What can language models teach us about structure in human language?

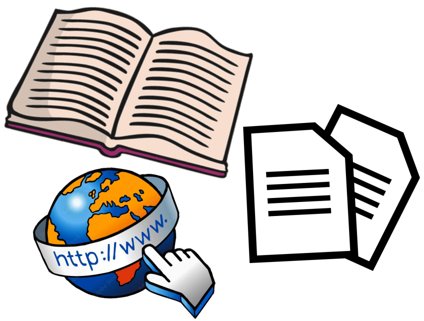
Kempner Institute
For the Study of Natural & Artificial Intelligence

Harvard University

Isabel Papadimitriou

An exciting empirical development

Language data



Imitation training

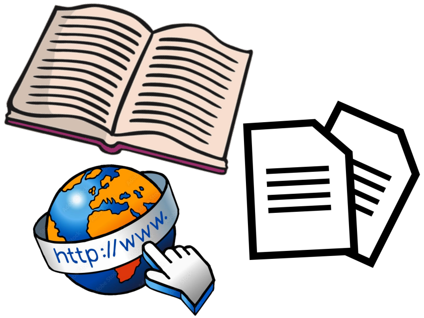
Neural model

Linguistic capabilities

- Amazing, powerful, versatile
- Governed by intricate systems

An exciting empirical development

Language data



Imitation training

Neural model

Linguistic capabilities

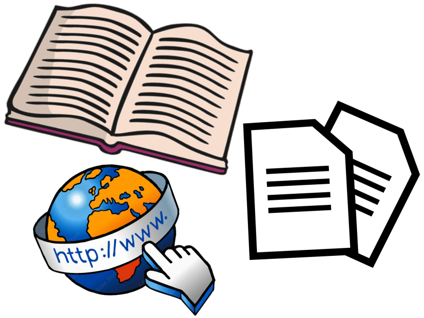
- Self-supervised learning – we don't know what we are going to get
- Model creates language system

Complex linguistic capabilities

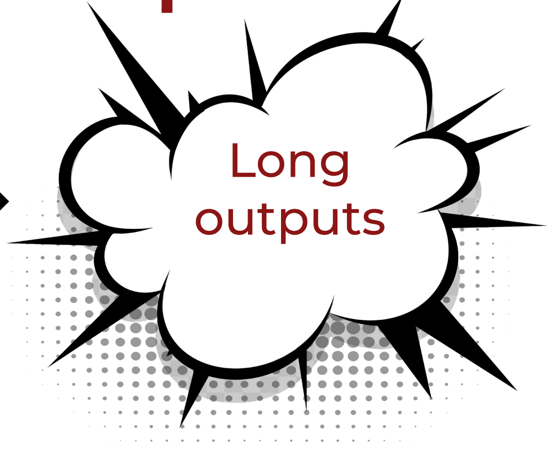
Language data

Imitation training

Linguistic capabilities

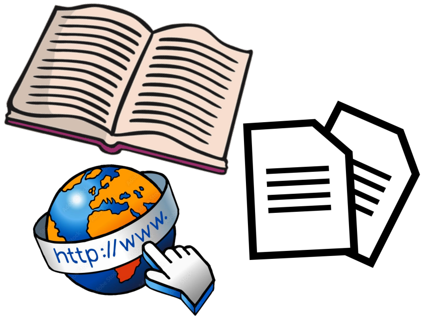


Neural model



Complex linguistic capabilities

Language data



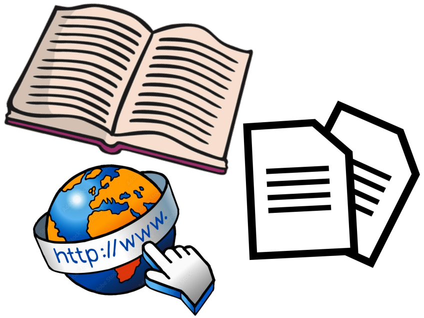
Imitation training

Neural model

Linguistic capabilities

Complex linguistic capabilities

Language data



Imitation training

Neural model

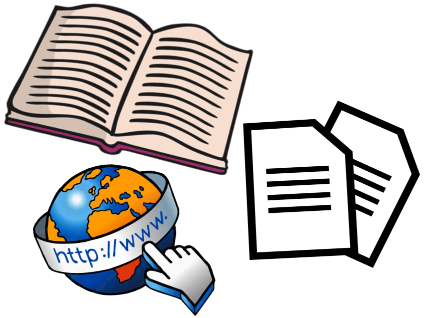
Linguistic capabilities

Complex linguistic capabilities

Language data

Imitation training

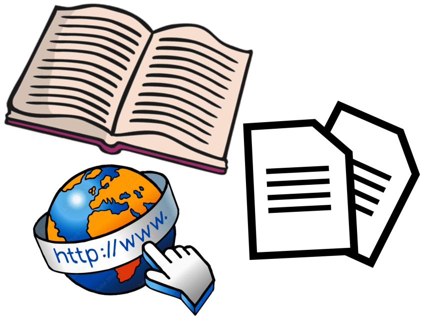
Linguistic capabilities



Neural model

Complex linguistic capabilities

Language data



Imitation training

Neural model

Linguistic capabilities

Complex linguistic capabilities

Language data

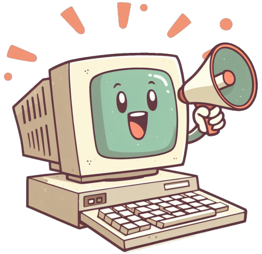
Imitation training

Linguistic capabilities

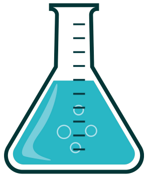
An exciting phenomenon for linguists to explore!

Language models and human language

Can language models be a tool for linguistics?

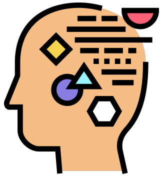


Two roles for LMs in linguistics:



Empirical testbeds

- **Intervention experiments** on language learning and production



Functional theoretical examples

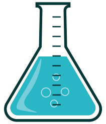
- A working example of how a linguistic process ***could*** be represented

To expand and enrich our hypothesis space

Three methodologies using LMs to explore questions of language structure:

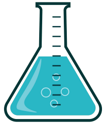
1) **Structural injection**:

   testing different linguistic learning biases



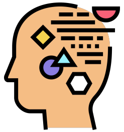
2) **Impossible language learning:**

   what do LMs learn more easily?



3) **Subjecthood in LMs:**

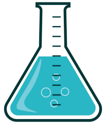
   how are grammatical roles organized in latent space?



Three methodologies using LMs to explore questions of language structure:

1) **Structural injection**:

   testing different linguistic learning biases



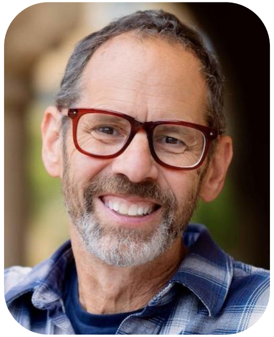
2) **Impossible language learning:**

   what do LMs learn more easily?

