What we can learn about language from exploring multilingual language models

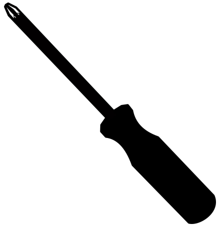
Isabel Papadimitriou

Stanford NLP

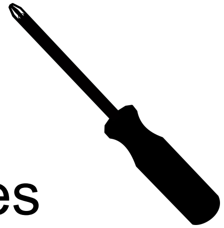
Some context

At last, we have language models that model language (pretty well!)

This gives us two things: a mystery, and a scientific tool



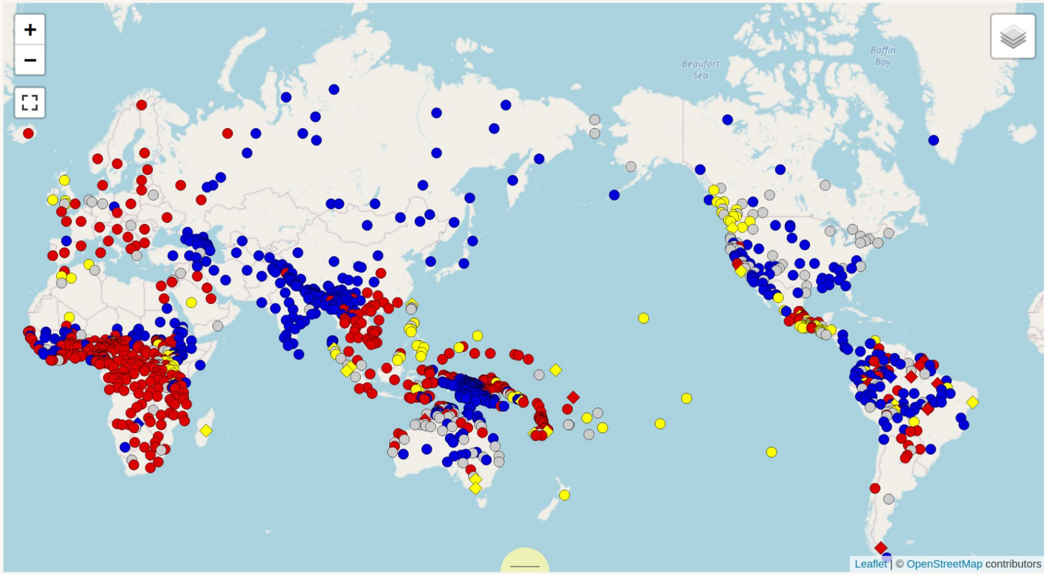
How are language models a tool?



- We have language learners that learn in front of our eyes
- We can investigate this in ways we never could before
- By looking into their representations…
  - We can relate complex **linguistic properties**
- By observing learning under controlled conditions…
  - We can investigate the **inductive learning biases** that contribute to language learning

This talk:

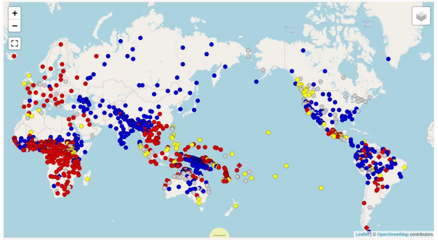
Using a multilingual lens to approach these questions



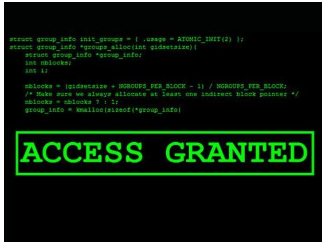
This talk:

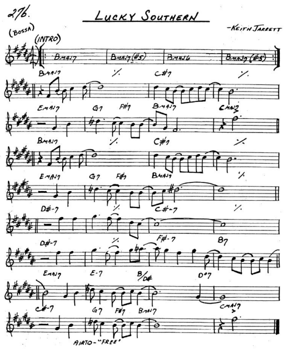
Using a multilingual lens to approach these questions

Human language



Code



Music

Structural Primitives

{ { } [( )] }

(1 2 3) { 1 2 3 }

Language Variation and Universals

Concrete

- How to understand multifaceted, cross-lingual properties?
- **LM Embedding spaces** provide a plausible testing ground.

Abstract

- What **inductive learning biases** make good language learners?
- What are the abstractions that underlie language?

Can we really prove anything?

No

- But an LM is a **concrete theory** for how to model a language
- We can investigate it, and it's outside the box
- Computational models provide **possibilities**, and **interesting cases** we'd not considered

[Baroni 2021, *On the proper role of linguistically-oriented deep net analysis in linguistic theorizing*]

Representing subjecthood

- A discrete category, but with subtleties and complexities
- One coherent continuous space
- How does this work?

Transfer learning with syntactic primitives

{ { } [( )] }

- Pretrain on non-linguistic data
- Create learners with known inductive biases
- A window into language learning

Representing subjecthood

- A discrete category, but with subtleties and complexities
- One coherent continuous space
- How does this work?

Transfer learning with syntactic primitives

{ { } [( )] }

- Pretrain on non-linguistic data
- Create learners with known inductive biases
- A window into language learning

Property: subjecthood

- Who does what to who, being the subject vs the object
- Subjecthood is relevant in basically every utterance, and is handled differently in different languages

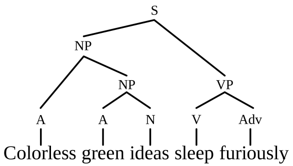
Subjecthood is complicated!

Intransitives

The **glass** broke

Isabel broke the **glass**

Discrete



"There is…"

Animacy

He ran all day

The **fridge** ran all day

Case

…

Passive voice

The **cat** jumped on to the perch

The **perch** was jumped on to by the cat

Volitionality

Mary punched **Sam**

Mary liked **Sam**

Mary forgot **Sam**

[Comrie 1989 *Language Universals and Linguistic Typology*]
[Hopper and Thompson 1980 *Transitivity in Grammar and Discourse*]

Multilingual Language Models

- They represent different **words**, in different **contexts**, in different **languages**
- All in one high-dimensional space

How do they do this for subjecthood?

Subjecthood in Multilingual Language Models

- Subjecthood is a concrete handle for looking into LM internals
- LMs give us a concrete view of how multilingual subjecthood ***can*** be represented and influenced

Deep Subjecthood: Higher-Order Grammatical Features in Multilingual BERT

Isabel Papadimitriou
Stanford University
isabelvp@stanford.edu

Ethan A. Chi
Stanford University
ethanchi@cs.stanford.edu

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Kyle Mahowald
University of California, Santa Barbara
mahowald@ucsb.edu

(EACL 2021)

Three questions:

- Is subjecthood a universal category?
- Is subjecthood a discrete category?
- What happens with typological variation?

