Successes and Failures of Compositionally in Neural Networks for Language

Tal Linzen
Department of Linguistics and Center for Data Science
New York University

Can neural networks for language generalize compositionally?

- It depends!
- For models trained ("pretrained") to predict missing words:
  - Is generalization measured on the original objective or on transfer to a new task?
- What is meant by compositionality? Distinguishing lexical and structural generalization

When generating text, does GPT-2 parrot its training corpus?

Novel words constructed using English morphological rules, used in a syntactically appropriate way:

(11) a. I love **Klymits** , but it has been nearly impossible for us to find them in stores .
b. The **Sarrats** were lucky to have her as part of their lives

(12) a. these small **townites**
b. so many **Brazilianisms**

(McCoy, Smolensky, Linzen, Gao & Celikyilmaz, 2021, *arXiv*)

When generating text, does GPT-2 parrot its training corpus?

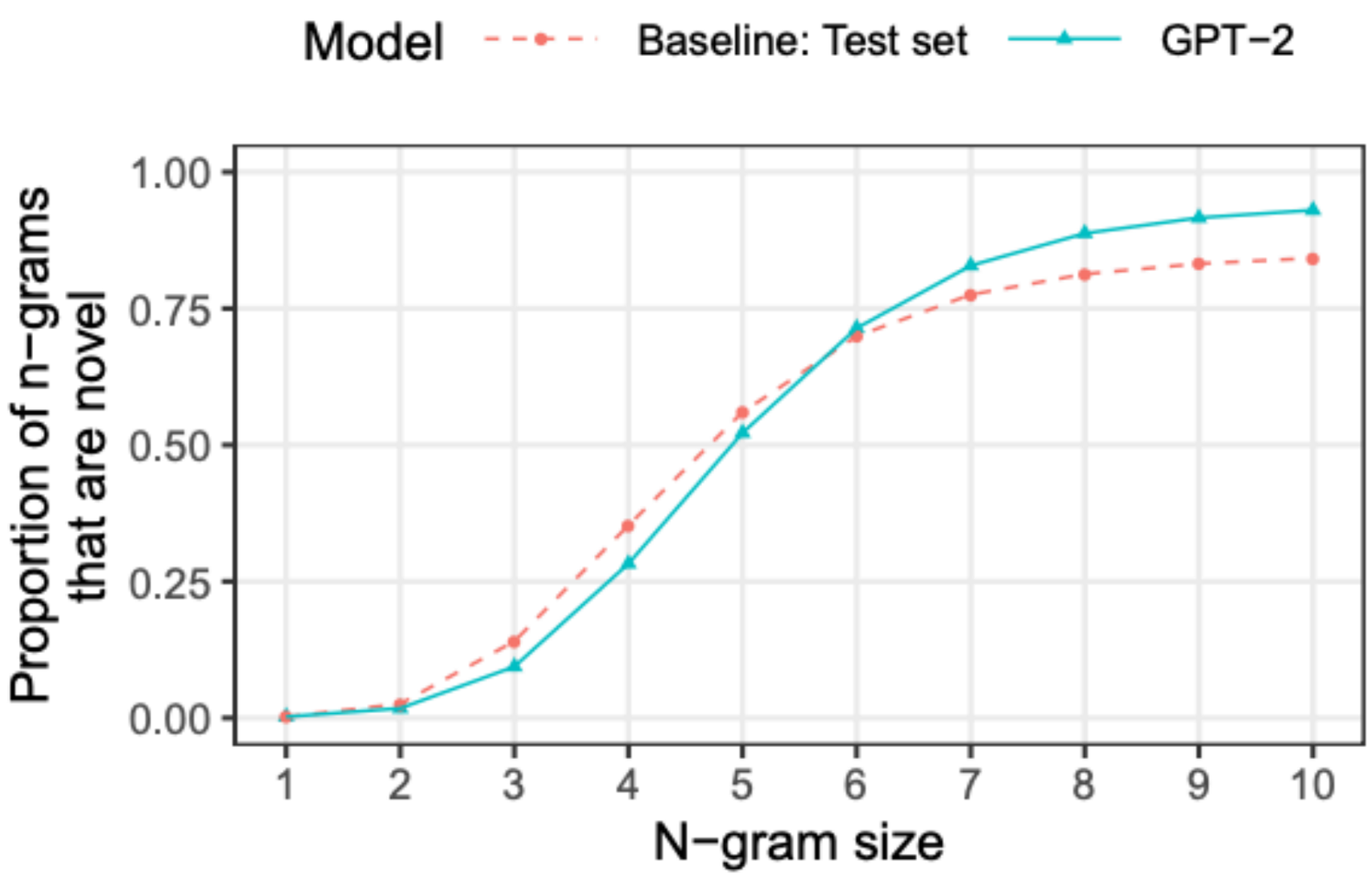
A verb seen in training only in the passive voice is used in the active voice:

(i) *Transformer:* They then **dry-docked** at Sasebo on 22 January 1916 to be fitted with an additional 4.5 cm / 40 anti-aircraft (AA) guns.

(ii) *Training example:* Ostfriesland was **drydocked** in Wilhelmshaven for repairs , which lasted until 26 July

(McCoy, Smolensky, Linzen, Gao & Celikyilmaz, 2021, *arXiv*)

When generating text, does GPT-2 parrot its training corpus?

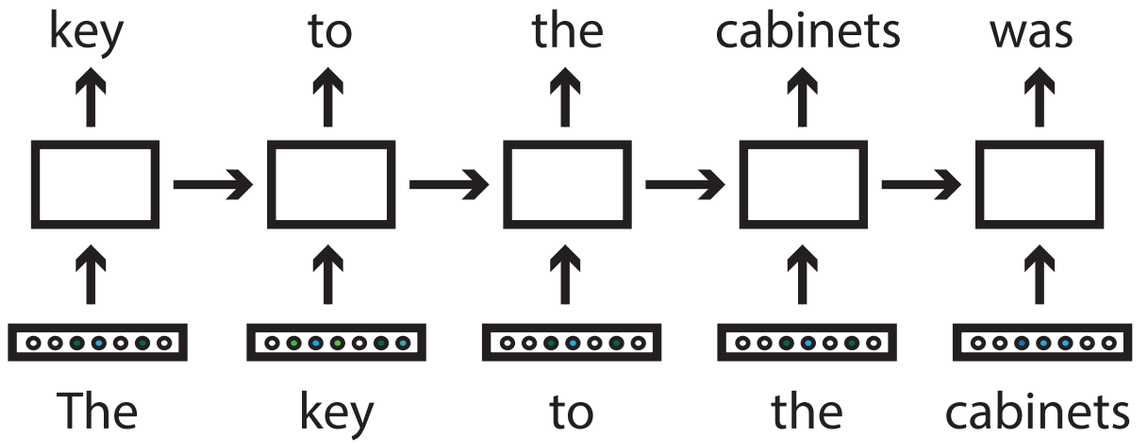


(McCoy, Smolensky, Linzen, Gao & Celikyilmaz, 2021, *arXiv*)