3) **Subjecthood in LMs:**

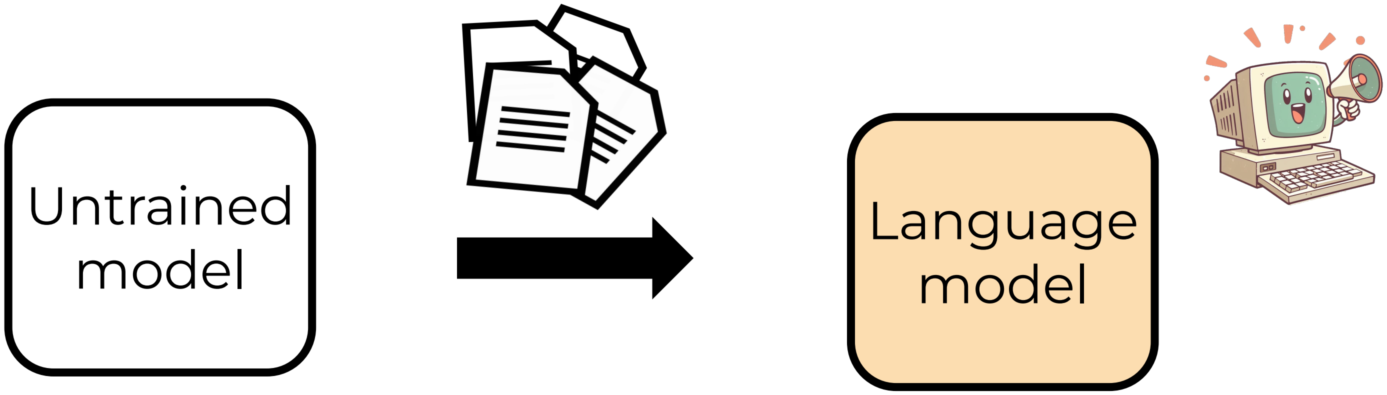
   how are grammatical roles organized in latent space?

Collaborator



Dan Jurafsky

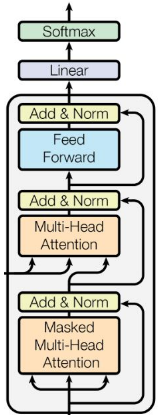
Question: What inductive biases make learning language easier?

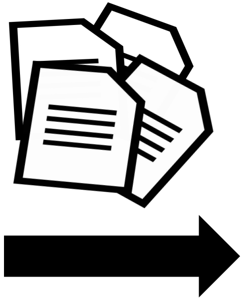


- Every learner has **inductive biases**
- In humans: big question
  In transformers: we also don't know

Question: What inductive biases make learning language easier?

(not a blank slate)

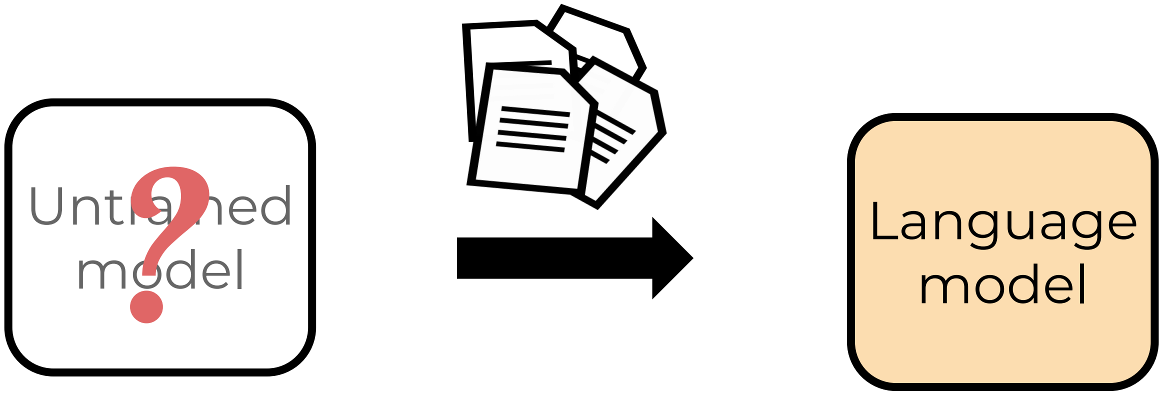




Language model

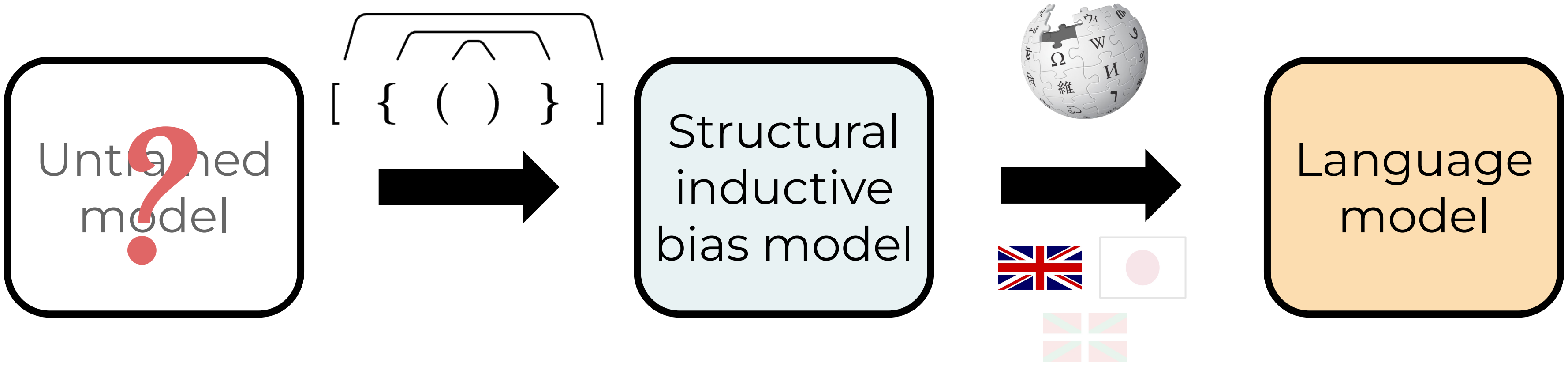
- Every learner has **inductive biases**
- In humans: big question
  In transformers: we also don't know

Question: What inductive biases make learning language easier?



- Every learner has **inductive biases**
- In humans: big question
  In transformers: we also don't know

Structural injection: Controlling inductive bias through **training**



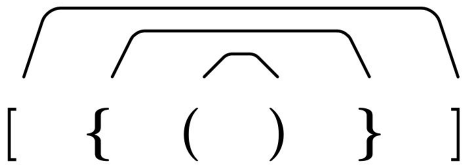
Method: 1) Train on formal language

2) Train on natural language

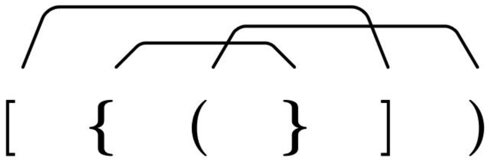
Testing different inductive biases

(and we could test more!)

