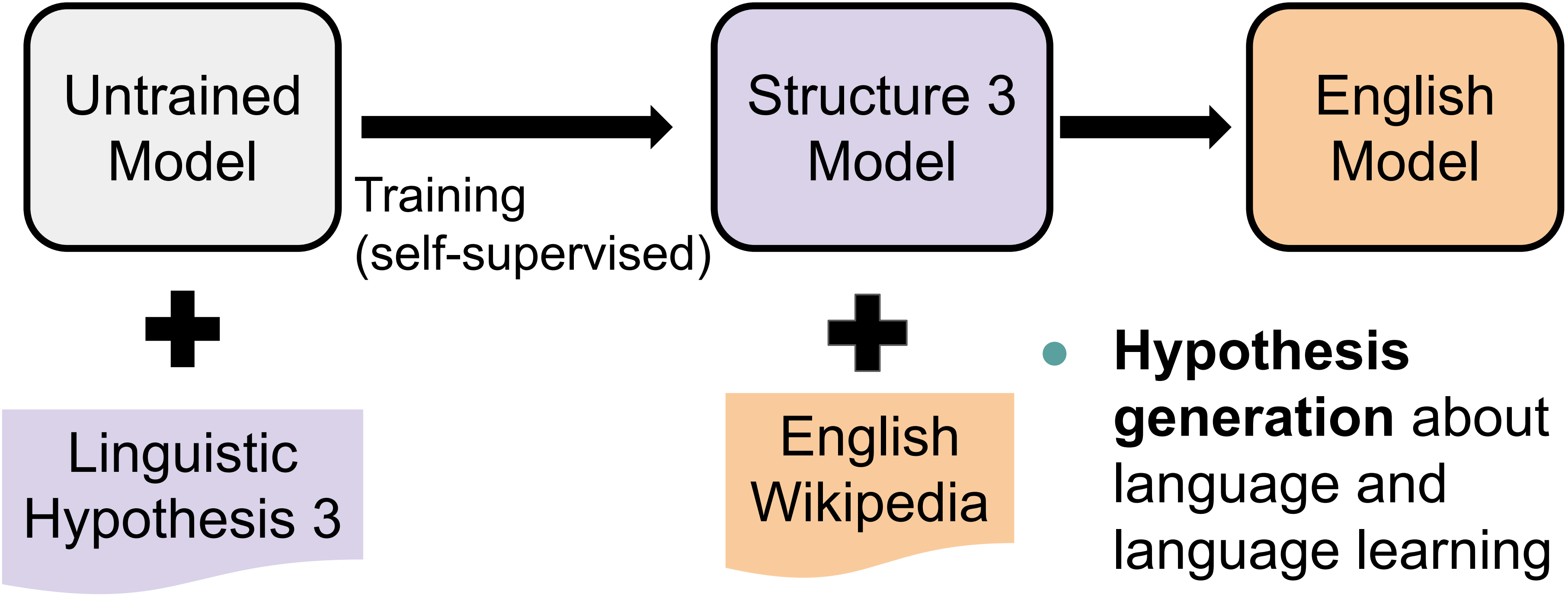
Structural Transfer: a testbed for linguistic structure hypotheses



Structural Transfer: a testbed for linguistic structure hypotheses



Structural Transfer: a testbed for linguistic structure hypotheses



- **Hypothesis generation** about language and language learning

Using structural transfer learning to explore the **role of structure** in language and language learning

Transfer learning in NLP

- Recent NLP: pretrain so much, that the task can be described in language. **Prompting**

Transfer learning now

- Looking beyond the **dominant languages** where we can do things like prompting
- And for understanding **structure**

Structure and language

- Structure is characteristic of human language
- Most obviously in syntax
- But also beyond syntax
  - Meaning, discourse, reference, information structure
- What structural biases are sufficient for language learning?
- (beyond this talk) Role of communication and language use in creating structure

Three hypotheses about language

1) Recursion

Constituents “Clumping”

The cat sat on the mat

I think that the cat sat on the mat

You always accuse me that I think that the cat sat on the mat

Three hypotheses about language

1) Recursion

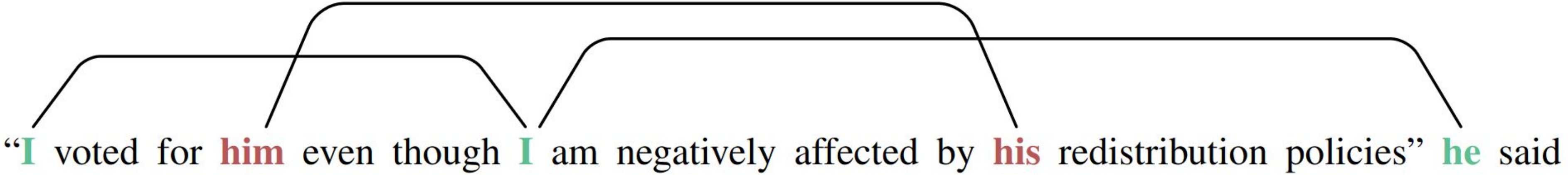
Nesting   Context-free

The lawyer that the man that the dog bit hired was disbarred

Three hypotheses about language

2) Crossing links and dependencies

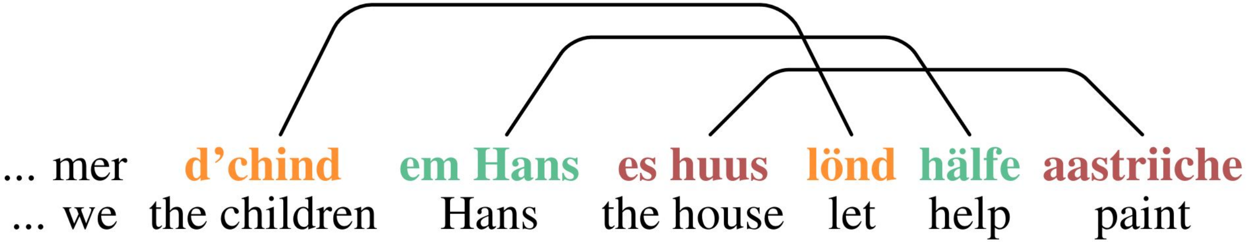
Linking in meaning and reference



Three hypotheses about language

2) Crossing links and dependencies

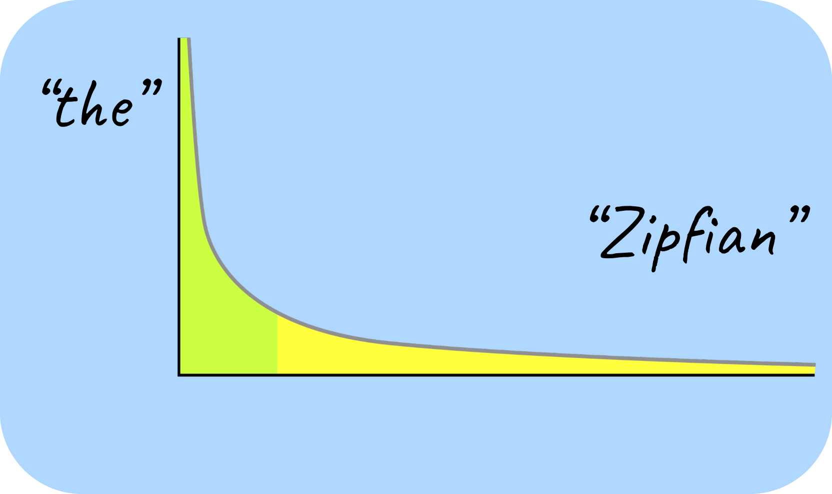
And syntactic structures



[Schieber 1985]