Probing a language model's syntactic representations using the number prediction task

The length of the forewings… **Singular**

The keys to the cabinets… **Plural**



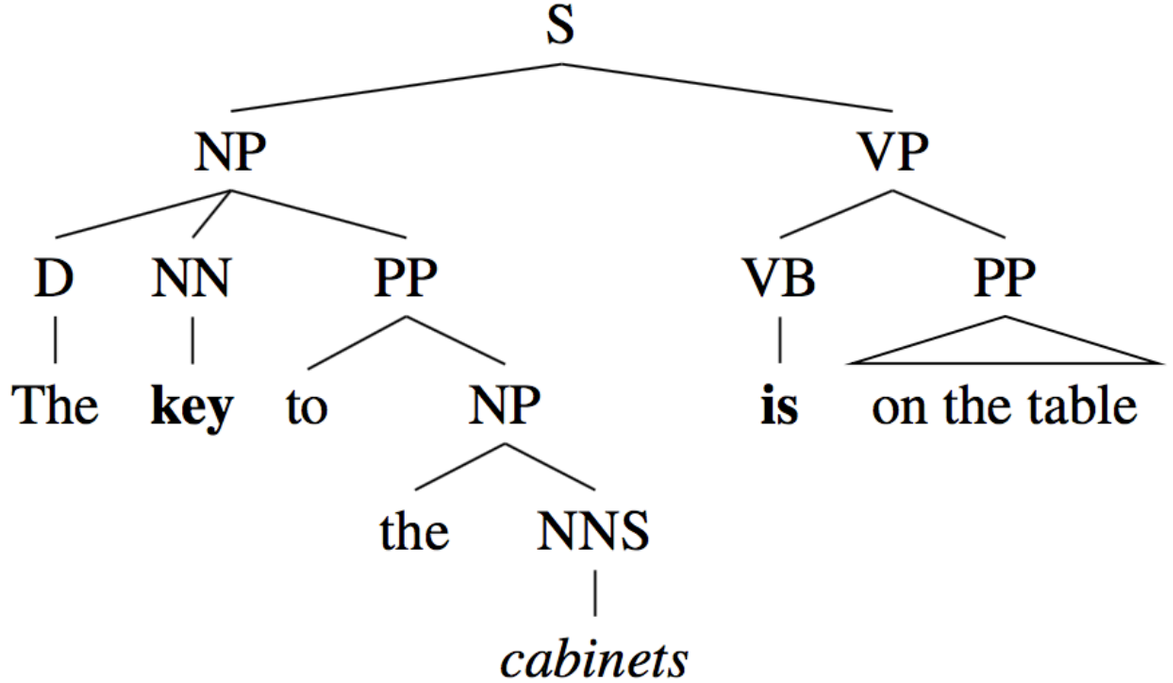
$$\hat{P}(w_n = w^k \mid w_1, \ldots, w_{n-1})$$

The keys to the cabinets…. P(***were***) > P(***was***)?

(Bock & Miller, 1991; Elman, 1991; Linzen, Dupoux & Goldberg, 2016)

This is easy, with the right representations

The **key** to the **cabinets** **is** on the table**.**



Are neural networks that are not designed around such structural representations able to do this task?