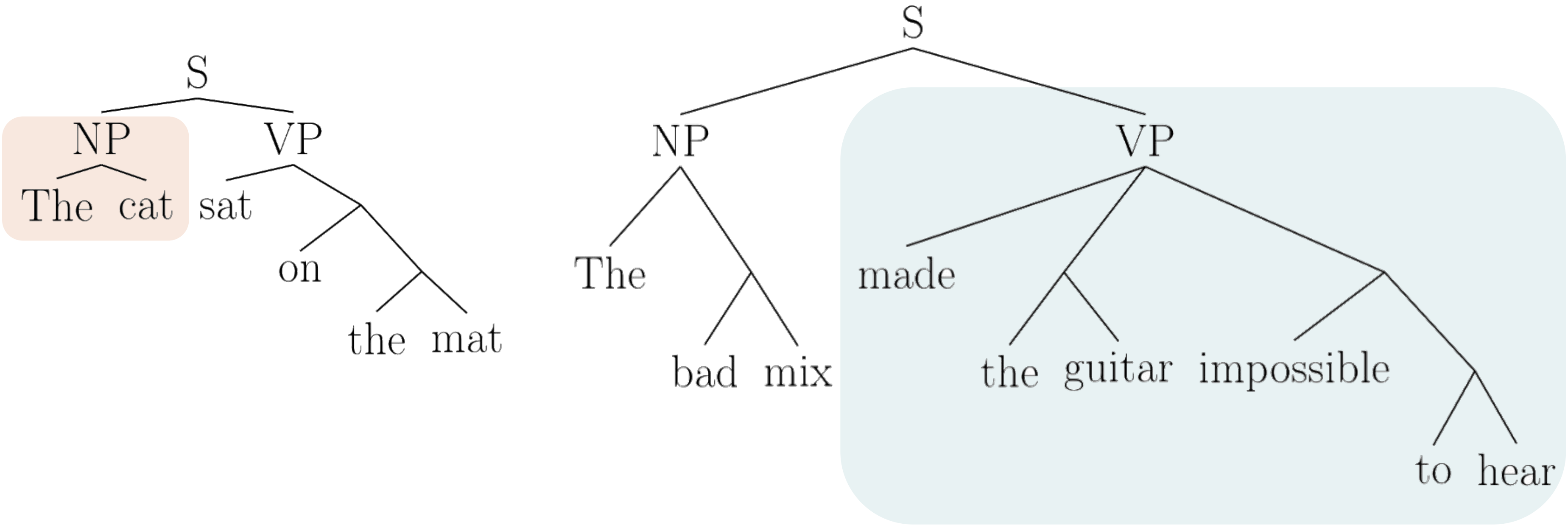
- **Recursion, nesting parentheses** (context-free)



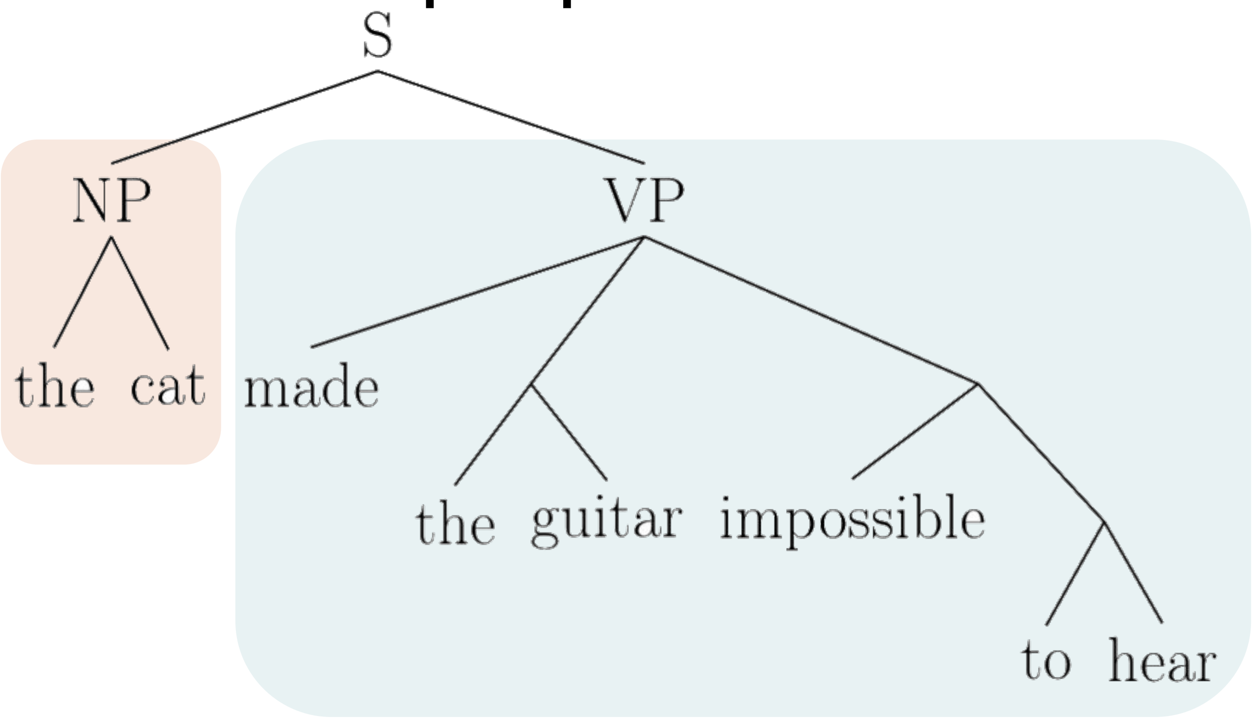
- **Crossing dependencies** (non-context-free)



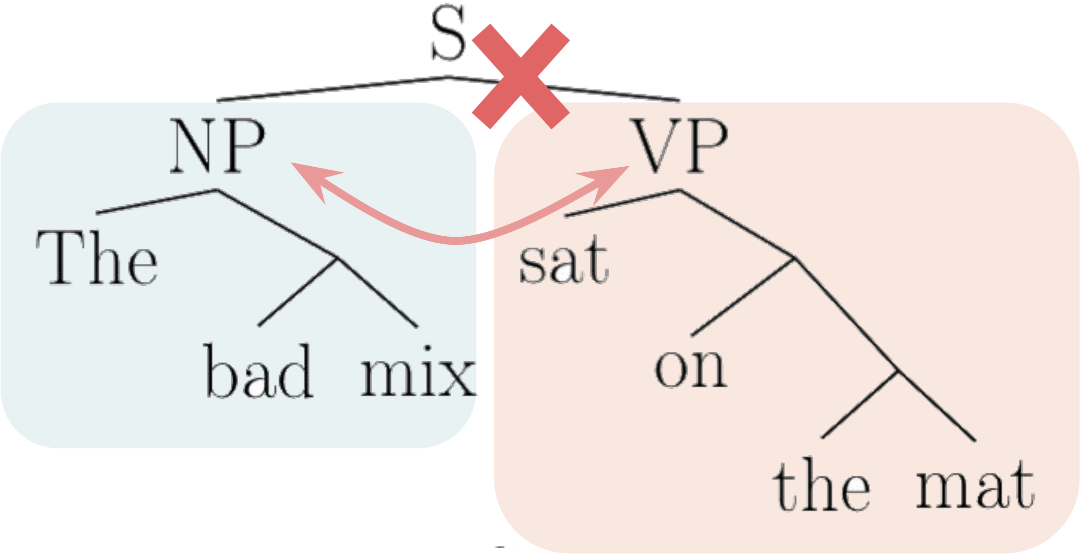
Recursion: Language is nesting, tree-structured