Three hypotheses about language

3) Zipfian vocabulary distribution



Outline

- What structural biases are useful for human language learners?
  - Disentangling the effects of **recursive** and **linking** structures
- How does vocabulary distribution transfer as a structural bias?
  - The structural effect of vocabulary

Outline

- What structural biases are useful for human language learners?
  - Disentangling the effects of **recursive** and **linking** structures

- How does vocabulary distribution transfer as a structural bias?
  - The structural effect of vocabulary

Exploring inductive bias

? + Pretty limited linguistic exposure → Language learning

Use transfer learning to test different **structural inductive learning biases**

Transfer learning methodology

Untrained GPT-2 model

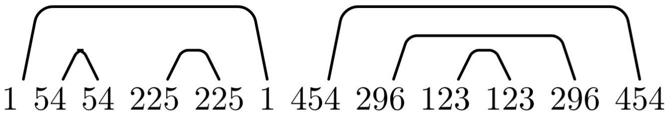
➔

Structurally-biased learner

Vocabulary

✚

Symbolic Language



✚

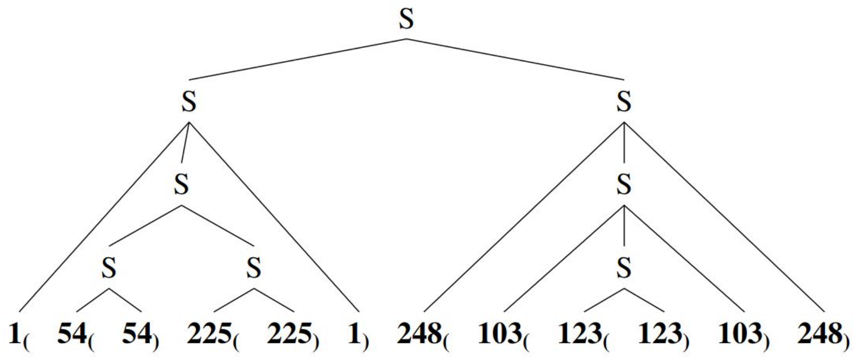
WIKIPEDIA
The Free Encyclopedia

How fast/well can it learn language?

Symbolic pretraining languages

Nesting Parentheses

Nested Parentheses Primitive

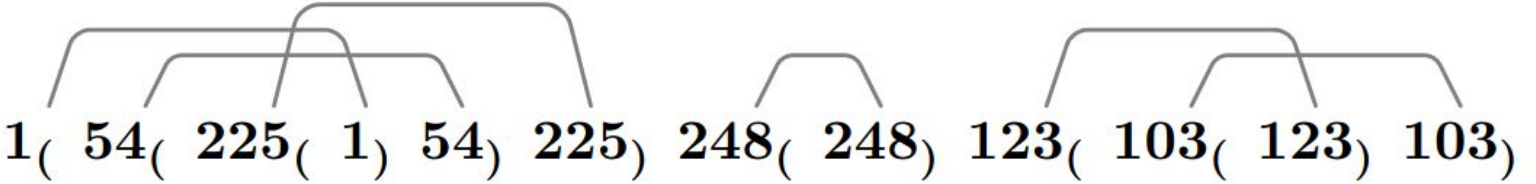


- Well-nested, matching pairs
- Constituents

Symbolic pretraining languages

Nesting Parentheses	Crossing Dependencies

Crossing Dependencies



- Tokens have to **match**, but not **nest**
- Where does the structure come from?
  - **Dependency length distribution:** sample from empirical distances of nesting parentheses

Symbolic pretraining languages

	Nesting Parentheses	Crossing Dependencies
Controls:	Random	Regular repetition

Simple Repetition Primitive

Randomly sample *k* words, then repeat them, then randomly sample *k* words…

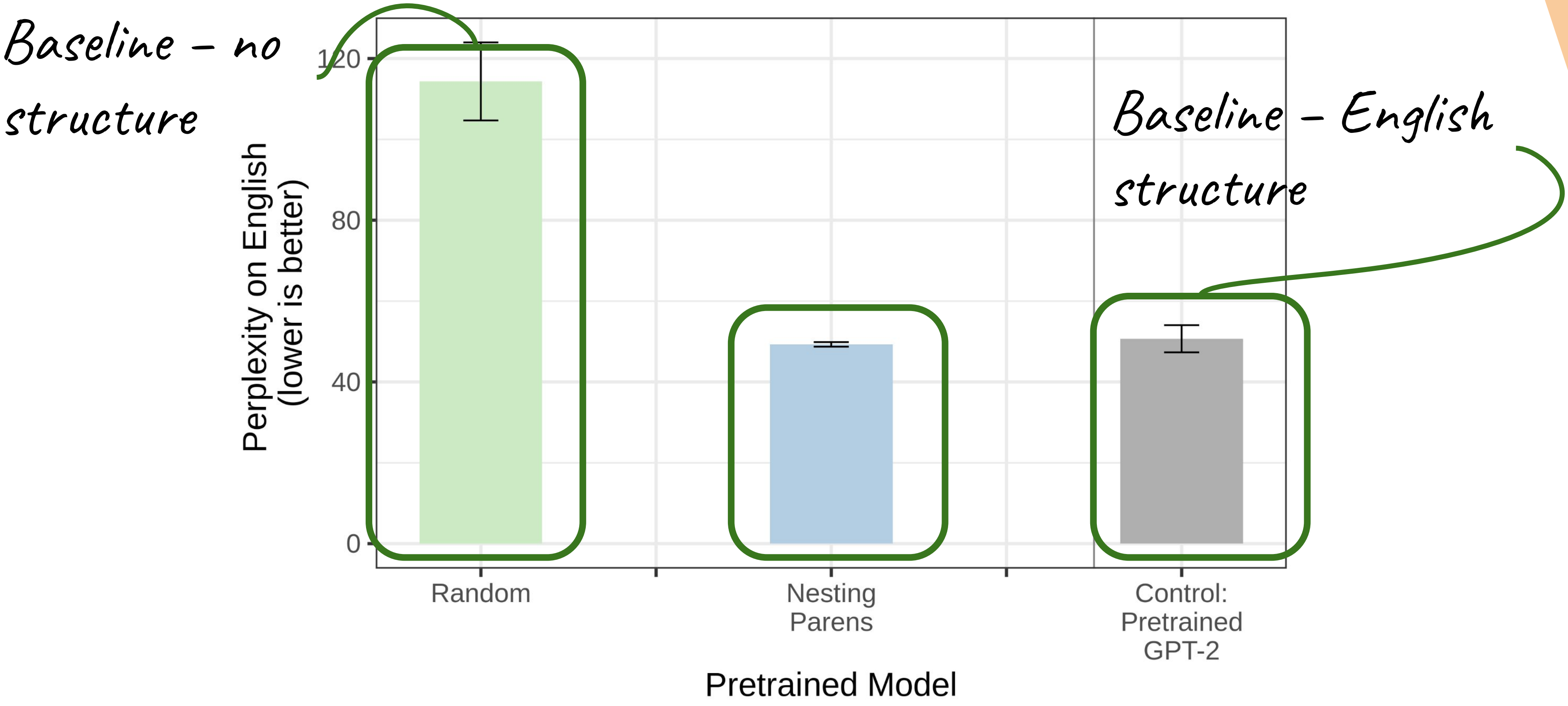
499 472 300 345 272 $\boxed{\textbf{499 472 300 345 272}}$ **309 17 15**