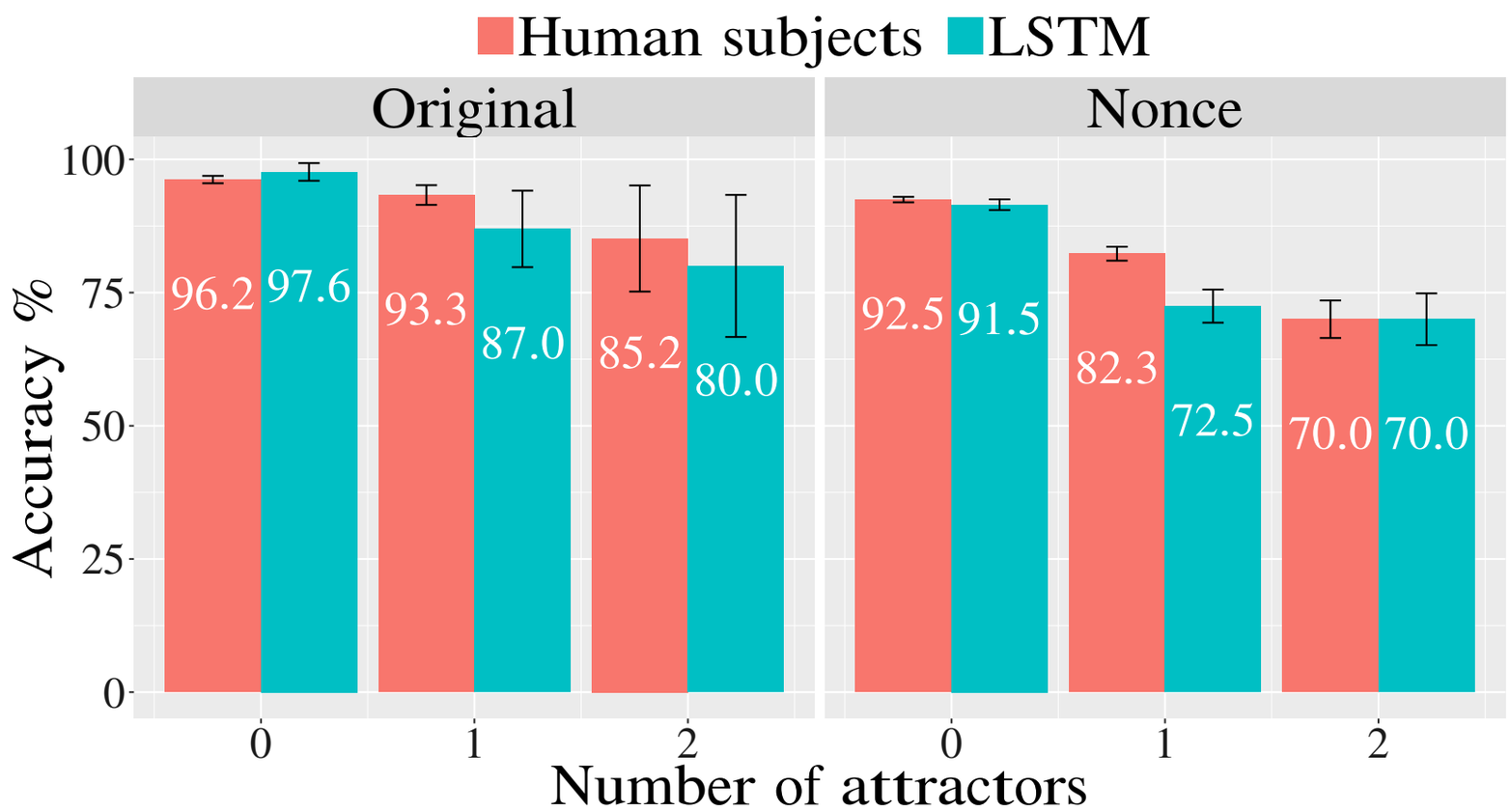
Test sentences

- The angry brown **dogs** that sit by the **cat bark/barks** furiously.

- The colorless green **ideas** that sit by the **honor sleep/sleeps** furiously.

(Gulordava, Bojanowski, Grave, Linzen & Baroni, 2018, *NAACL*)

Language models generalize grammatical rules to novel sentences



(Gulordava, Bojanowski, Grave, Linzen & Baroni, 2018, *NAACL*)

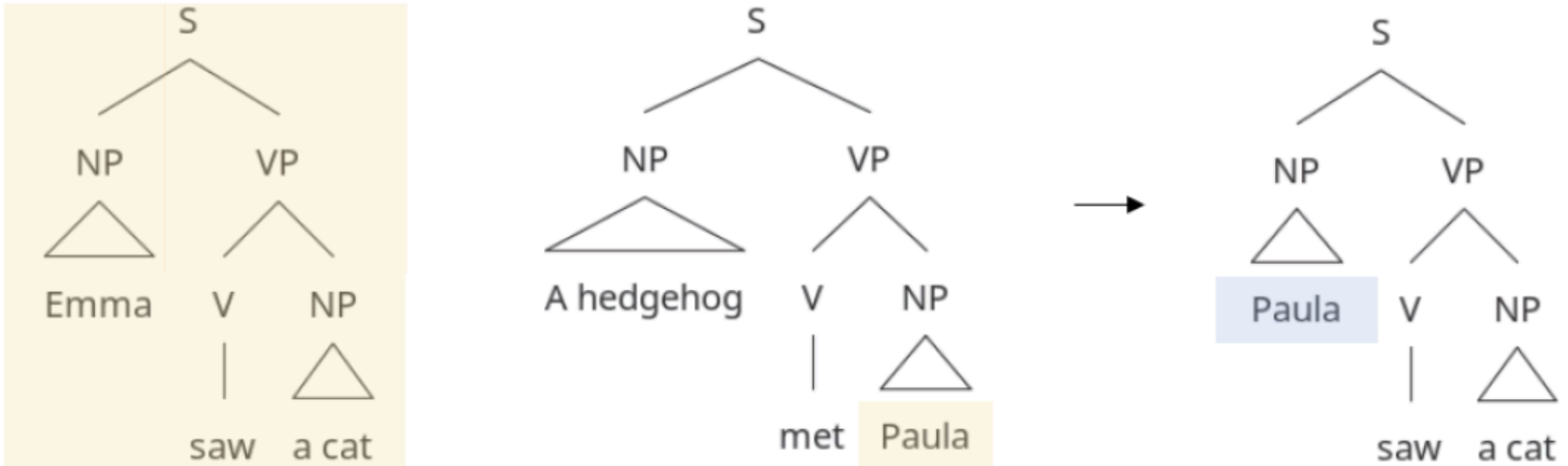
Can neural networks for language generalize compositionally?

- It depends!
- For models trained ("pretrained") to predict missing words:
  - Is generalization measured on the original objective or on transfer to a new task?
- What is meant by compositionality? Distinguishing lexical and structural generalization

Lexical generalization: a familiar word in a new context

In training set

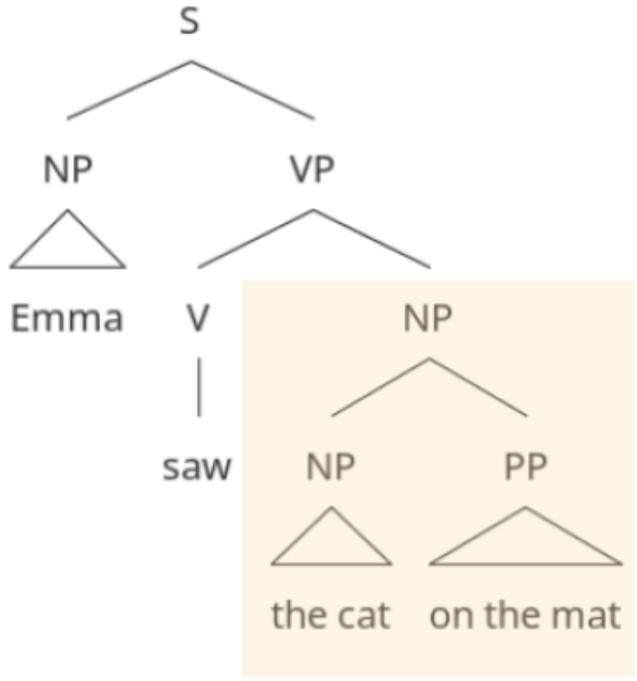
In generalization set

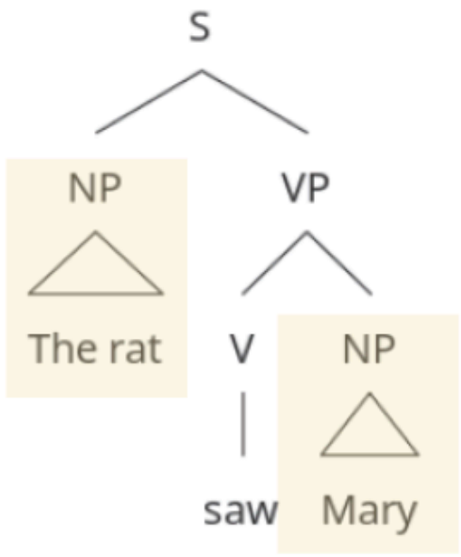


(Kim & Linzen, 2020, EMNLP)