Language is tree structured, context-free properties

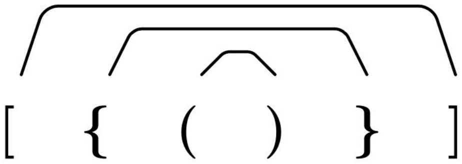


Language is also full of complex, crossing dependencies

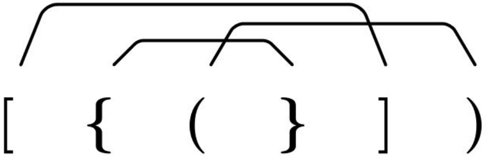


Testing different inductive biases

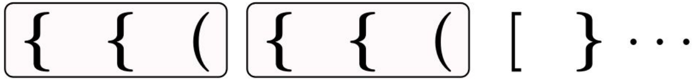
- **Recursion, nesting parentheses** (context-free)



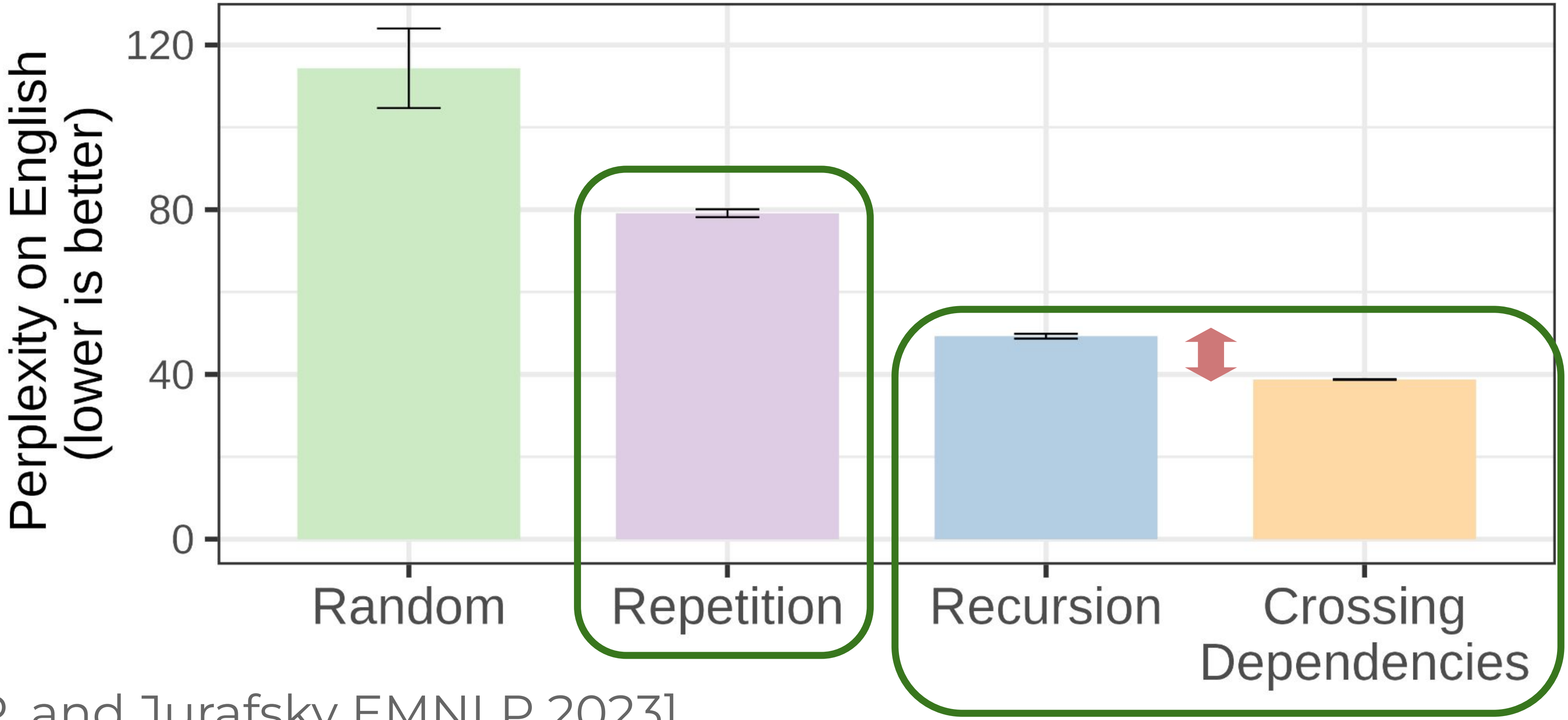
- **Crossing dependencies** (non-context-free)



- Baseline: **Repetition** (regular)

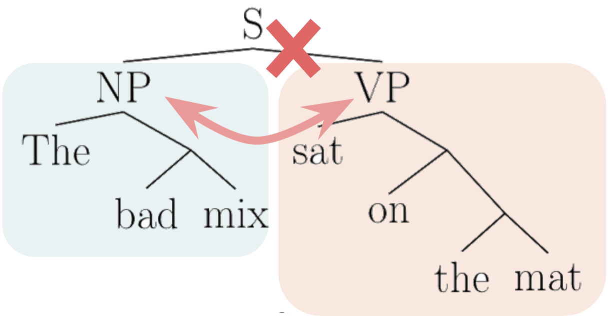


Formal inductive biases for language



[P. and Jurafsky EMNLP 2023]

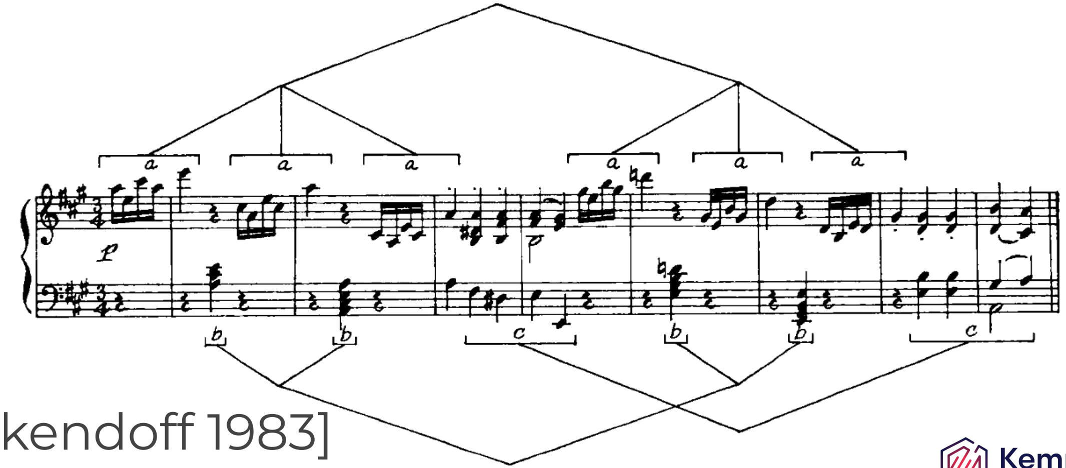
Crossing dependencies as a language primitive



- LM experiments bring to light the possible importance of these structures for cognitively scaffolding language learning