(Example is for k=5, we do k=10 in experiments)

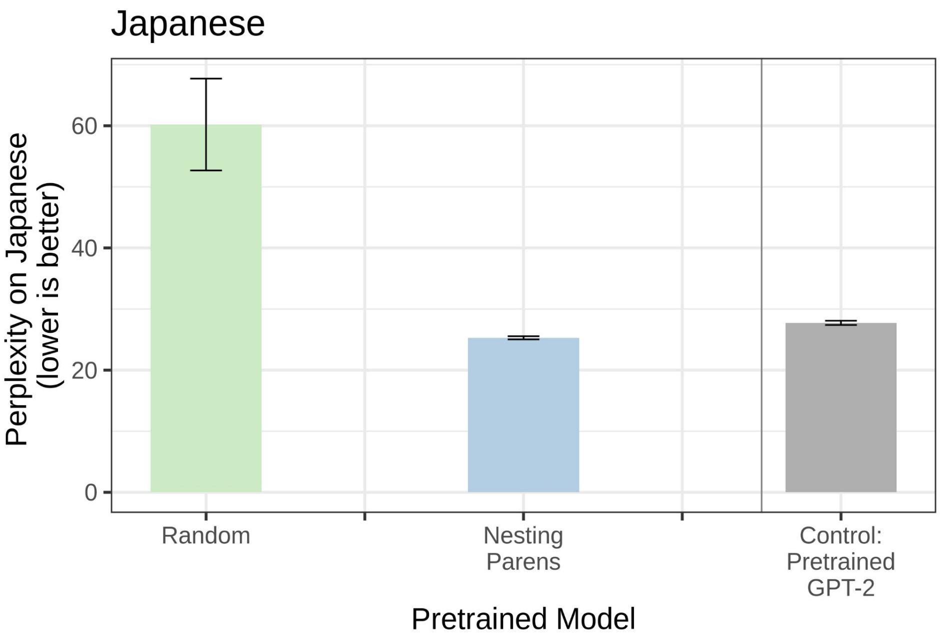
Symbolic pretraining languages

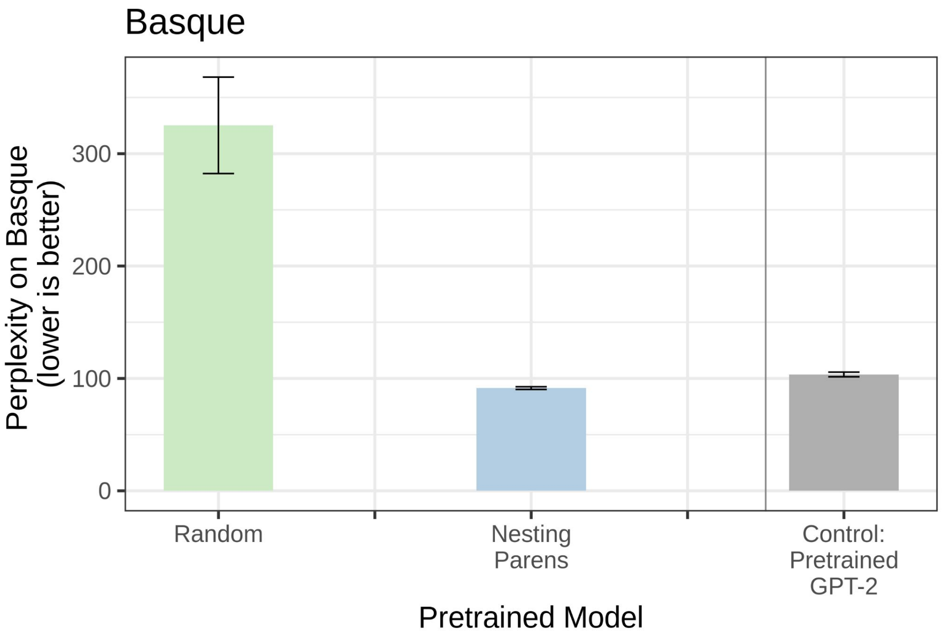
	Nesting Parentheses	Crossing Dependencies
Controls:	Random	Regular repetition

Nesting structure helps language learning



Multilingual case – Japanese and Basque





Symbolic pretraining languages

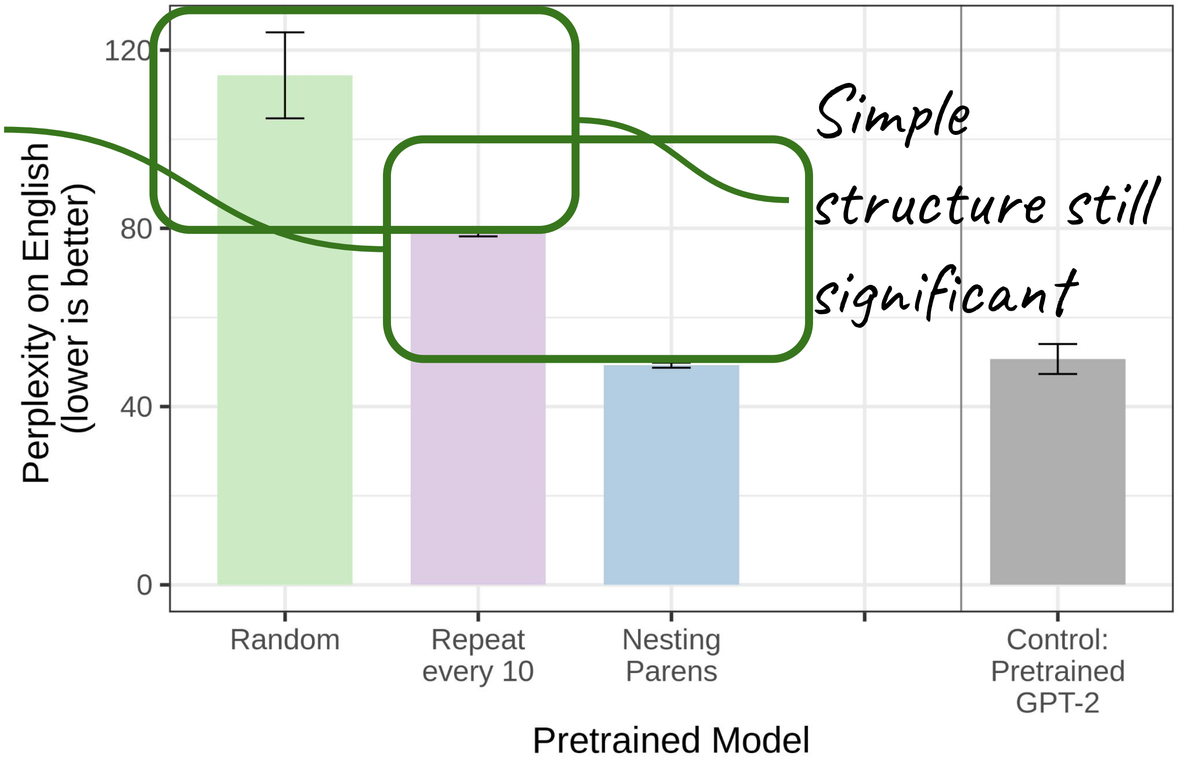
	Nesting Parentheses	Crossing Dependencies
Controls:	Random	**Regular repetition**

Question: does **nesting** really help? Or would any structure help?