Main experimental tool:

- Train a <u>binary classifier</u> on mBERT embeddings to distinguish **subjects** from **objects** in one language
- Zero-shot transfer classifiers from one language to another

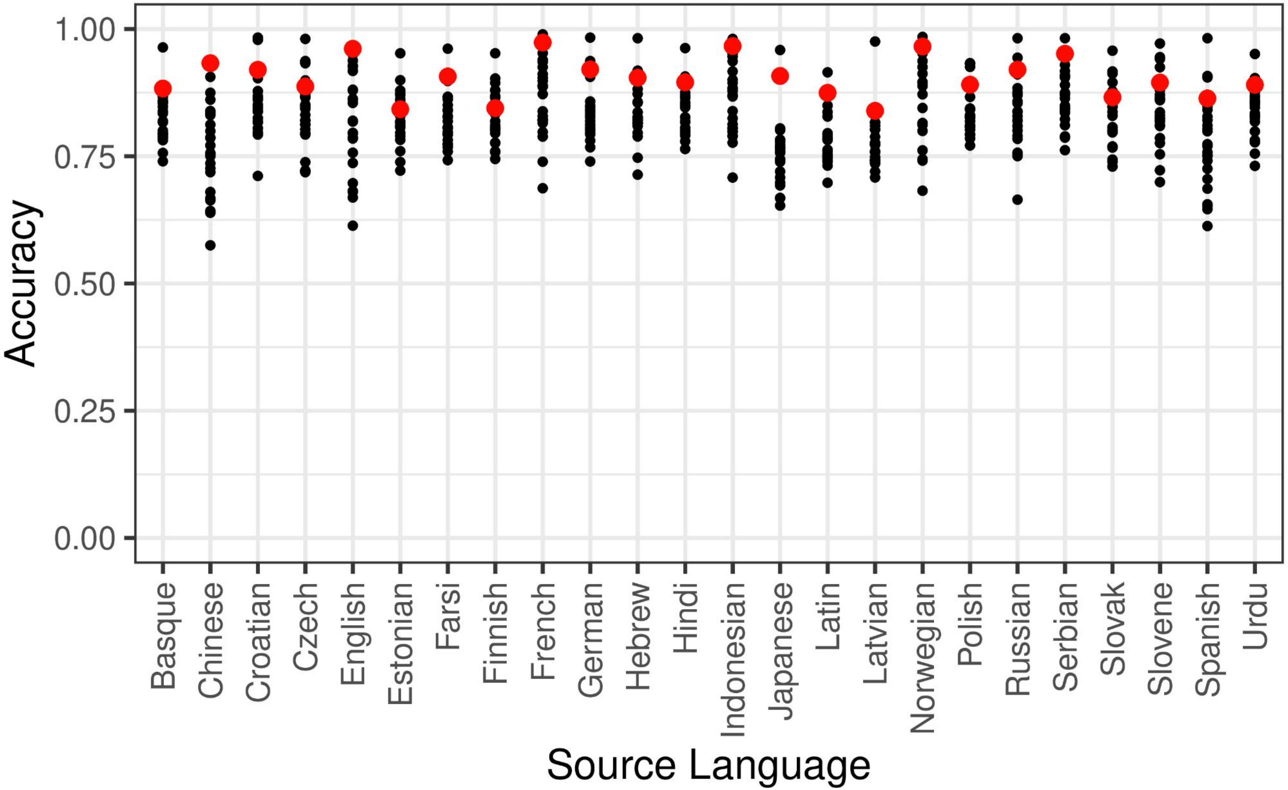
✅ *Subject* / *Object*
The **cat** is chasing the **dog**

→

?? / *??*
Τον **σκύλο** τον κυνηγάει η **γάτα**

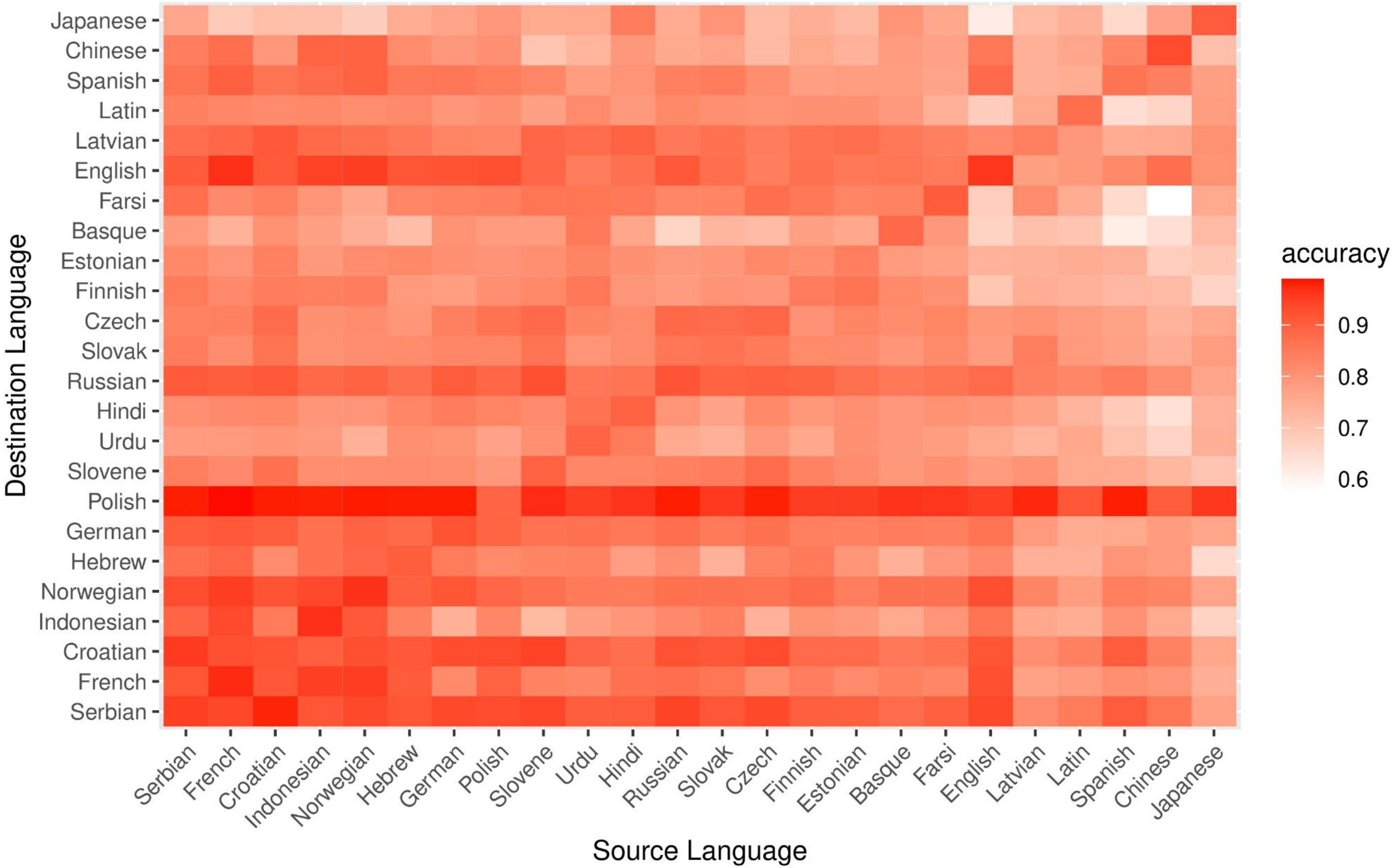
?? / *??*
这只**猫**在追那条**狗**

Cross-lingual accuracy is comparable to in-language



Red dots are in-language accuracy, black dots are cross-language

Cross-lingual accuracy is comparable to in-language



Parallel, Multilingual Subjecthood

- **Linguistic generalization** in pretrained LMs:
  - Encode subjecthood separately from language

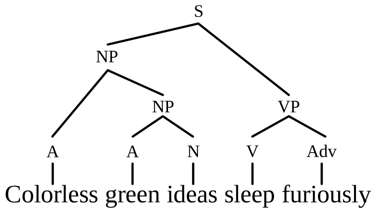
- Subjecthood is available to a learner as a universal

Three questions:

- Is subjecthood a universal category?
- Is subjecthood a discrete category?
- What happens with typological variation?