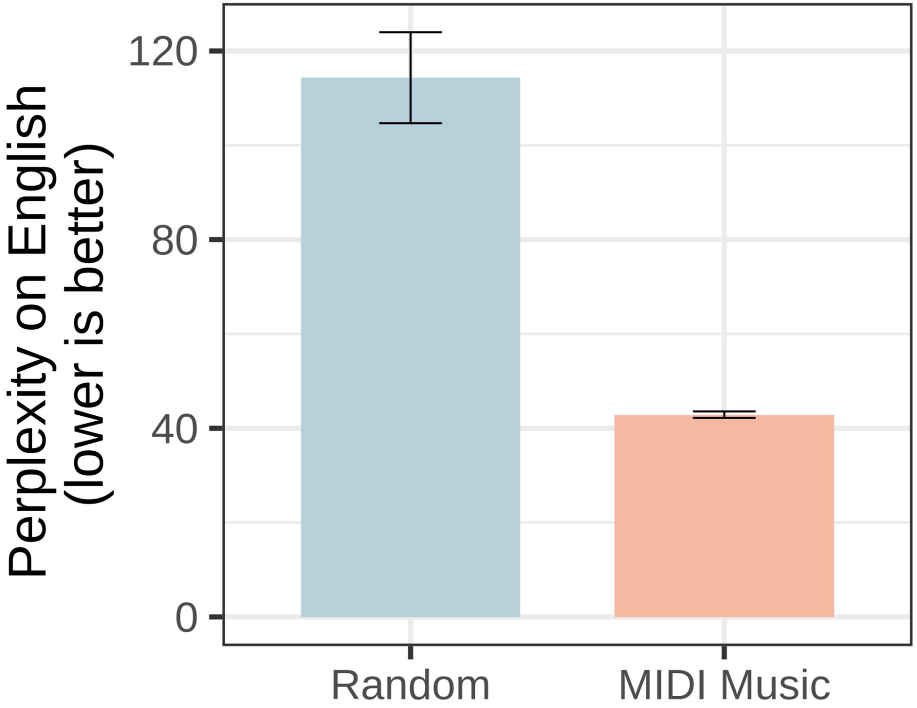
Coda: where do inductive biases come from?

- In humans: perhaps innate, perhaps joint learning, perhaps related to **other cognitive processing**
- Can other cognitive domains act as inductive bias?



[Lerdahl and Jackendoff 1983]

Abstractions transfer from music to language



[Papadimitriou and Jurafsky EMNLP 2020]

💡 Language models are hypothesis generators:

- for the cognitive representation of language
- for language inductive biases

Three methodologies using LMs to explore questions of language structure:

1) **Structural injection**:
   testing different linguistic learning biases

2) **Impossible language learning:**
   what do LMs learn more easily?

3) **Subjecthood in LMs:**
   how are grammatical roles organized in latent space?

Collaborators

Julie Kallini

Kyle Mahowald

Richard Futrell

Chris Potts

Can language models learn impossible languages?

- Main idea: train models on corpora that we have altered to have **unattested characteristics**
- Can statistical language models learn altered languages as well as English? **– No!**
- Why not?

For example: LMs disprefer counting rules

Counting-based inflection: *Hop languages

NoHop language

singular/plural marker

`He clean S his very messy bookshelf`

TokenHop language

`He clean his very messy books S he lf`

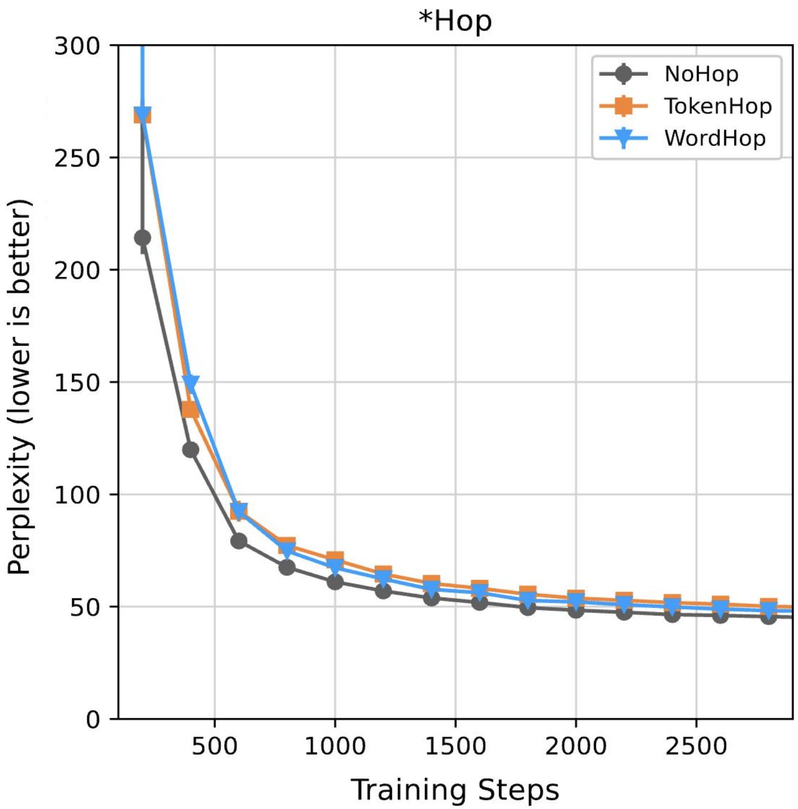
4 tokens

WordHop language

`He clean his very messy bookshelf S`

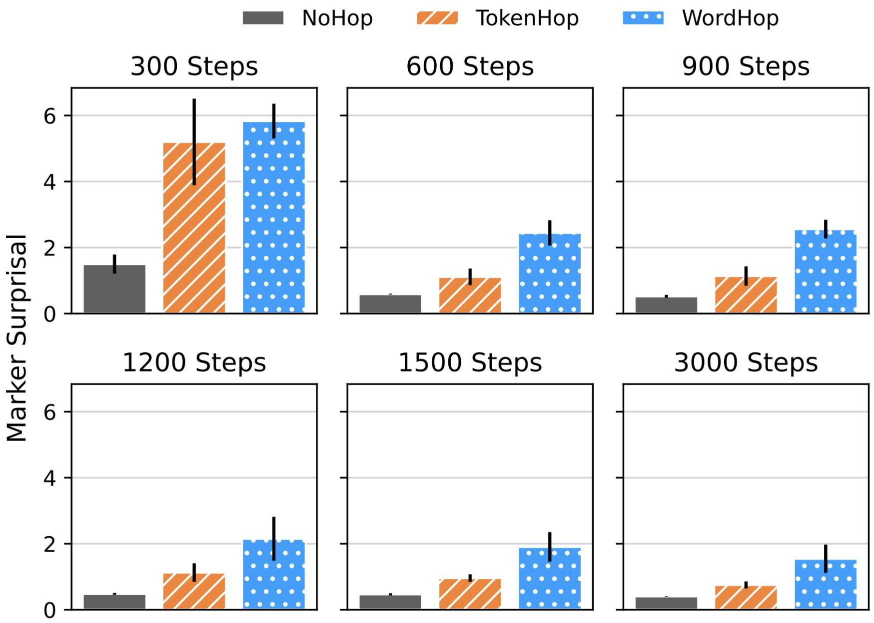
4 words

Worse *overall perplexity* for hop languages



Consistently higher surprisal of marker

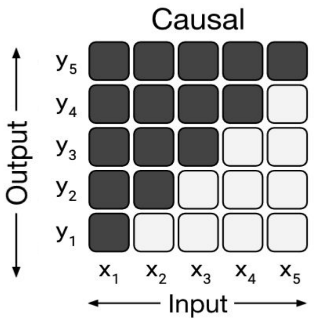
$S(\boxed{\text{S}})$



Why do we see these human-like effects with impossible languages?