499 472 300 345 272 $\boxed{499\ 472\ 300\ 345\ 272}$ 309 17 15

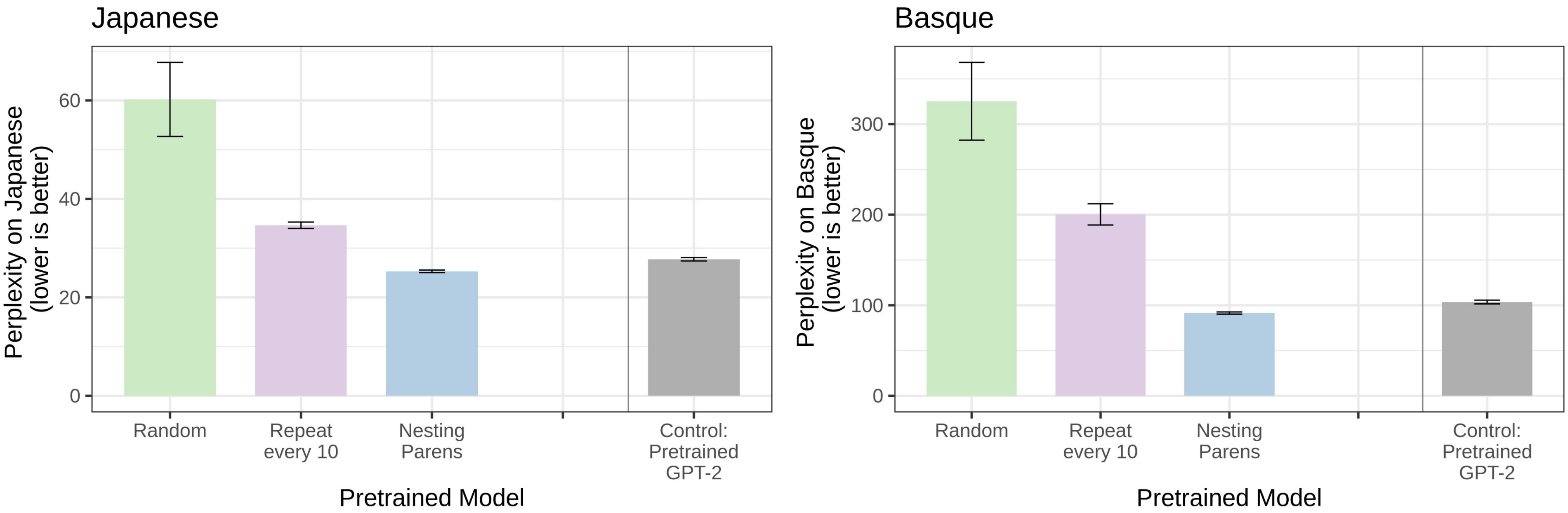
Not just any structure has this effect

Complexity of nesting is necessary



Simple structure still significant

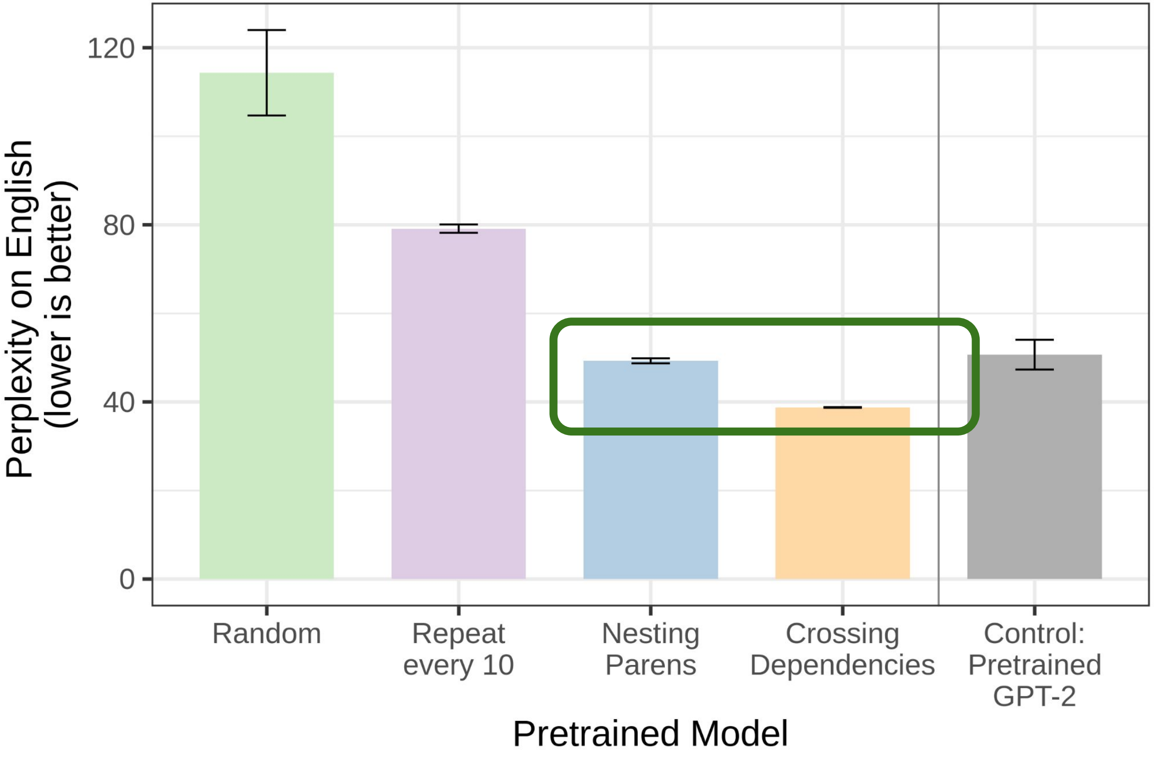
Again, a multilingual effect



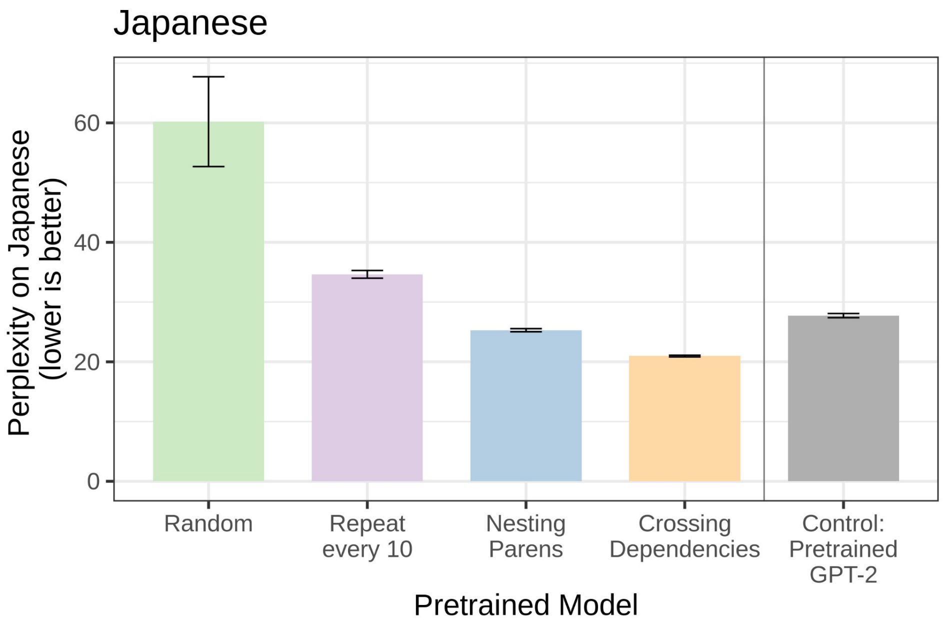
Symbolic pretraining languages

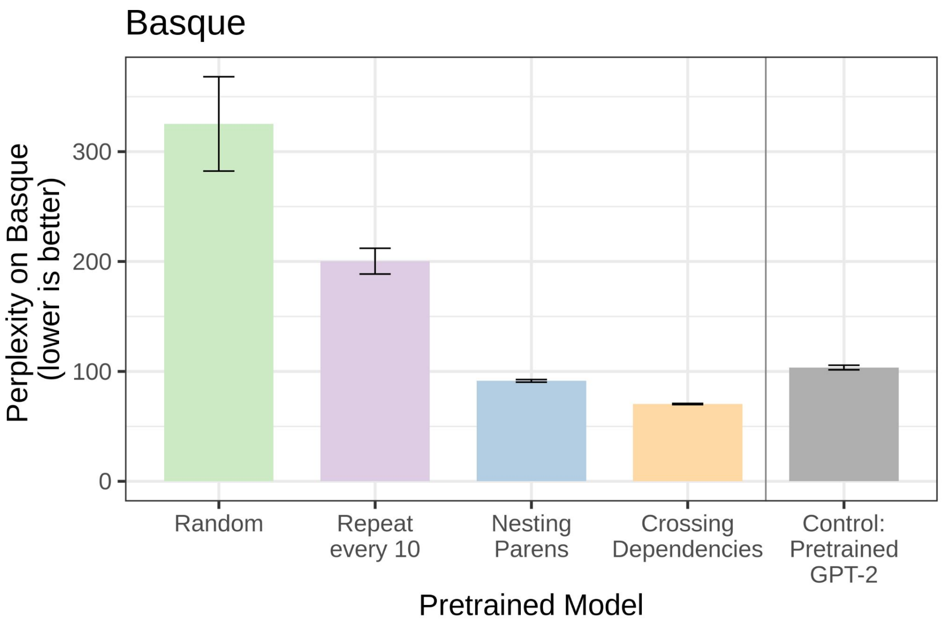
	Nesting Parentheses	Crossing Dependencies