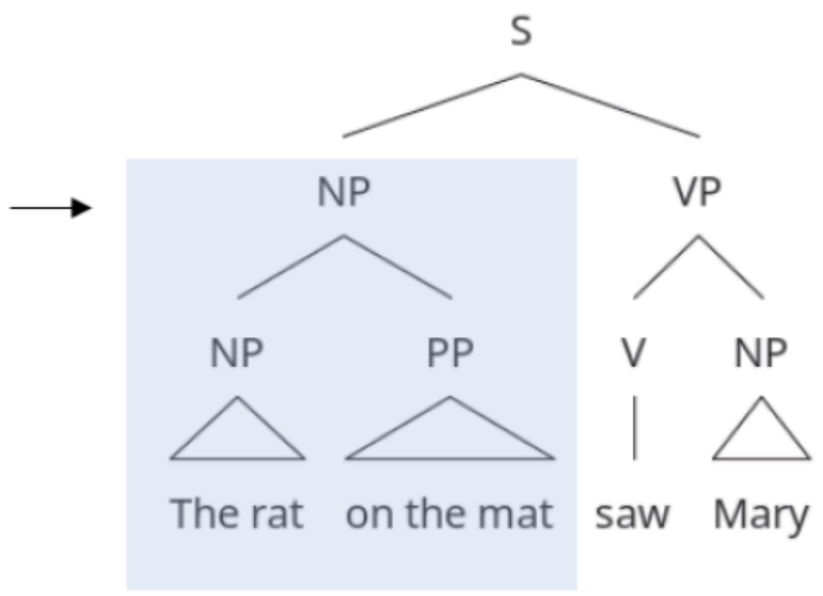
Structural generalization: a new combination of familiar structures

In training set





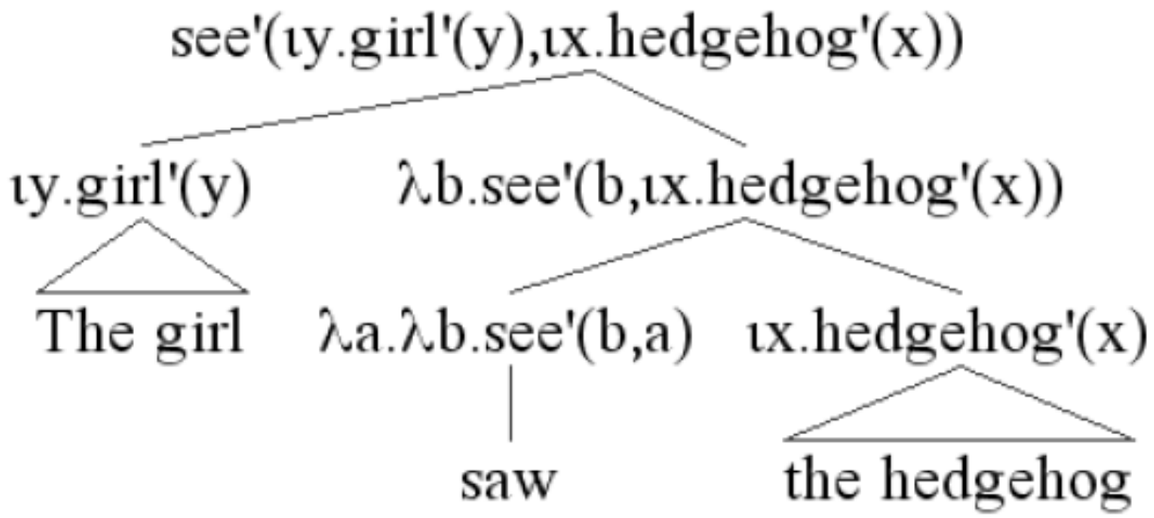
In generalization set



Evaluating compositionally in semantic parsing

Goal: convert a sentence into a normalized logical form (meaning representation); can then be used to query a database, send commands to a robot, etc

The girl saw the hedgehog =
The hedgehog was seen by the girl =
It was the hedgehog that the girl saw = …



(Kim & Linzen, 2020, EMNLP)

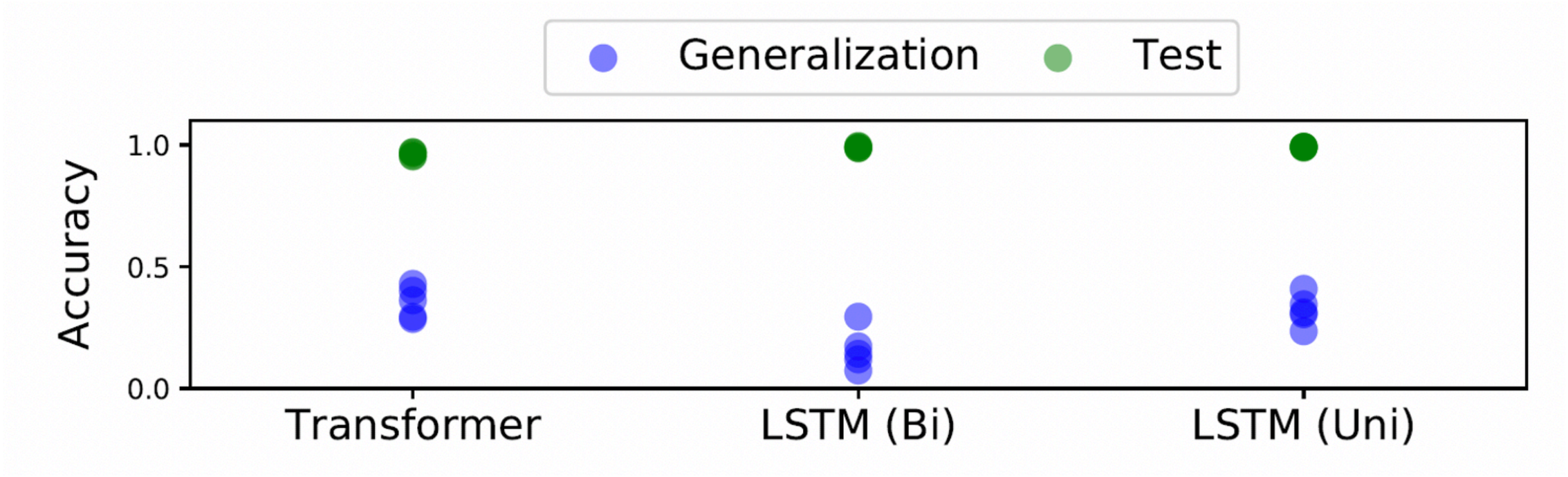
COGS: Benchmark for compositional generalization in semantic parsing

- Models trained “from scratch” on the training set and evaluated on the generalization set

	Case	Training	Generalization
Lexical generalization	Subject → Object (common noun)	*Subject* A **hedgehog** ate the cake.	*Object* The baby liked the **hedgehog**.
	Object → Subject (common noun)	*Object* Henry liked a **cockroach**.	*Subject* The **cockroach** ate the bat.
Structural generalization	Depth generalization: PP modifiers	*Depth 2* Ava saw the ball **in the bottle on the table**.	*Depth 3* Ava saw the ball **in the bottle on the table on the floor**.