- Transformers don't share all human learning capabilities/biases
- But some things are the same – eg, they can't see the future
- Experiments let us explore **where aspects of human language arise from**

Attention masking



Why do we see these human-like effects with impossible languages?

- One of the main takeaways from these experiments:

  Likely importance of **information locality** in defining features of language

Three methodologies using LMs to explore questions of language structure:

1) **Structural injection**:

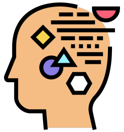
   testing different linguistic learning biases

2) **Impossible language learning:**

   what do LMs learn more easily?

3) **Subjecthood in LMs:**

   how are grammatical roles organized in latent space?



Collaborators

Kyle Mahowald

Richard Futrell

Ethan A. Chi

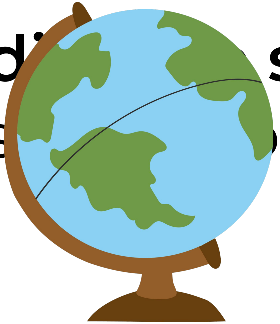
Eli Pugh

Ishan Shah

Subjecthood in language models

1) A **multilingual concept** for probing representations

2) Determined by both **d** … **syntactic** rules, as well as complicated, inters … **rototype features**



subject

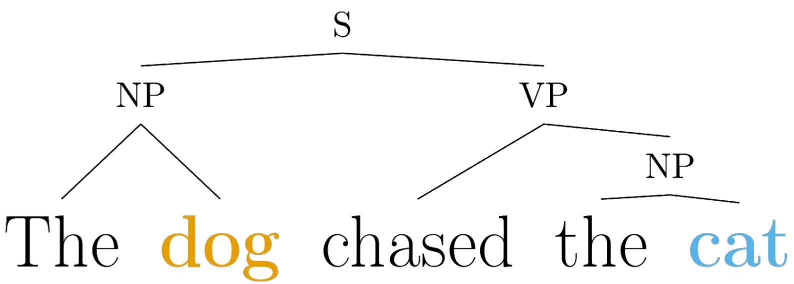
object

The dog chased the cat

Subjecthood in language models

1) A **multilingual concept** for probing representations

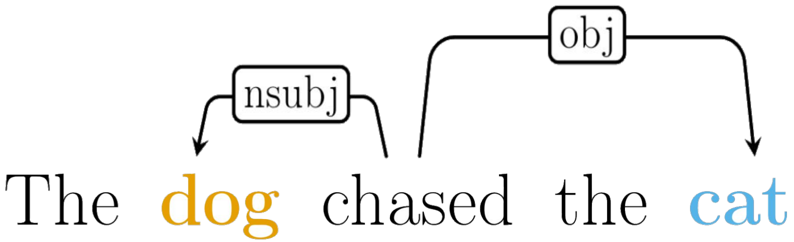
2) Determined by both **discrete syntactic** rules, as well as complicated, intersecting **prototype features**



Subjecthood in language models

1) A **multilingual concept** for probing representations

2) Determined by both **discrete syntactic** rules, as well as complicated, intersecting **prototype features**



Subjecthood in language models

1) A **multilingual concept** for probing representations

2) Determined by both **discrete syntactic** rules, as well as complicated, intersecting **prototype features**

In English,
word order
(stay tuned)

The dog chased the cat

⟷

The cat chased the dog

Subjecthood in language models

1) A **multilingual concept** for probing representations

2) Determined by both **discrete syntactic** rules, as well as complicated, intersecting **prototype features**

Intransitives

The glass broke

Subjecthood in language models

1) A **multilingual concept** for probing representations

2) Determined by both **discrete syntactic** rules, as well as complicated, intersecting **prototype features**

Intransitives, **passive voice**

The perch was jumped onto by the cat

Subjecthood in language models

1) A **multilingual concept** for probing representations
2) Determined by both **discrete syntactic** rules, as well as complicated, intersecting **prototype features**

Intransitives, passive voice, **animacy**

The onion made the man cry

Subjecthood in language models

1) A **multilingual concept** for probing representations

2) Determined by both **discrete syntactic** rules, as well as complicated, intersecting **prototype features**

Intransitives, passive voice, animacy, **volitionality**

Mary punched/liked/**forgot** Sam

Subjecthood in language models

1) A **multilingual concept** for probing representations

2) Determined by both **discrete syntactic** rules, as well as complicated, intersecting **prototype features**

Intransitives, passive voice, animacy, volitionality,

case marking, information structure, … …

Subjecthood in language models

1) A **multilingual concept** for probing representations

2) Determined by both **discrete syntactic** rules, as well as complicated, intersecting **prototype features**

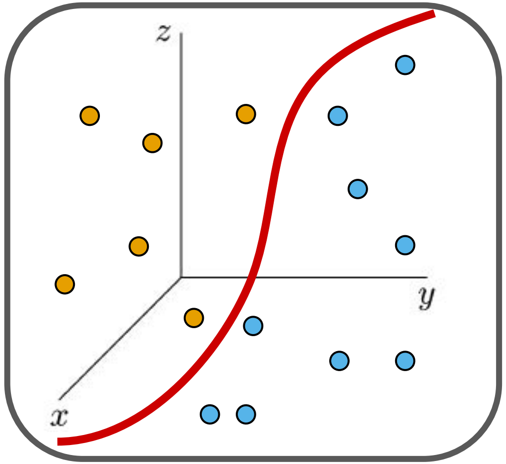
Let's see how language models represent it!

Method: find subjecthood representation

Annotated Corpus

LM latent space

The **dog** chases the **cat**

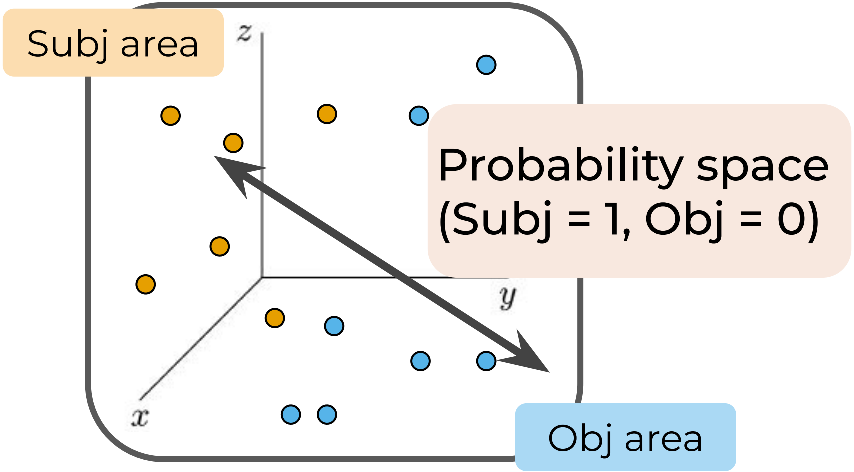


Train a classifier

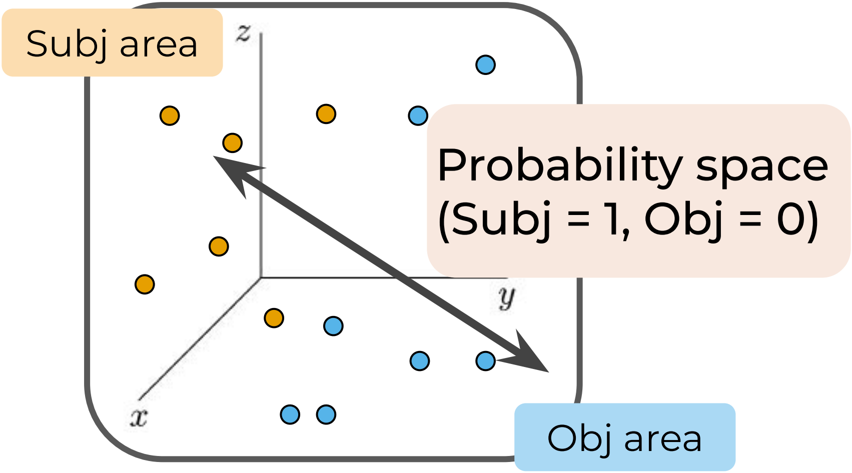
Method: find subjecthood representation

Annotated Corpus

The **dog** chases the **cat**



Two questions about subjecthood representation:



1) How does it work across languages?