Controls:	Random	Regular repetition

Crossing links, without nesting, provide a better inductive bias



This is also true across languages



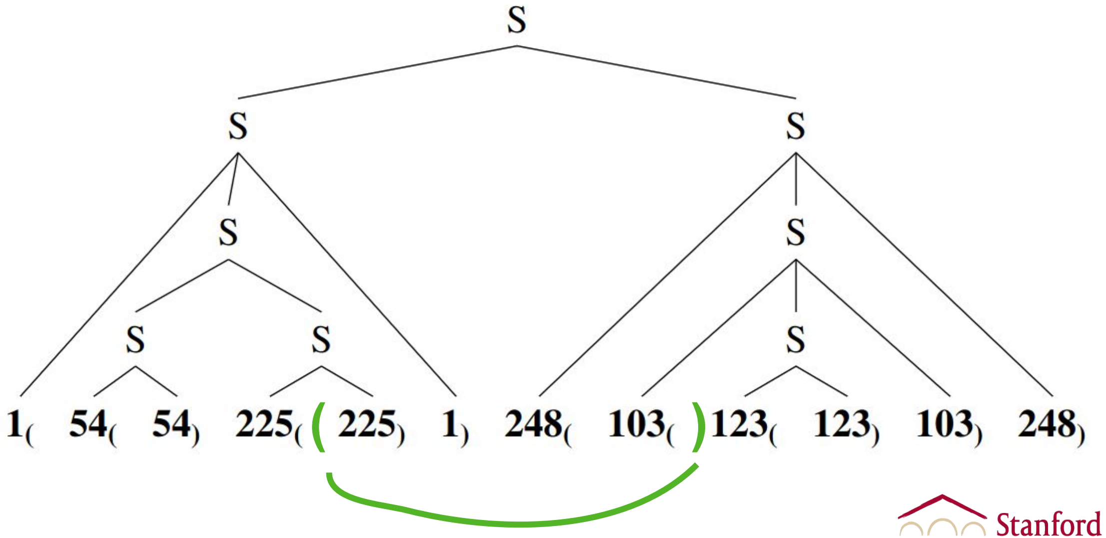


The kinds of structure that make language are multifaceted

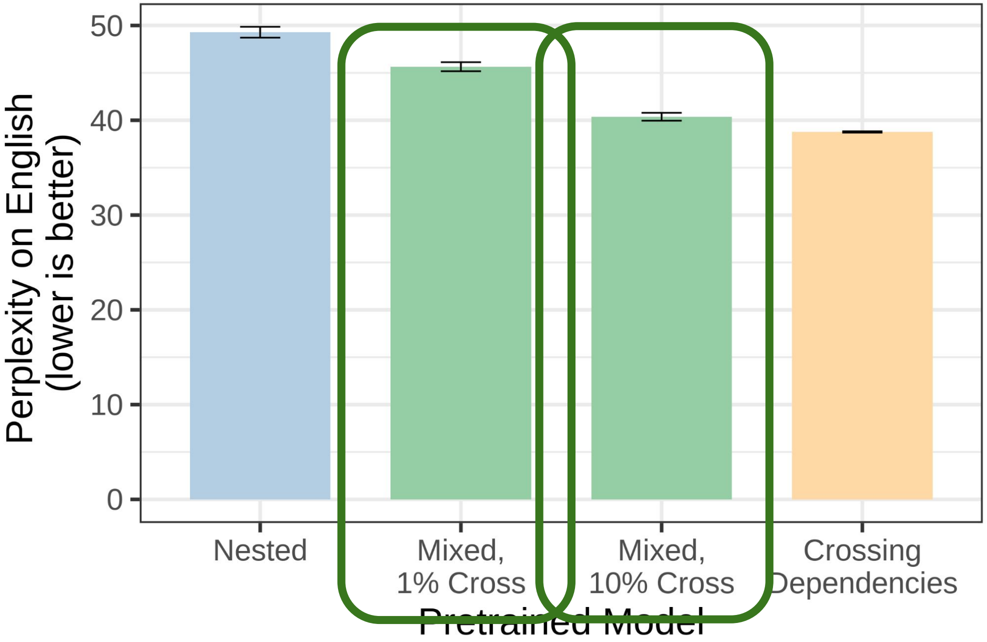
- Structural transfer lets us explore hypotheses about structure in language
- Language as a learnable system, independent of linguistic theory