But is subjecthood a simple binary issue?

Discrete



vs.

Prototype

Animacy,
Passive voice,
Volitionality,
Agency,
Case,
...

- Different views on how to think of subjecthood
- Multilingual LMs can help us tease out this conflict

Main experimental tool:

- Train a <u>binary classifier</u> on mBERT embeddings to separate **subjects** from **objects** in one language

Probabilities: Subjecthood projection space

Object
P(subject) = 0

Subject
P(subject) = 1

- ***How***, not ***if***, the classifier encodes subjecthood

Main experimental tool:

- Train a <u>binary classifier</u> on mBERT embeddings to separate **subjects** from **objects** in one language

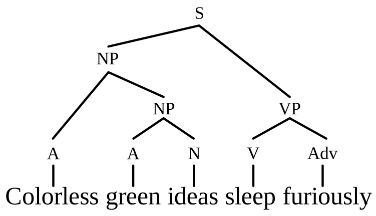
Probabilities: Subjecthood projection space

Object
P(subject) = 0

Subject
P(subject) = 1

- ***How***, not ***if***, the classifier encodes subjecthood

Discrete



vs.

Prototype

Animacy,
Passive voice,
Volitionality,
Agency,
Case,
…

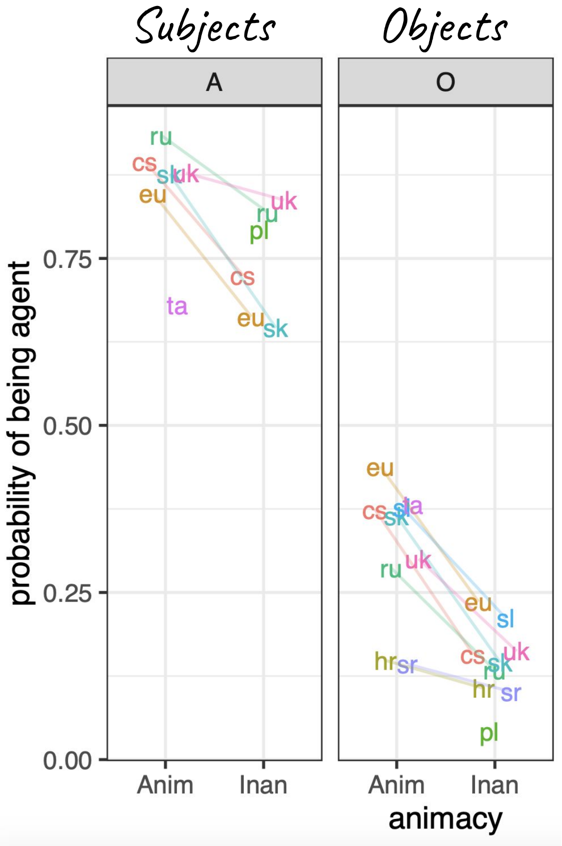
- Do probe probabilities reflect the effect of other features?

Classifier probabilities show animacy effects

Animacy

He ran all day

The **fridge** ran all day



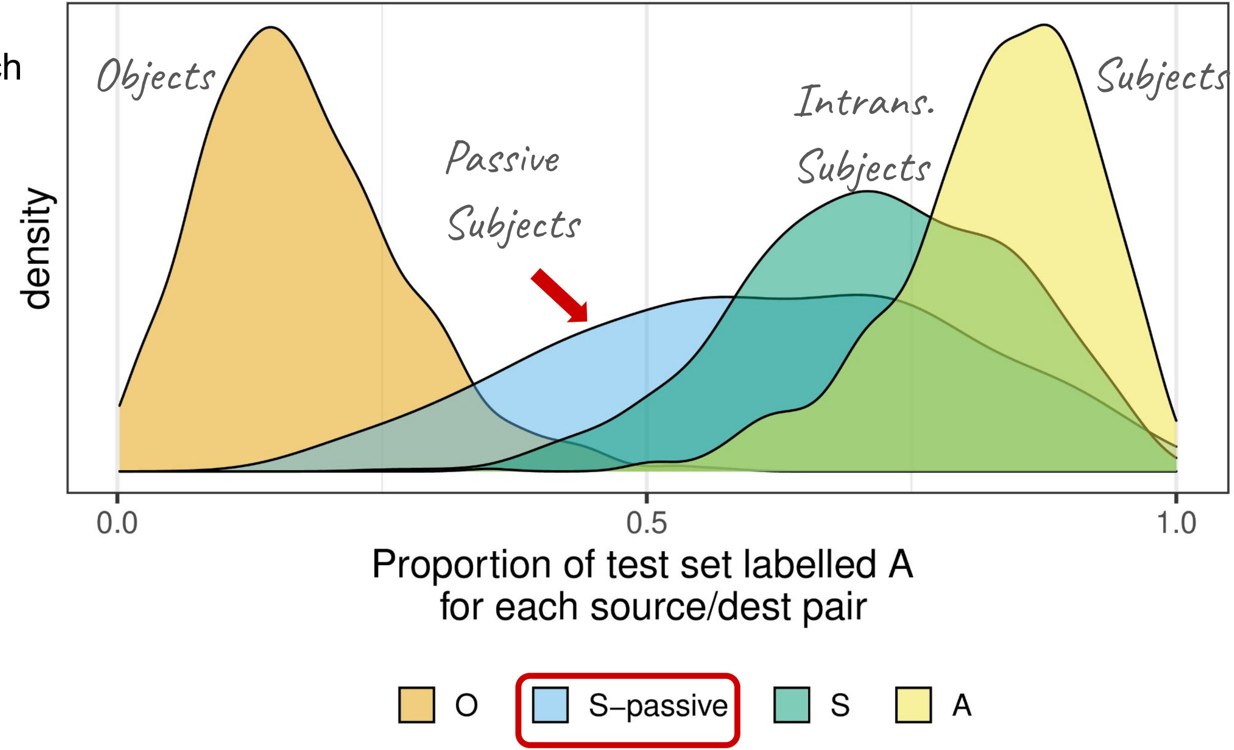
*Even when **controlling for syntactic role**, animacy has an effect*

Classifier probabilities show passive voice effects

Passive voice

The **cat** jumped on to the perch

The **perch** was jumped on to by the cat



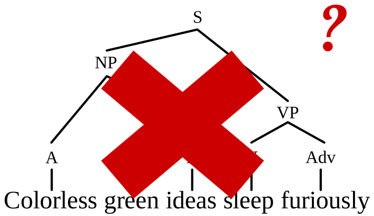
- We see **prototype effects** in mBERT embeddings

- **Many factors** play into making something a subject

Also look at the effect of **case**.
Future work: discourse, information structure (given/new)

But is it all just prototypes?

Discrete



vs.