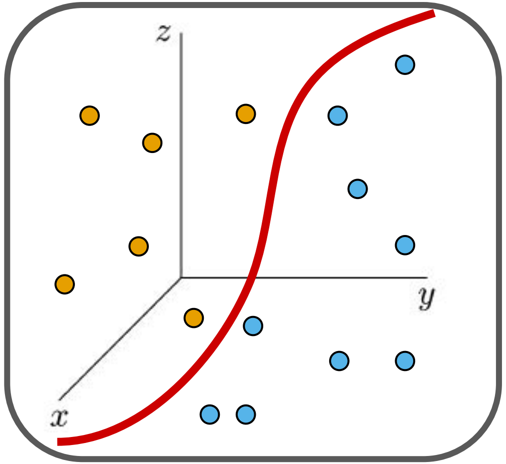
2) Prototypes, or discrete syntactic process?

Is representation parallel across languages?

Train language

Multilingual LM

The dog chases the cat

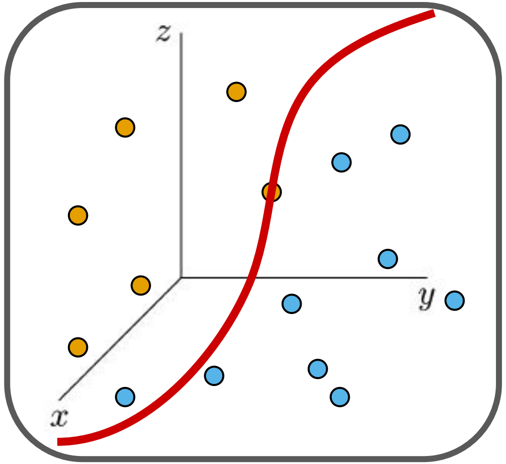


Is representation parallel across languages?

Test language

Multilingual LM

Ο σκύλος κυνηγάει την γάτα

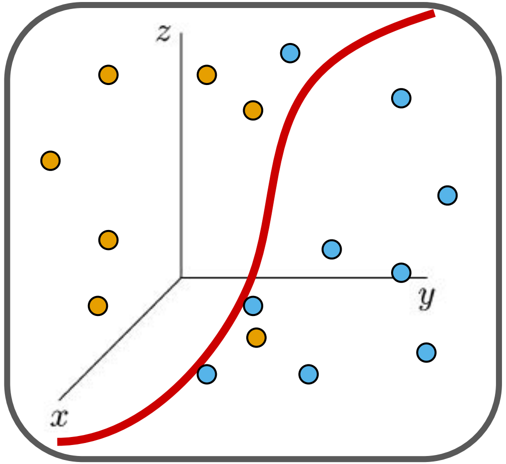


Is representation parallel across languages?

Test language

狗追猫

Multilingual LM



generally pretty good ✅ (24x24 langs)

Model learns a **parallel grammatical abstraction**

[Papadimitriou et al 2021]

But can models encode how languages *differ*?

Languages differ in how they treat **intransitives**:

The glass broke

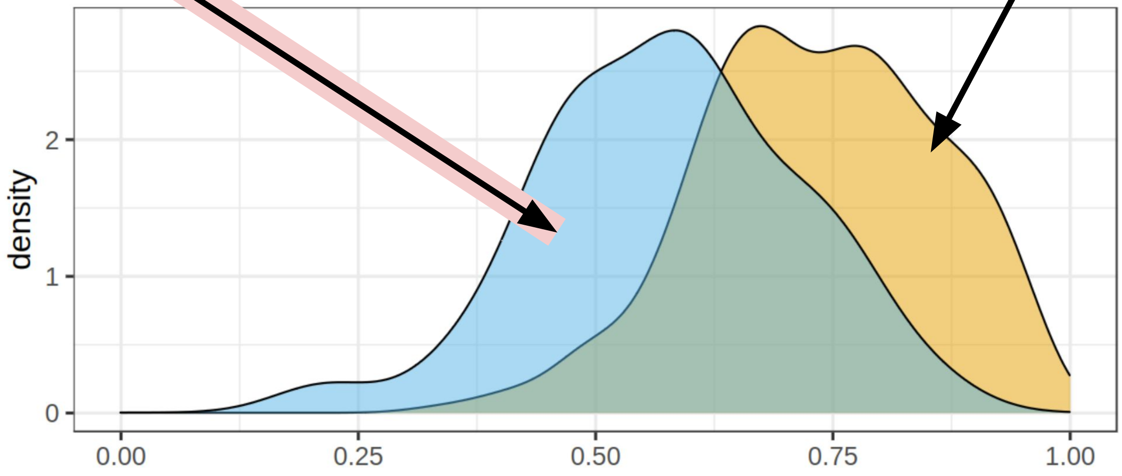
Is this a **subject** or an **object**?

Model recovers language differences

Languages like Basque (intransitives ≈ objects)

Languages like English (intransitive = subject)

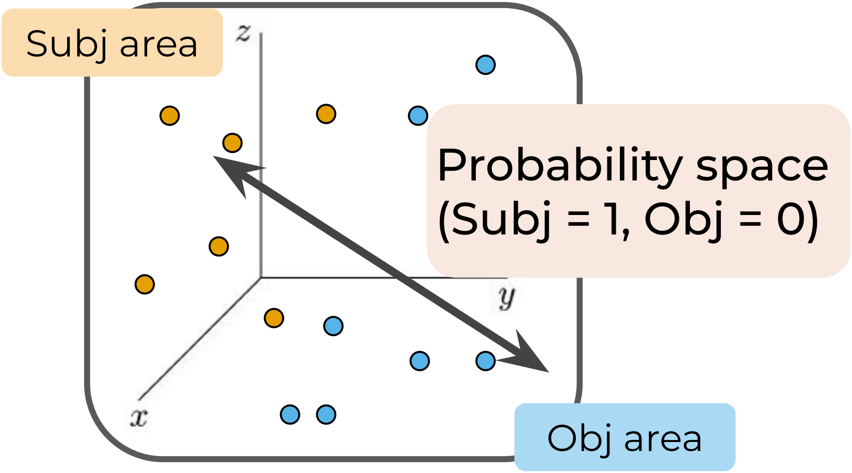
Same test set!