Mixing nesting and crossing parentheses

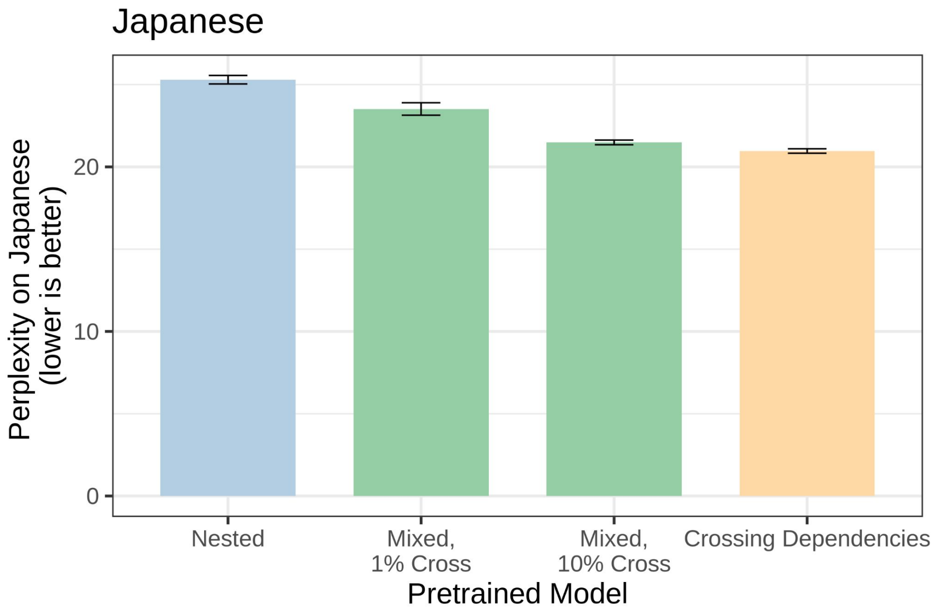
- A language that is mostly nesting, with 1%, or 10% of parentheses not following the structure

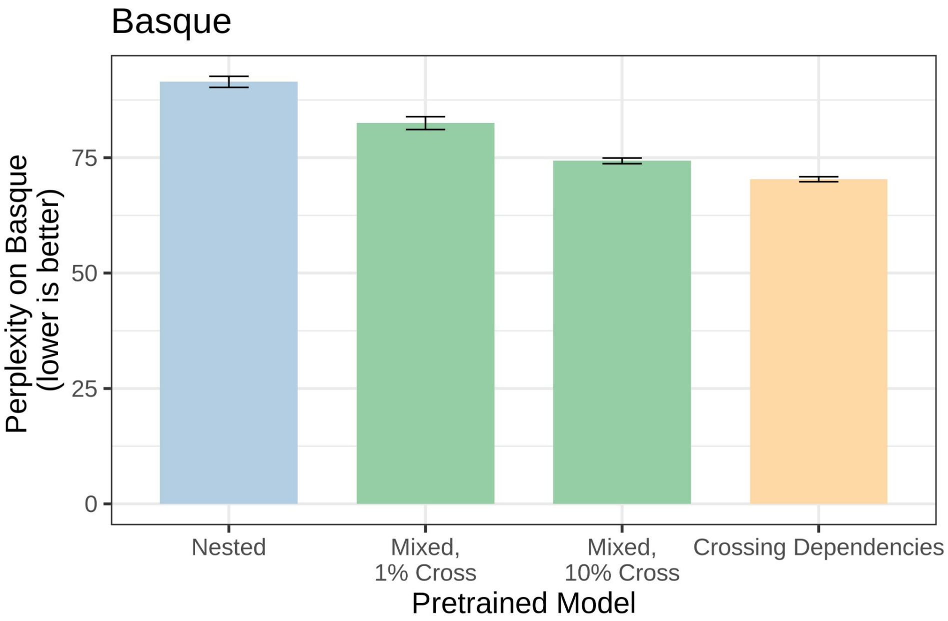


Slightly breaking constituent structure makes better language learners



Also a multilingual effect





Structural inductive bias through transfer learning

- Complex structural relationships are important in language
- Multiple crossing dependencies a good starting point for language learning
- Computational models as **hypothesis generators**: testing linguistic structure in theory-free ways

Outline

- What structural biases are useful for human language learners?
  - Disentangling the effects of **recursive** and **linking** structures
- How does vocabulary distribution transfer as a structural bias?
  - The structural effect of vocabulary

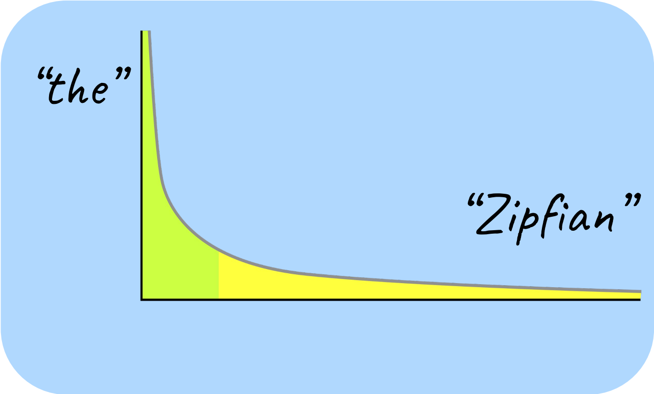
Outline

- What structural biases are useful for human language learners?
    - Disentangling the effects of **recursive** and **linking** structures

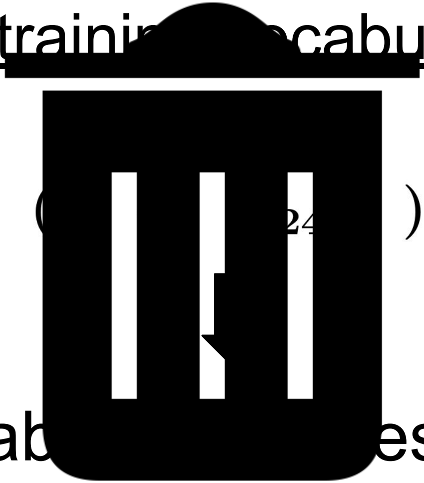
- How does vocabulary distribution transfer as a structural bias?
    - The structural effect of vocabulary