(Kim & Linzen, 2020, *EMNLP*)

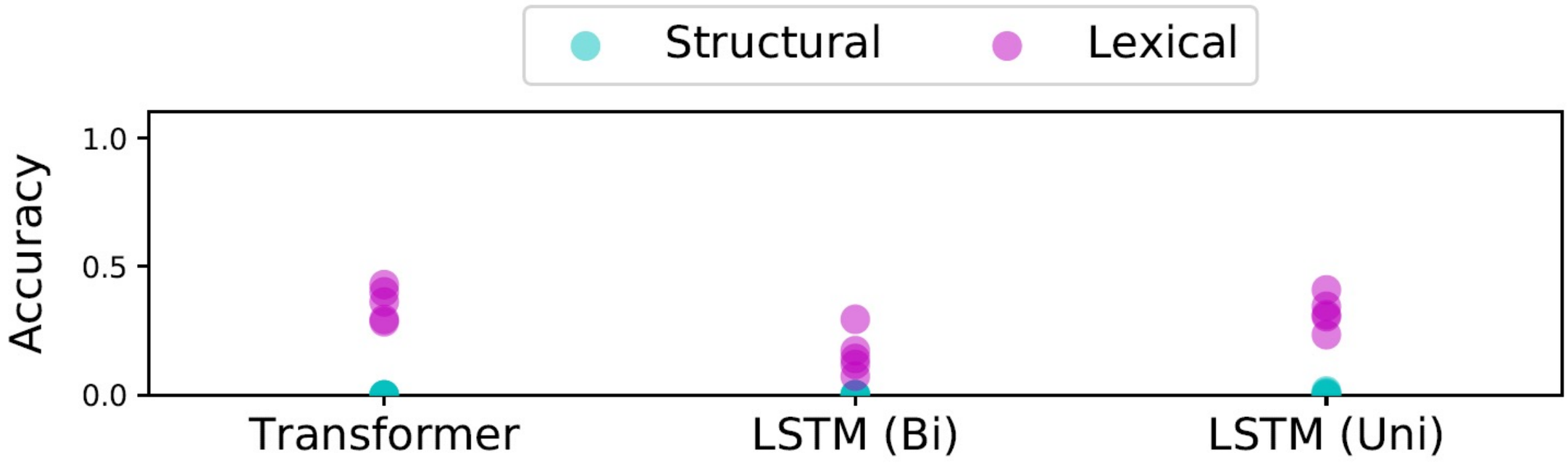
Results



(Exact match accuracy;
each dot represents a different random weight initialization)

Generalization is poor and varies across weight initializations

Results



Lexical generalization: a familiar word in a new context

Structural generalization: a new combination of familiar structures

(Kim & Linzen, 2020, EMNLP)

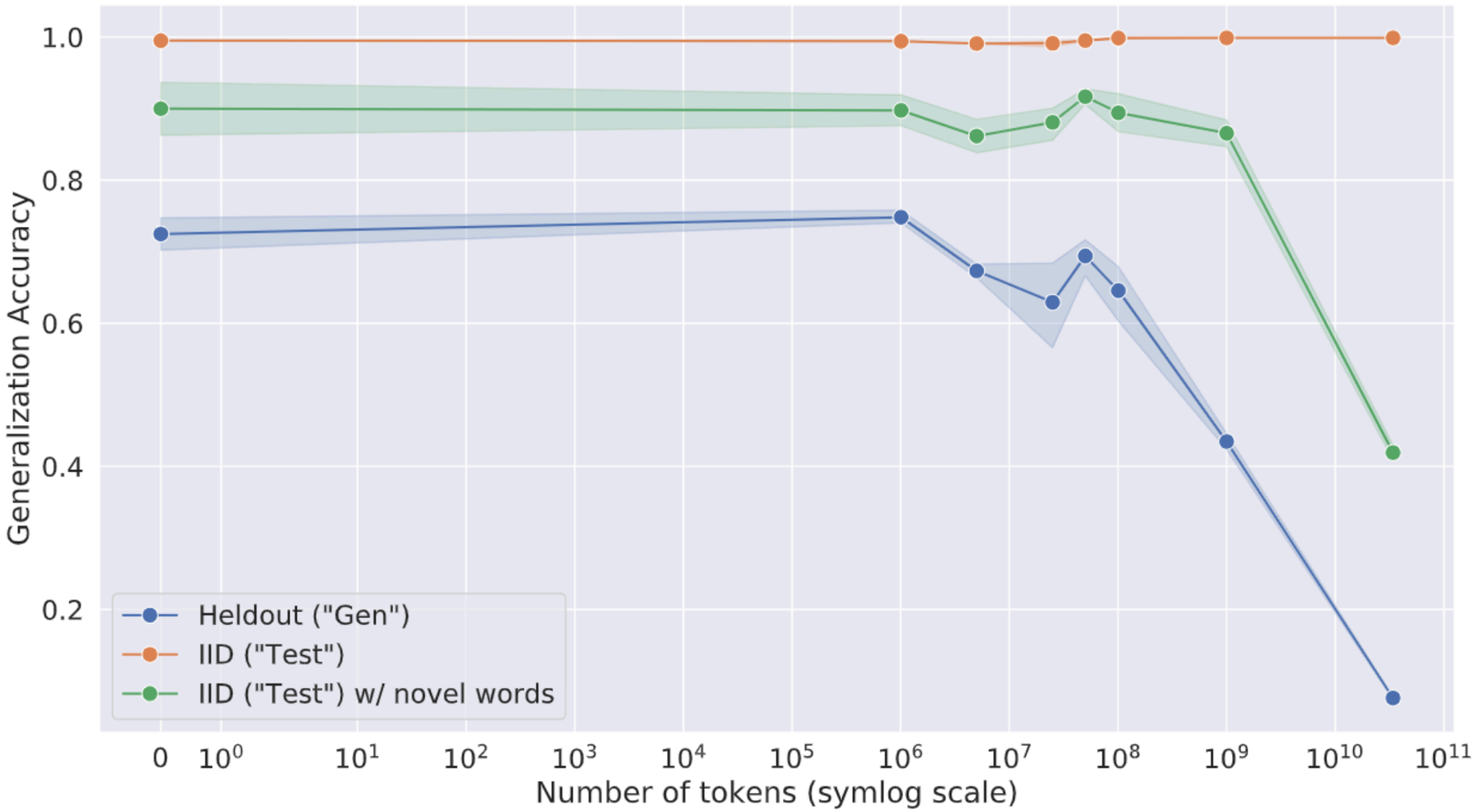
Measuring compositional generalization with a pretrained model is hard

- Is it plausible to assume that *hedgehog* never occurred in the object position in T5's training corpus?