Prototype

Animacy,
Passive voice,
Volitionality,
Agency,
Case,
…

When classifying grammatical role, BERT doesn't care about word order...except when it matters

Isabel Papadimitriou
Stanford University
isabelvp@stanford.edu

Richard Futrell
University of California, Irvine
rfutrell@uci.edu

Kyle Mahowald
The University of Texas at Austin
mahowald@utexas.edu

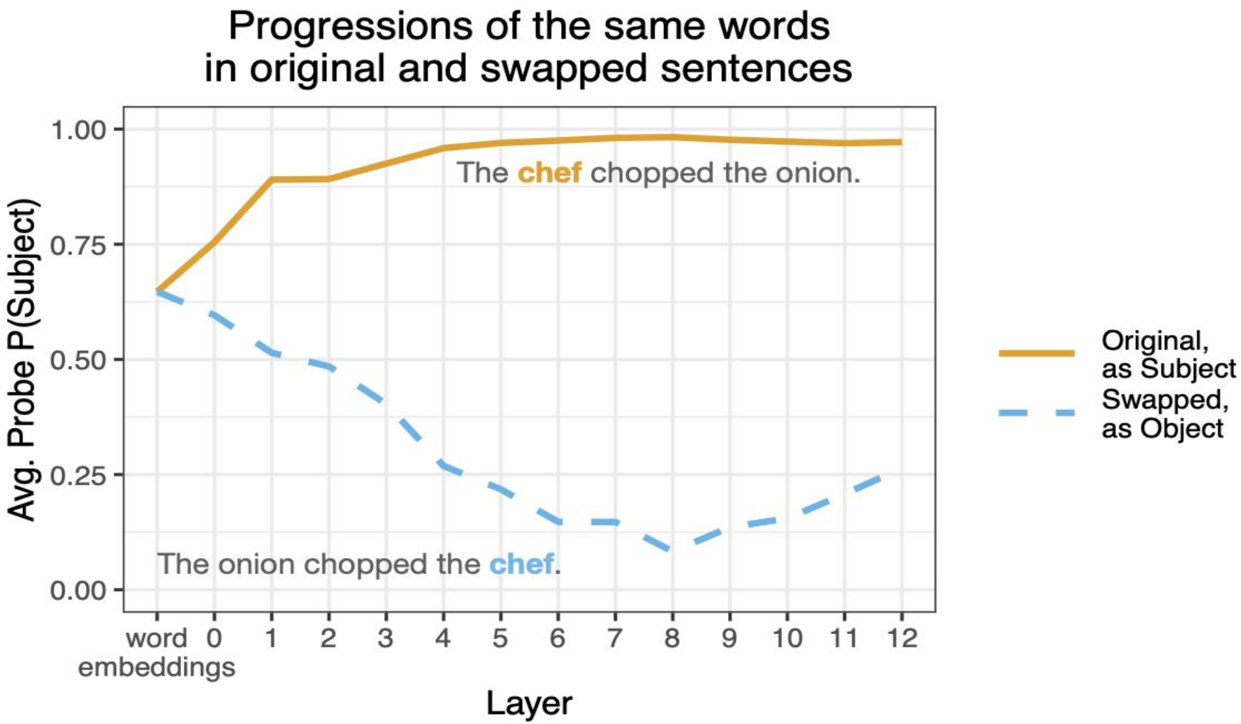
(ACL 2022)

What if we test the probe on the same sentences (*with the same prototype effects*) but we **swap the labels**?

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

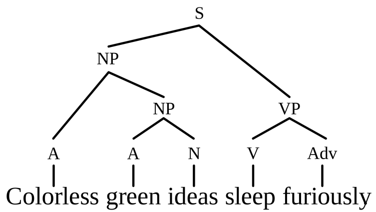
Will the probe tell them apart?

Yes – Representation differences that are caused **only** by syntactic word order



Both grammatical subjecthood and prototype effects

Discrete



&

Prototype

Animacy,
Passive voice,
Volitionality,
Agency,
Case,
...

- Future work: *How* can a representation embody both of these types of information?
- LMs as a tool to **better understand this middle ground**

Three questions:

- Is subjecthood a universal category?
- Is subjecthood a discrete category?
- What happens with typological variation?

Typological variation: **Intransitives**

- Subjecthood is encoded in parallel between languages
- But are the **particularities** of each language also encoded?
- **Do we see variation in treatment of intransitives?**

Typological variation: **Intransitives**

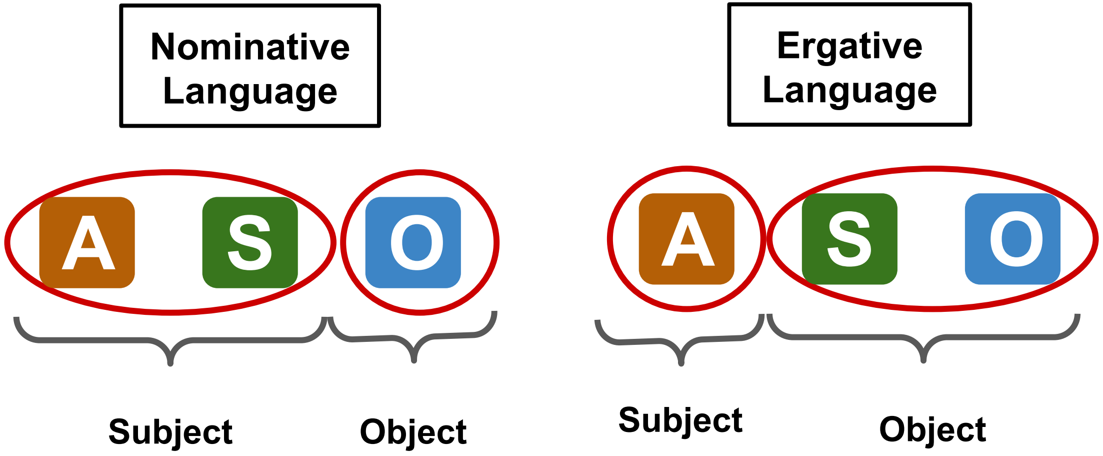
A O

Transitive: The **dog** chased the **cat**

S

Intransitive: The **glass** broke

Ergative languages treat **intransitive subjects** like **objects**



Typological variation: **Intransitives**

- Subjecthood is encoded in parallel between languages
- But are the **particularities** of each language also encoded?
- **Do we see variation in treatment of intransitives?**
  - Can higher-order information be represented in embedding space?