The lexicon in linguistics

- A good amount of structure is in the vocabulary:
- Vocabulary distribution
- Structure in meaning
- and also in grammar
  - Properties like transitive verb
  - Constructions, like “Let alone”
  - …



We throw out the vocabulary between pretraining and fine-tuning



Pretraini ocabulary

$(_1 \quad (_2 \quad (\quad \ldots \quad)_{249} \quad)_{250}$

Vocab es 0-499

What the model sees

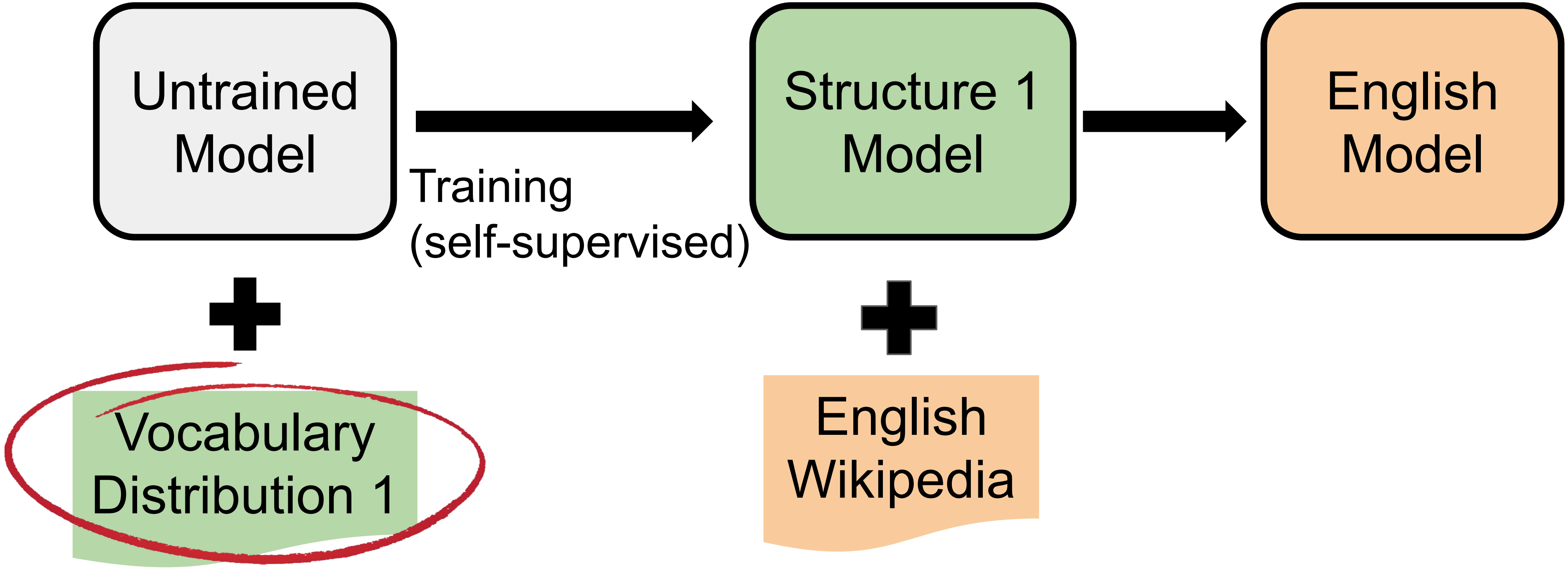
Fine-tuning vocabulary

cat, dog … book … ameliorate …

↓

Vocabulary indices 0-50K

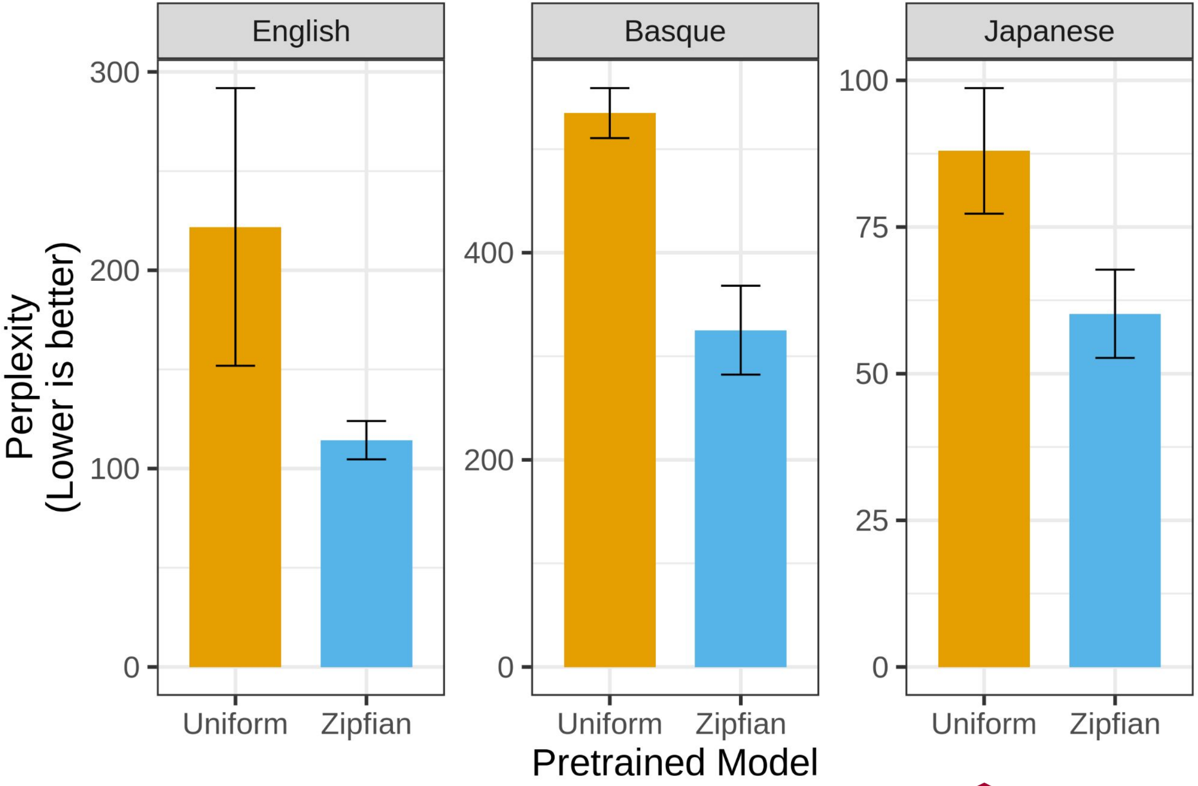
Structural Transfer: a testbed for linguistic structure hypotheses