Case		Training	Generalization
Lexical generalization	Subject → Object (common noun)	*Subject* A **hedgehog** ate the cake.	*Object* The baby liked the **hedgehog**.
	Object → Subject (common noun)	*Object* Henry liked a **cockroach**.	*Subject* The **cockroach** ate the bat.

(Najoung Kim, unpublished dissertation, 2021)

Pretraining might ***hurt*** compositional generalization, when words are truly novel



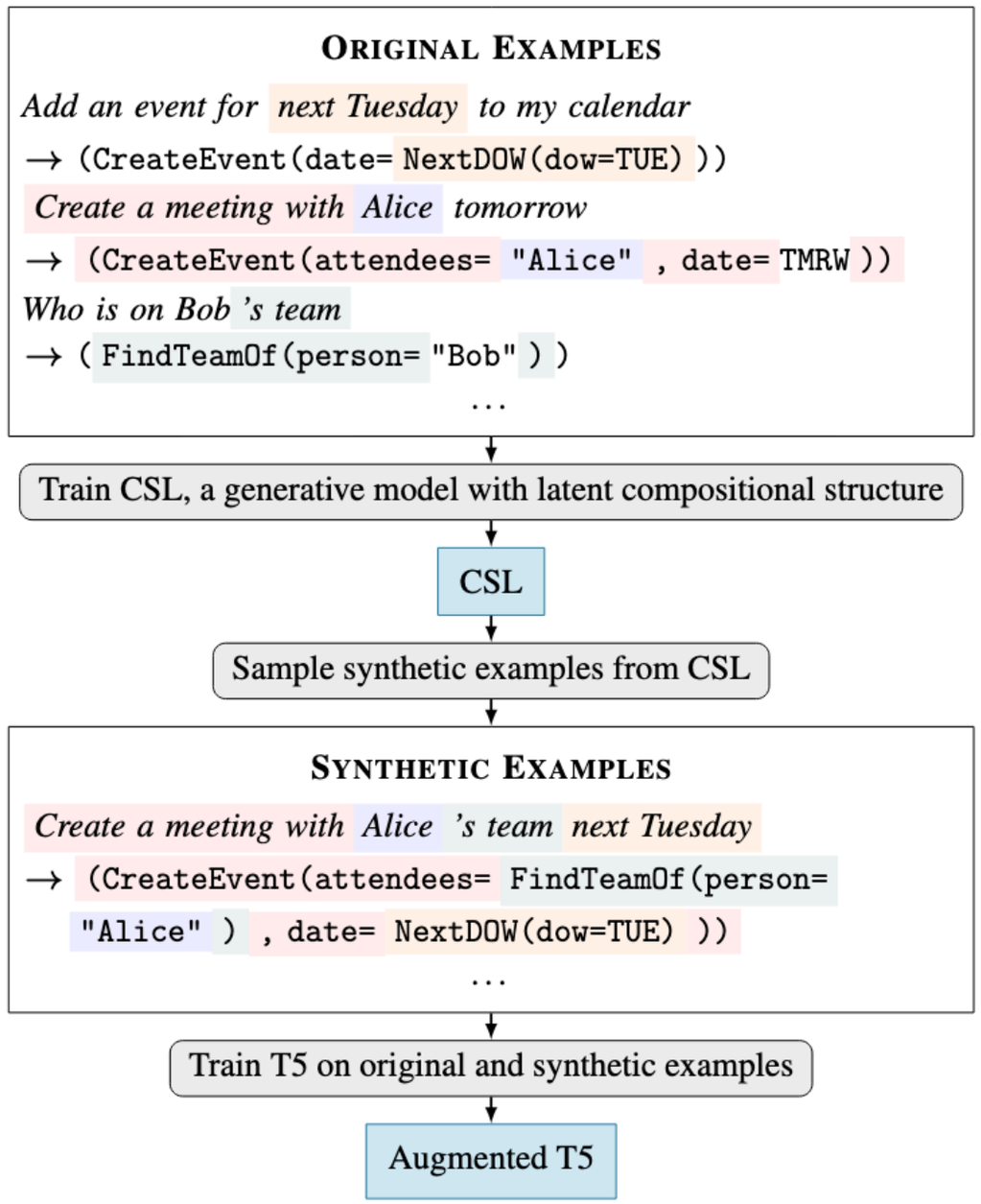
(Najoung Kim, unpublished dissertation, 2021)

COGS structural generalization is difficult across models

Model Class	Model	STRUCT Obj to Subj PP	STRUCT CP recursion	STRUCT PP recursion	LEX all 18 other types
seq2seq	BART	0	0	12	91
	BART+syn	0	5	8	93
	T5	0	0	9	97
	Kim and Linzen 2020	0	0	0	73
	Akyürek and Andreas 2021	0	0	1	96
	Zheng and Lapata 2022	0	12	39	99
	Conklin et al. 2021	0	0	0	88
	Csordás et al. 2021	0	0	0	95
	Qiu et al. 2021 *	100	100	100	100
structure-aware	Liu et al. 2021	93	100	99	99
	Weißenhorn et al. 2022	78	100	99	100