Hold out intransitives from classifier training

Probabilities: Subjecthood projection space

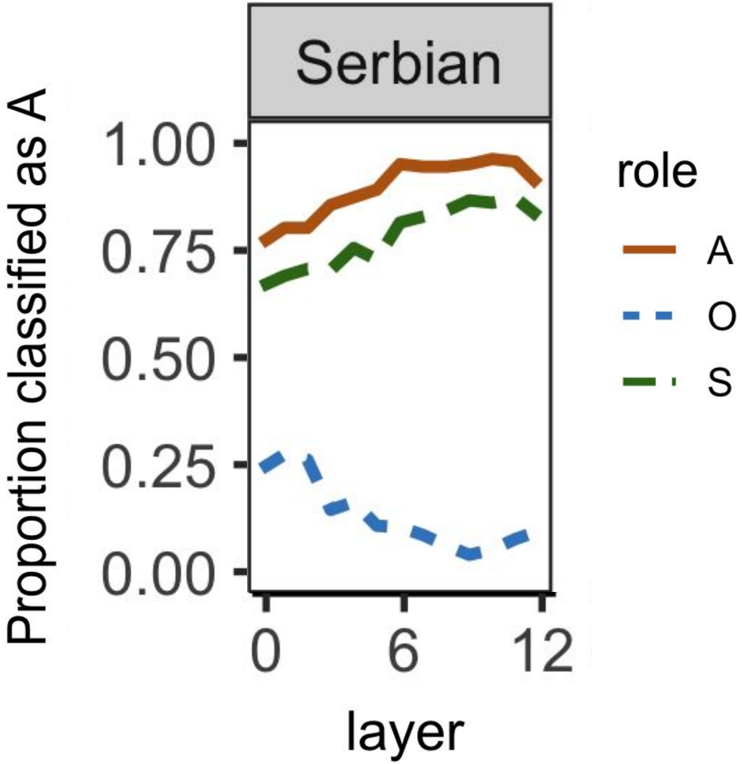
Transitive object
P(subject) = 0

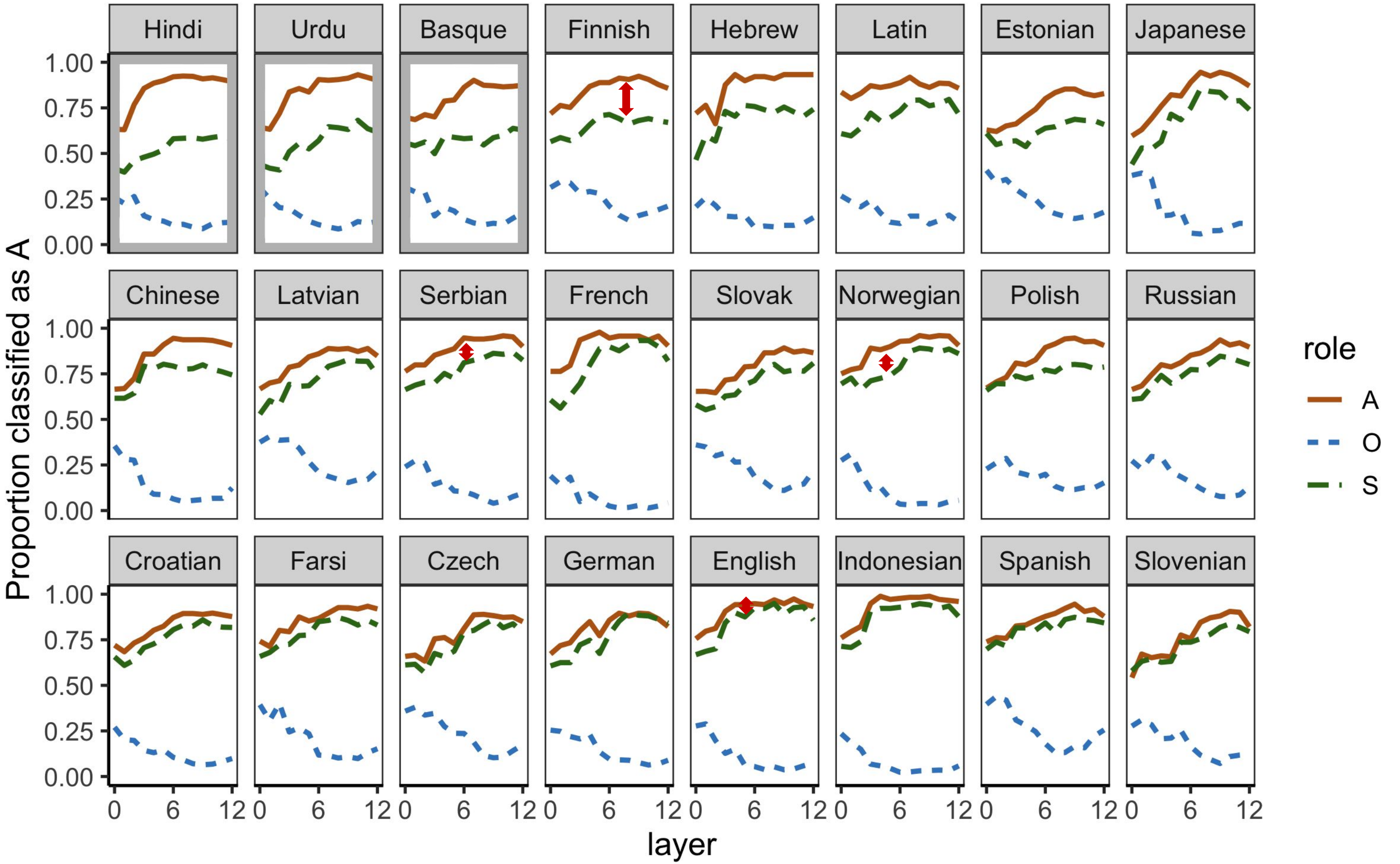
Transitive subject
P(subject) = 1

Unseen: Intransitives
P(subject) = ??

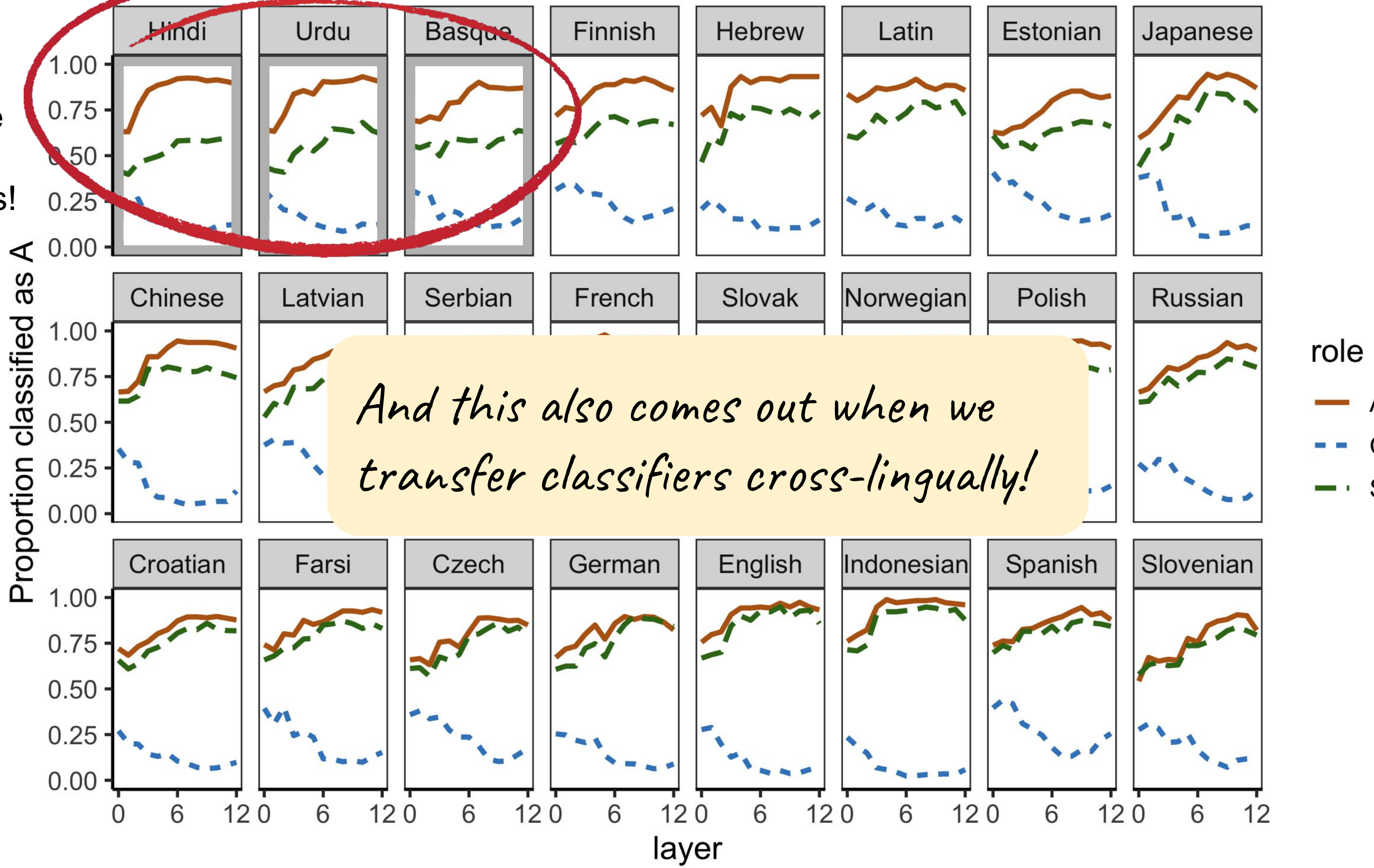
- Classifier probabilities show how intransitives align