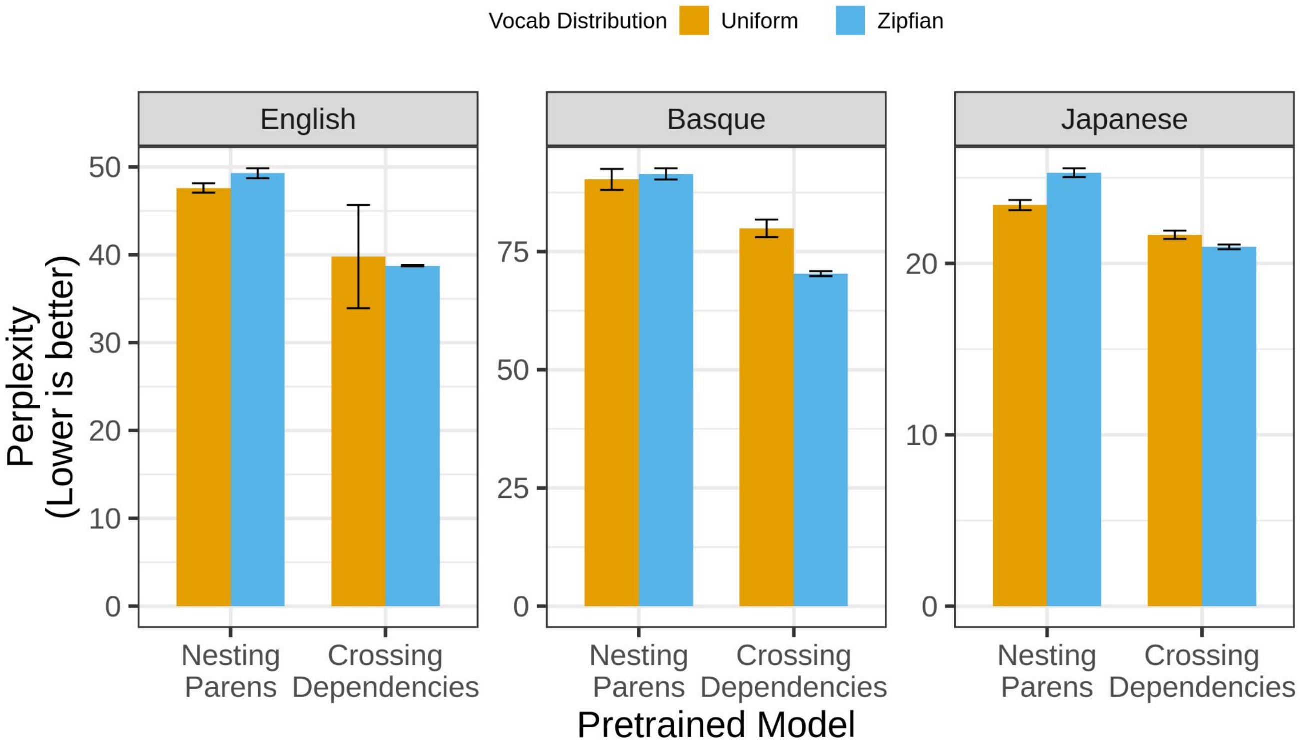
Does a Zipfian vocabulary distribution in pretraining have a **structural** effect?

- Even though we discard vocabulary information

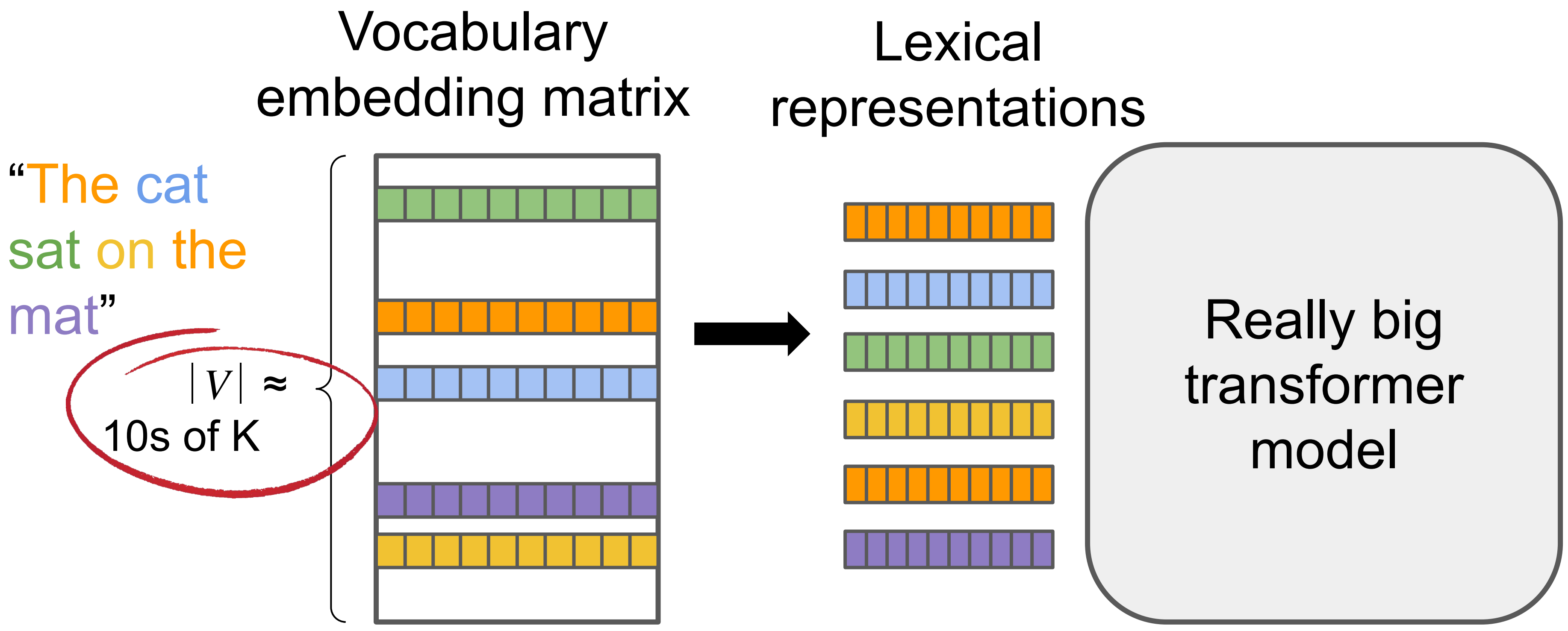
Yes, Zipfian information is transferred



… but does not necessarily combine with structure



The role of vocabulary in transfer learning is an interesting problem

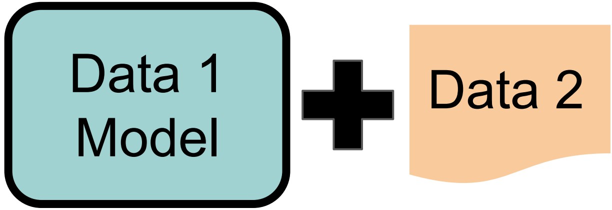


The role of vocabulary in transfer learning is an interesting problem

- A practical problem: without enough data, it’s hard to see a word often enough to learn a good vector
- A puzzle: how is structural information separated between vocabulary matrix and model weights?
  - Vocabulary information like distribution can have structural effects

Transfer learning, language, and structure

- Transfer learning is a test bed for understanding structure in language learning



- Computational models of cognitive processes can't prove anything – but they serve as interesting hypotheses generators

- It's an exciting time: machine learning opens up new avenues for exploring questions in language

Thanks!