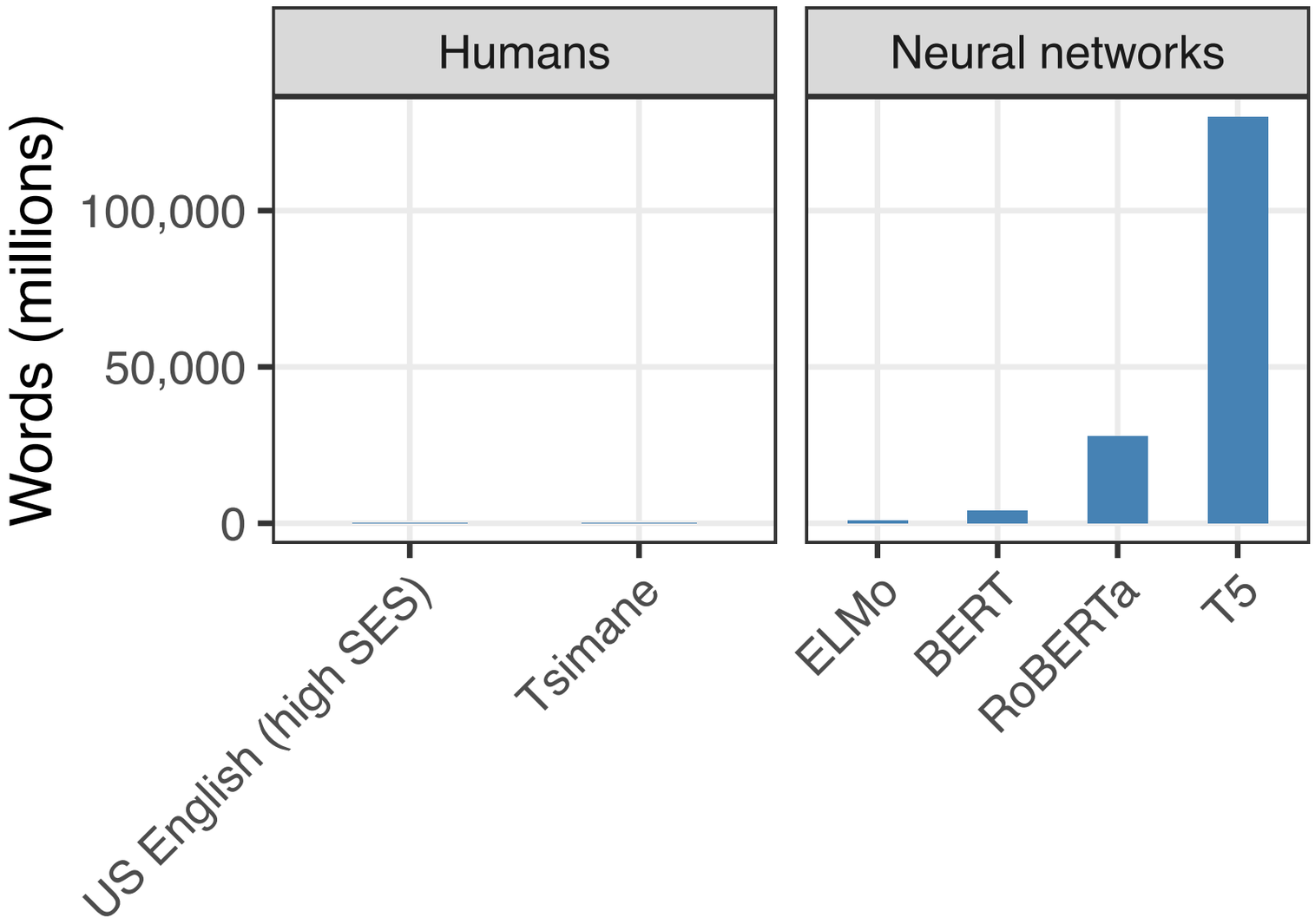
(Weißenhorn, Yao, Donatelli & Koller, 2022, **SEM, Compositional Generalization Requires Compositional Parsers*)

Data augmentation using a symbolic induced grammar: the robustness of a neural parser combined with a symbolic crutch



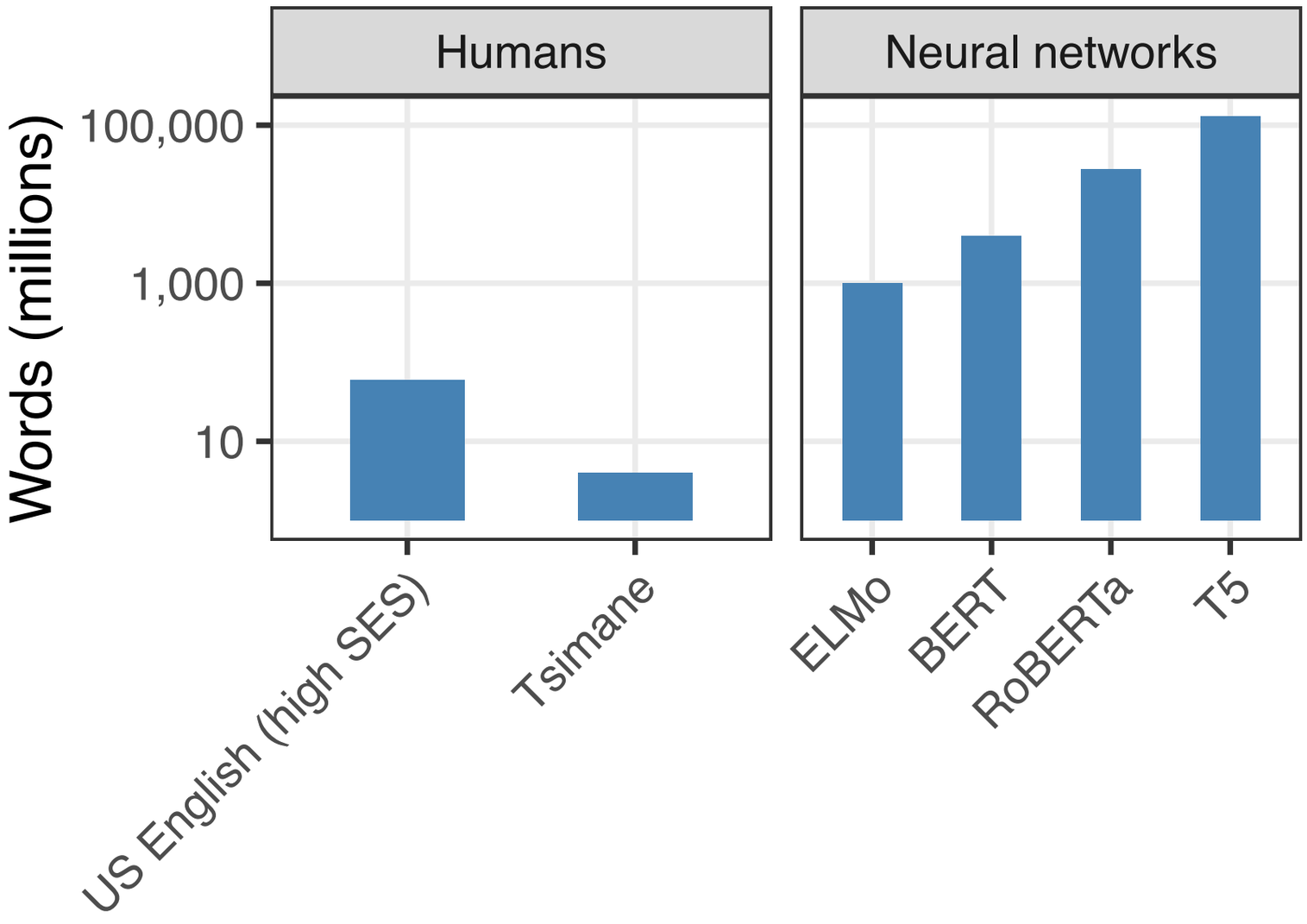
(Qiu et al, 2022, *NAACL*)