Transitive Subjects (**A**) > Intransitive Subjects (**S**) > Transitive Objects (**O**)




Hindi
Urdu
Basque
Finnish
Hebrew
Latin
Estonian
Japanese
Chinese
Latvian
Serbian
French
Slovak
Norwegian
Polish
Russian
Croatian
Farsi
Czech
German
English
Indonesian
Spanish
Slovenian
Proportion classified as A
1.00
0.75
0.50
0.25
0.00
layer
0
6
12
role
A
O
S

These are ergative languages!



And this also comes out when we transfer classifiers cross-lingually!

Classifiers Reflect Intransitive Alignment

- Alignment of **intransitives** is a feature of a grammar, **not of any one utterance**
- But it is apparent in embedding space, even when they are held out

Subjecthood: what we learned

- Subjecthood representation can be, and is, **multilingual**
- Prototype effects **co-exist** with discrete grammatical classes
- **Higher-order information** (like what to do with intransitives) is represented in the same space as meaning

Future work: better understanding the geometric expression of these properties

Representing subjecthood

- A discrete category, but with subtleties and complexities
- One coherent continuous space
- How does this work?

Transfer learning with syntactic primitives

{ { } [( )] }

- Pretrain on non-linguistic data
- Create learners with known inductive biases
- A window into language learning

Learning Music Helps You Read: Using Transfer to Study Linguistic Structure in Language Models

Isabel Papadimitriou
Stanford University
isabelvp@stanford.edu

Dan Jurafsky
Stanford University
jurafsky@stanford.edu

(EMNLP 2020)

Main Question:

What structural inductive biases make a good language learner?

We can't really have blank-slate learners that work

- Small networks can't model the data well
- Large models come with many inductive biases

This paper:

- But, (if we're careful about data) pre-training creates a powerful learner with a known inductive learning bias

[Baroni 2021, *On the proper role of linguistically-oriented deep net analysis in linguistic theorizing*] untrained model, unknown inductive biases

Pretraining,
non-linguistic

Learner whose inductive biases we know

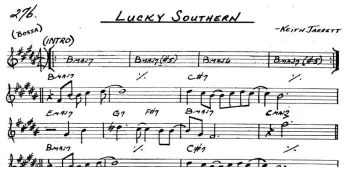
(Because we pretrained them in!)

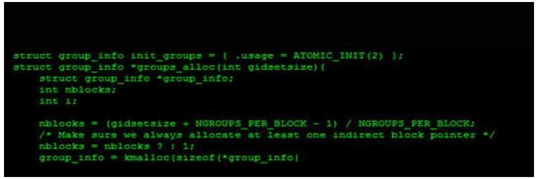
Transfer
learning

How well can this model
learn from **language** data?

Pretraining data

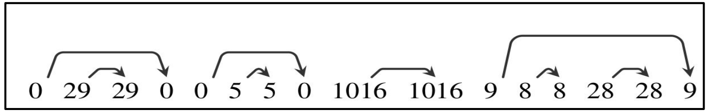
Real Data:

Music

Code

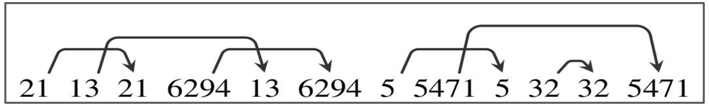


Synthetic Structural Primitives:

Hierarchical



Non-hierarchical



Transfer learning should be *constrained*

- We want to make sure that we're using inductive biases, not re-pretraining

- Two ways of constraining transfer learning:
  - Limit **data**
  - Limit **trainable parameters**

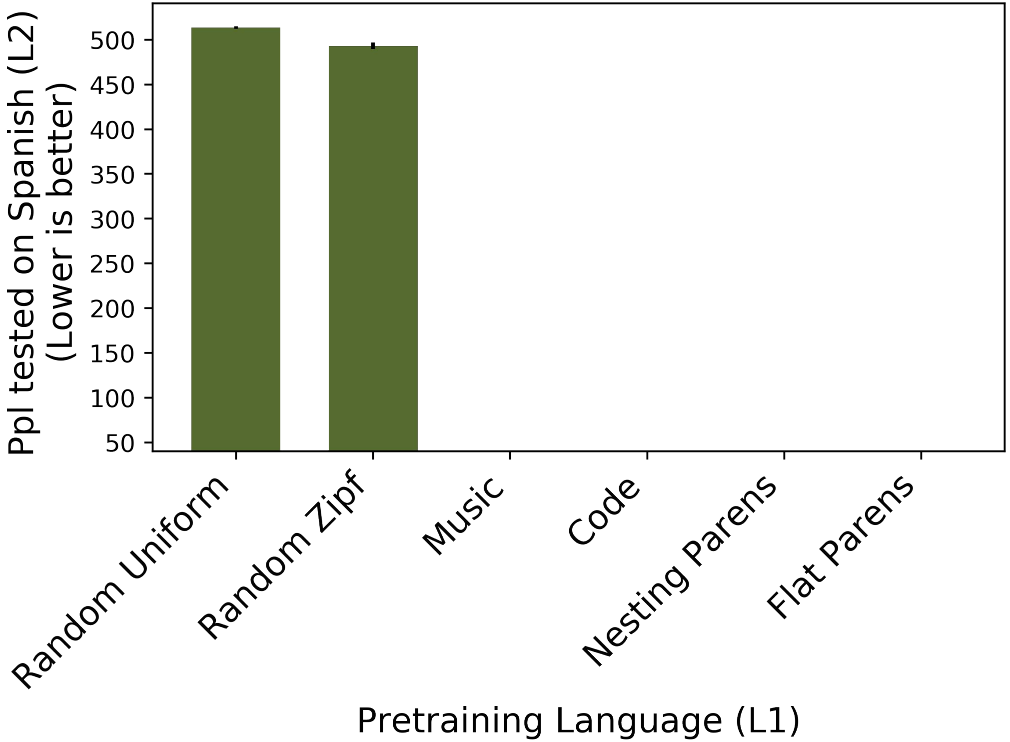
Transfer learning should be *constrained*

- We want to make sure that we're using inductive biases, not re-pretraining

- Two ways of constraining transfer learning:
    - Limit **data**
    - Limit **trainable parameters**

Freeze everything except **word embeddings**. Can LM internals be effectively repurposed?