← labelled obj labelled subj →

[Papadimitriou et al 2021]

Two questions about subjecthood representation:

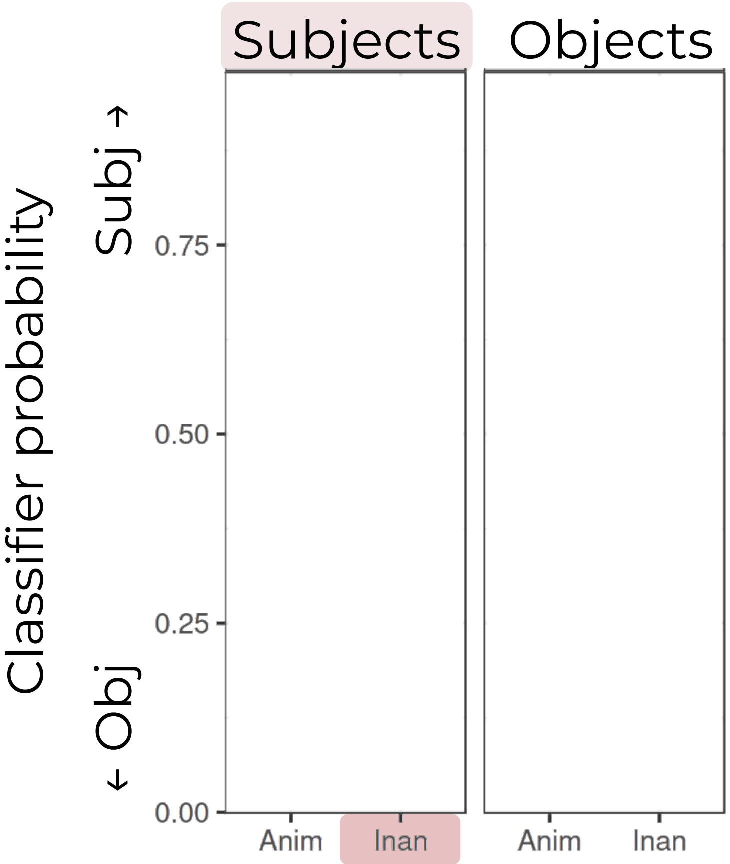


1) How does it work across languages?

2) **Prototypes, or discrete syntactic process?**

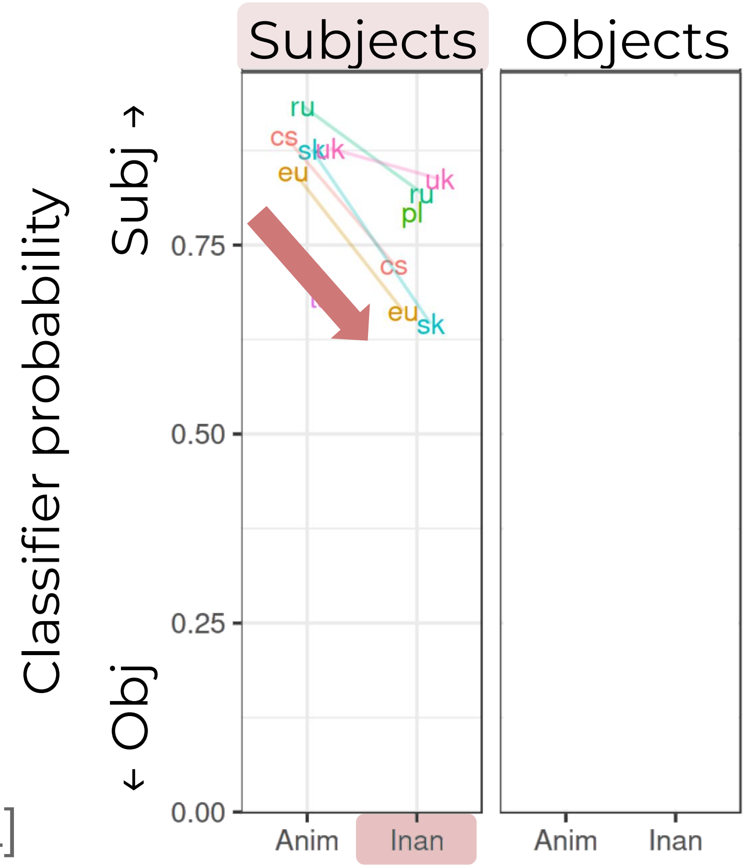
Long story short: both!

Features like **animacy** affect LM subjecthood



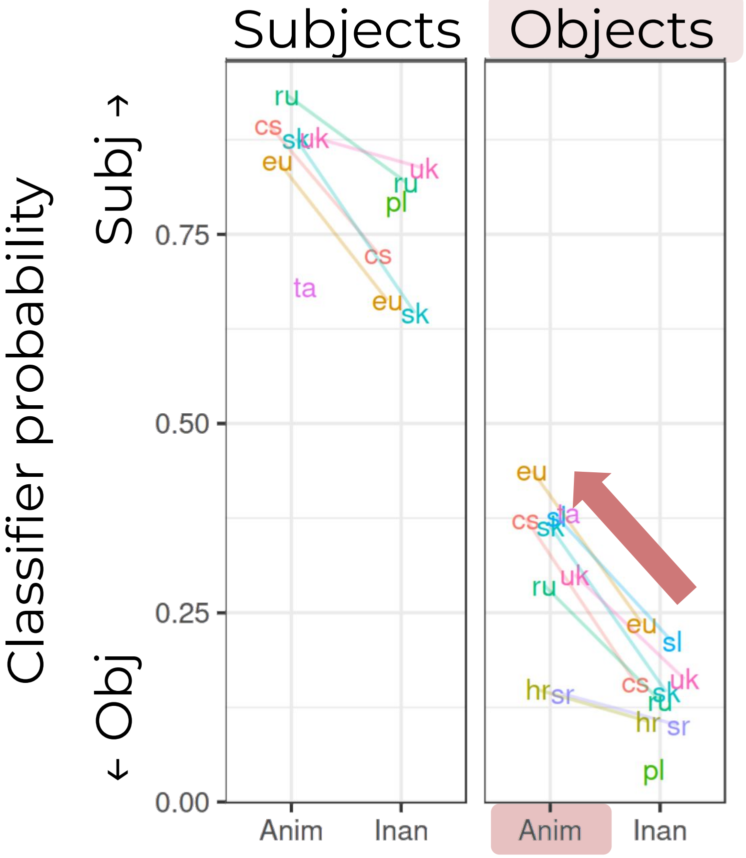
[Papadimitriou et al 2021 EACL]

Features like **animacy** affect LM subjecthood



[Papadimitriou et al 2021 EACL]

Features like **animacy** affect LM subjecthood



[Papadimitriou et al 2021 EACL]

But LM subjecthood is also sensitive to totally **discrete cues**

- What happens when we manually **swap** subjects and objects, but keep everything else **the same**?

The **chef** chopped the **onion**, The **onion** chopped the **chef**

Average $P(\text{subject}) =$ 98.2%

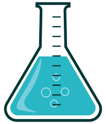
Average $P(\text{subject}) =$ 8.3% ↓

(average over corpus)

Three methodologies using LMs to explore questions of language structure:

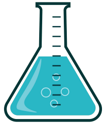
1) **Structural injection**:

   testing different linguistic learning biases



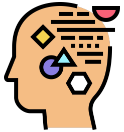
2) **Impossible language learning:**

   what do LMs learn more easily?



3) **Subjecthood in LMs:**

   how are grammatical roles organized in latent space?



LMs can help us experimentally probe the nature of language

- What conditions make language learning possible?
- Where do formal possible/impossible separations come from?
- How can a complex syntactic system be latently represented?

Thanks!