Does this matter? Sample efficiency in pretraining



(Linzen, 2020, *ACL*)

Does this matter? Sample efficiency in pretraining



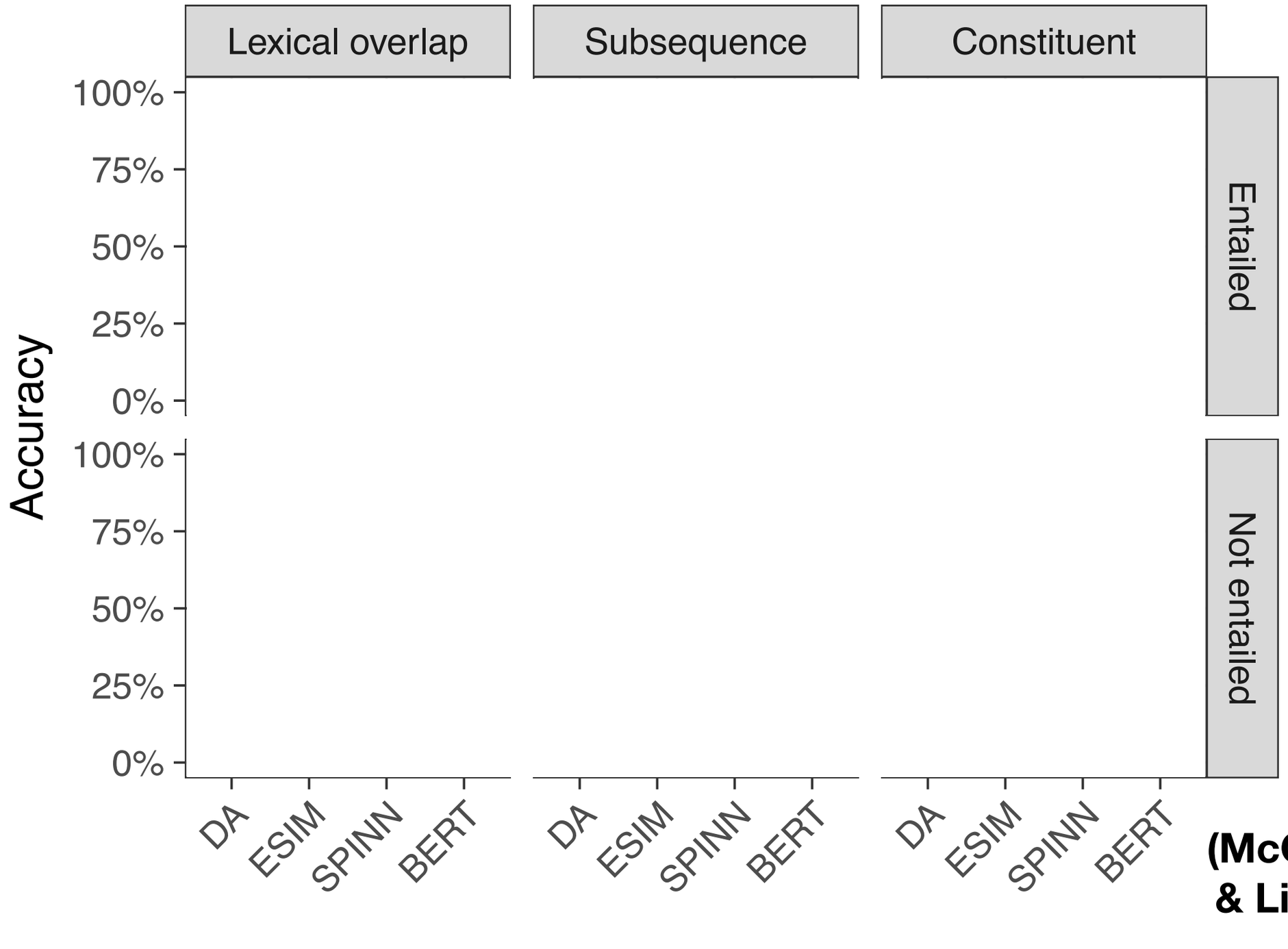
(Linzen, 2020, *ACL*)

Does this matter?

- For many semantic tasks (like semantic parsing) that are useful in practice, we often have a small fine-tuning dataset

- If we are able to give models a compositional inductive bias, they will generalize appropriately from a small amount of data (in pretraining or in transfer)

HANS: Heuristic Analysis of NLI Systems



Heuristic makes correct prediction

Heuristic makes incorrect prediction

(McCoy, Pavlick & Linzen, 2019, ACL)

HANS: Case-by-case results (where the heuristic makes an incorrect prediction)

Heuristic	Subcase	DA	ESIM	SPINN	BERT
Lexical	Subject-object swap	0.00	0.00	0.03	0.00

BERT trained on MNLI always predicts that

The lawyer advised the judge

entails

The judge advised the lawyer

(McCoy, Pavlick & Linzen, 2019, ACL)

Conclusions

- Language models do much more than memorize, especially when evaluated on their training objective
- They can parse ***colorless green idea*** sentences and generate novel text: compositional generalization!
- Caveat: even *n*-gram models with smoothing do more than memorize

Conclusions

- Transfer to new tasks (e.g. semantic parsing) is less impressive

- There is a fundamental distinction between lexical generalization and structural generalization

- Is structural generalization even useful if you've seen billions of English sentences, with basically all possible structures?

- What about "in-context learning" / "language models are few shot learners"? Probably more like fine-tuning than tests on the original objective