Random Baselines – Randomly sampled tokens



- Control - How far can we get with just word embeddings?
- Vocabulary distribution has a significant effect

Music and Code

Music: MIDI

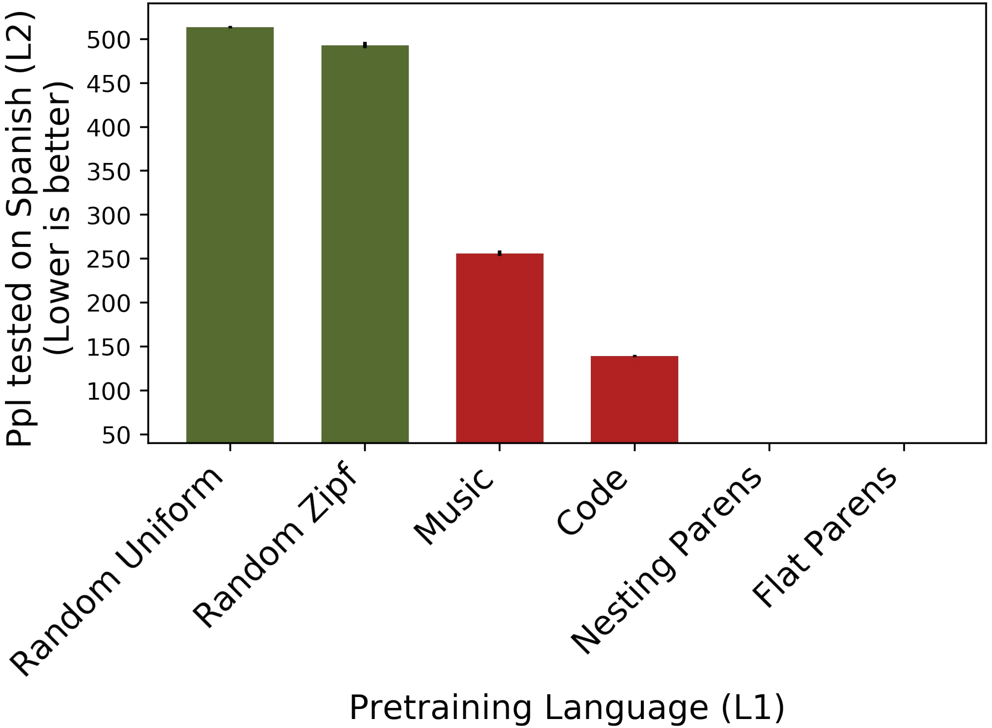
```
SETVELOCITY_29 NOTEON_47 SHIFTMS_180 SETVELOCITY_57 NOTEON_66...
```

No comments

Code: Java

```
if ( coordFactor == 1 ) { return sum ; } else { result =
...
```

- Non-linguistic, structured data, with **different surface forms**
- Is this structural bias helpful for language modeling?

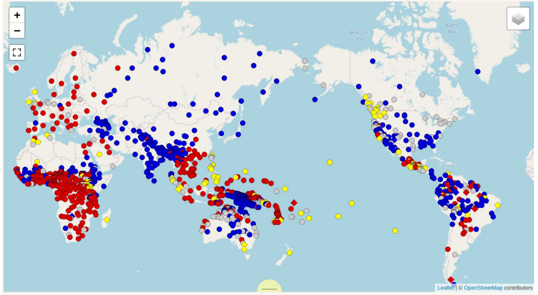


- Impressive improvement in perplexity
- MIDI surface form is very different (and vocabulary is just 310 tokens)
- But music and language have structural similarities

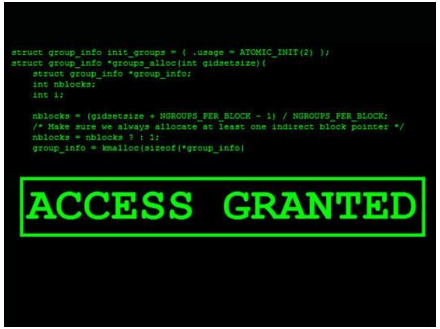
Is this because of hierarchical structures?

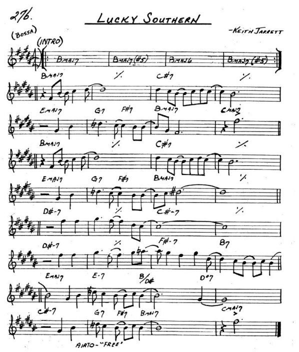
{ { } [( )] }

Human language



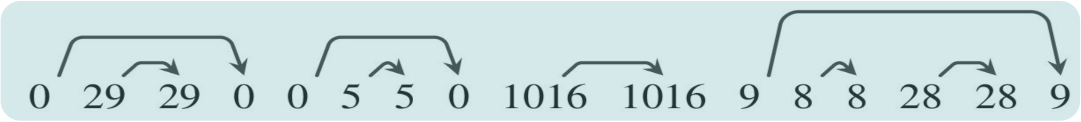
Code



Music

This is testable

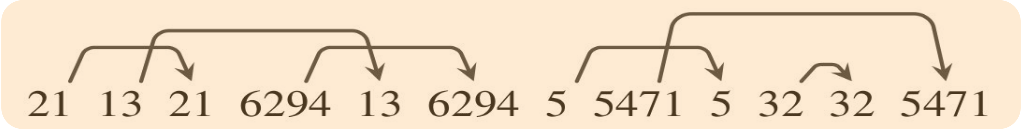
Pretrain on a simple hierarchical structure:

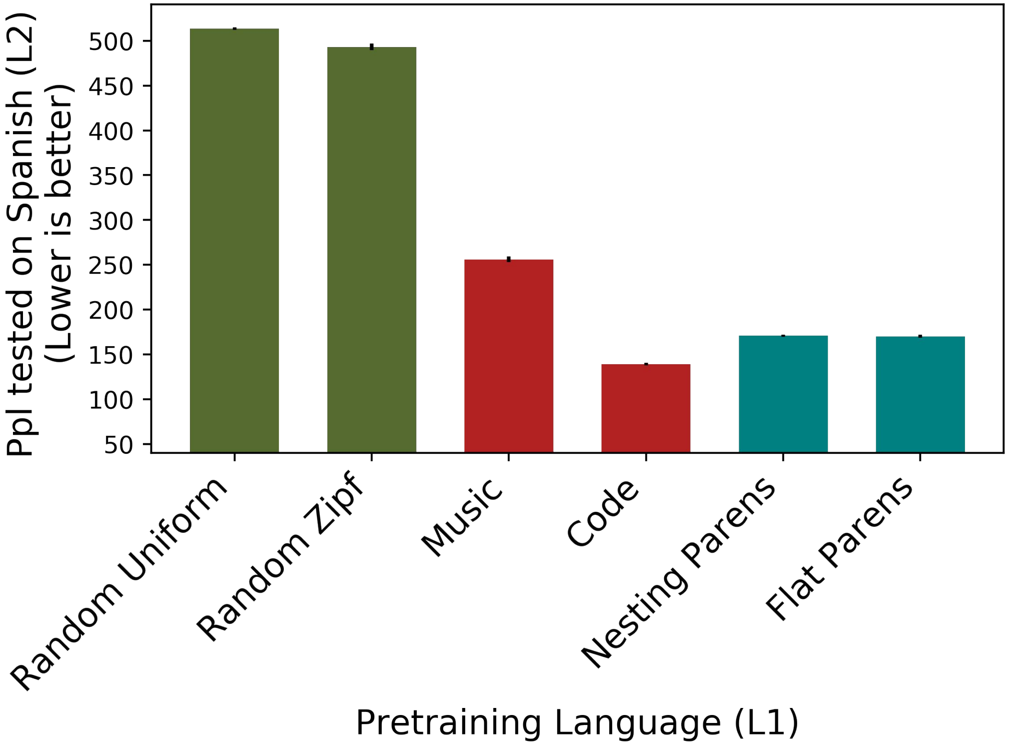
Nesting (Recursive) Parentheses:



But also have a control:

Flat Parentheses:



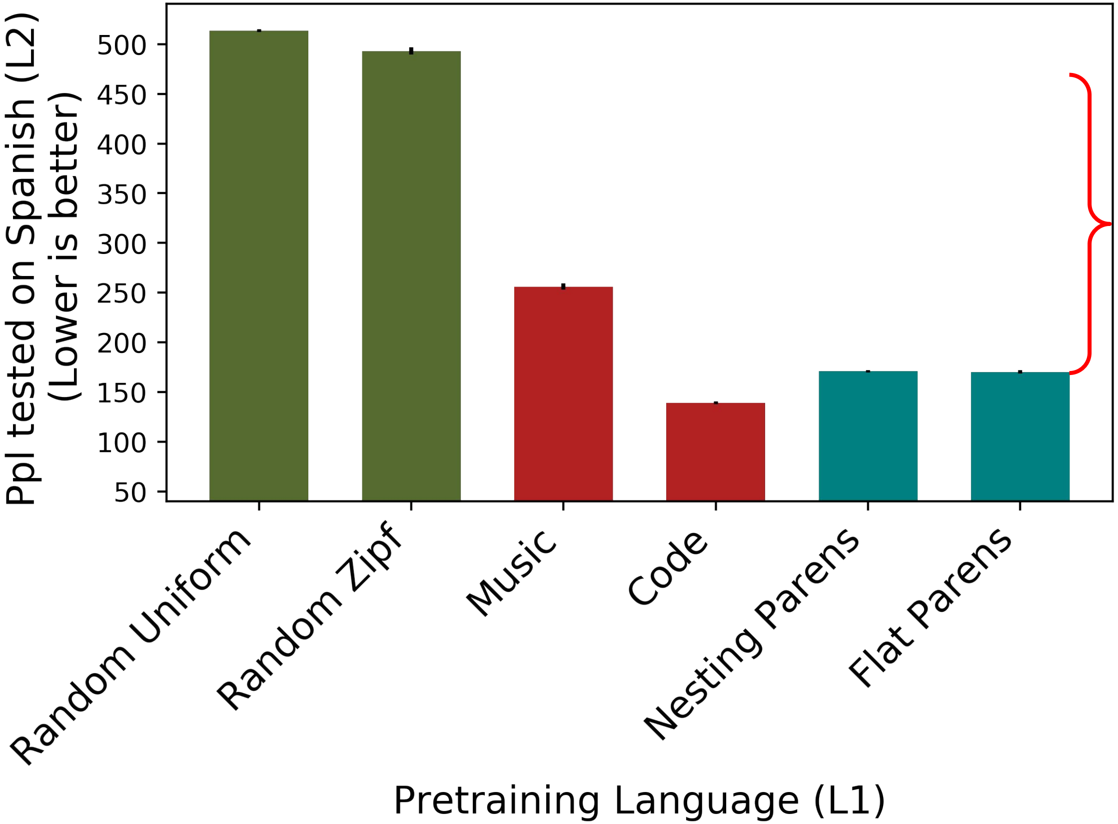


- **Simple underlying structure** causes huge increase in performance compared to random
- Flat parentheses are as good as hierarchical parentheses

Parentheses inductive bias is much better than random

- But the Flat Parentheses corpus is very similar to the Random corpus

Example: 21 13 21 6294 13 6294 5 5471 5 32 32 5471 …

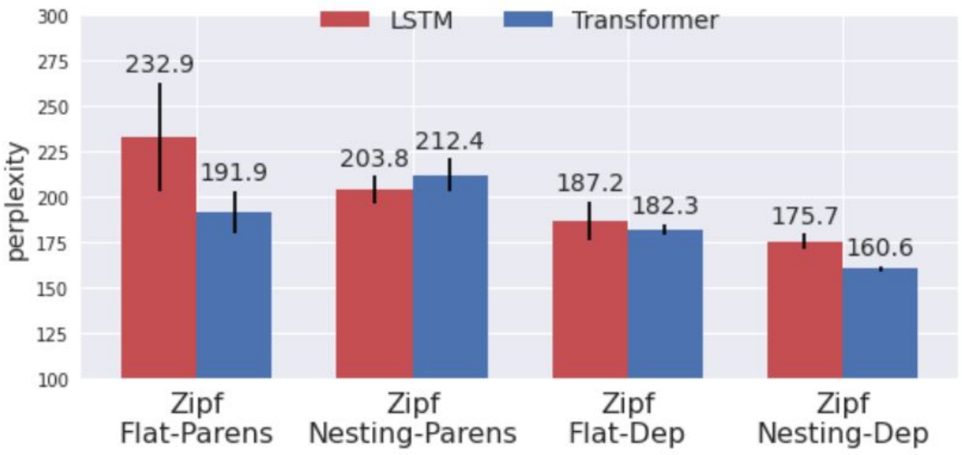


Difference from placing a random token twice instead of once

What to take away from these experiments?

- A structural inductive bias **(but not necessarily hierarchical)** helps learn language
- Flat, head-to-head dependencies are an important learning bias to consider

Results have been reproduced in transformers



(b) Comparison of dependency structures.

[Ri and Tsuruoka, 2022, *Pretraining with Artificial Language:Studying Transferable Knowledge in Language Models*]

[Chiang and Lee, 2021, *On the Transferability of Pre-trained Language Models: A Study from Artificial Datasets*]

Flat parentheses in the wild

Kundan Krishna, Jeffrey Bigham, Zachary C. Lipton (2021) *Does Pretraining for Summarization Require Knowledge Transfer?*

- Take a nonsense (random) corpus,
- Create “summarization” input-output pairs with **simple summarization-type dependencies**
- Good downstream performance!

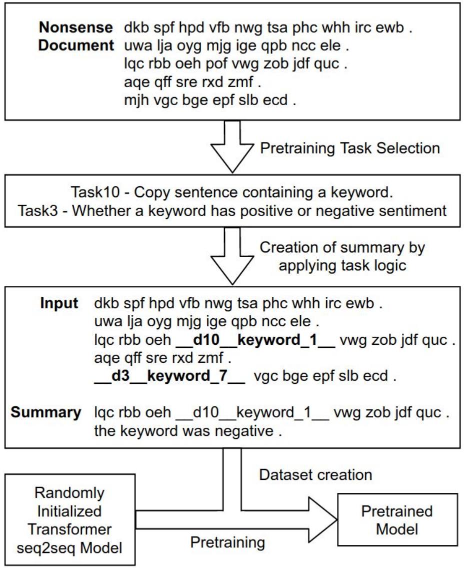


Figure 1: Procedure to create pretraining dataset using the nonsense corpus and our proposed pretraining tasks

How about more **language-like structures**?

- When we transfer between languages, transfer is correlated with **typological syntactic distance**

There's a correlation – but can we test causes?

[Wu*, Papadimitriou*, Tamkin*, 2022, *Oolong: Investigating What Makes Crosslingual Transfer Hard with **Controlled Studies***]

- Subjecthood: One embedding space can encode
  - A property **generalized** across languages
  - A discrete property also influenced by **prototype** effects
  - **Higher order features** of the language

{ { } [( )] }

- Structural primitives:
  - We're in a unique position – we can make **powerful learners** imbued with **known inductive biases**
  - Flat dependencies are an important and interesting bias

What can we learn from LMs?

- The embedding space of multilingual LMs suggests how the **complexities and dualities** of universal properties like subjecthood might function

- Pretraining with structural primitives demonstrates what **starting points** make language learning possible

Thanks!