

What we want to do is (to) glimpse into how LMs learn language

Youn-Gyu Park (The University of Texas at Austin)

Introduction

English pseudoclefts refer to sentences where a relativizer (usually a *wh*-word) introduces a clause which is followed by a copula and a pivot (Collins 1991; den Dikken 2006; Flickinger and Wasow 2013).

- (1) a. I want to eat an apple.
b. **What** I want to eat **is** an apple.

Meanwhile, a VP can also be realized in the pivot position. In such cases, it can take two possible forms (den Dikken 2006).

- (2) a. **What** I want to do **is eat** an apple.
b. **What** I want to do **is to eat** an apple.

[bare-inf pivot]
[to-inf pivot]

Interestingly, the distribution of *bare*-infinitives (*bare*-inf) is **more restrictive** than *to*-infinitives (*to*-inf); Unlike *to*-infs, it is ungrammatical for a *bare*-inf to follow a copula in most cases (Biber et al. 2002; Huddleston and Pullum 2002), thus *to*-infs gain the upper hand in terms of the frequency effect (cf., Goldberg 2006).

- (3) a. *My goal **is finish** my Ph.D.
b. My goal **is to finish** my Ph.D.

Research question

Then, what would happen if LMs are trained on a dataset that lacks *bare*-inf pseudoclefts?

In other words, under the frequency effect expected, do LMs require a **direct evidence** for idiosyncratic constructions, such as *bare*-inf pseudoclefts, in language acquisition?

Linguistic backgrounds

bare-infs vs. *to*-infs

In English, *bare*-infs can be licensed in restricted syntactic environments in which *to*-ints are usually allowed to appear.

- (4) a. I want *(*to*) sleep.
b. John helped me (*to*) move out.

A copula does not permit a *bare*-inf complement except in pseudoclefts. In such cases as well, it can be substituted by its *to*-inf counterpart.

- (5) a. My goal **is** *(*to*) finish my Ph.D. in 5 years.
b. **What** I want to do **is** (*to*) finish my Ph.D.

Pivots in (semi-)pseudoclefts

Akin to *wh*-words in pseudoclefts, a limited set of syntactic units, such as *all*, *the thing*, can be used instead. Such constructions are called semi-pseudoclefts.

In (semi-)pseudoclefts, a pivot can be realized in various categories (den Dikken 2006; Flickinger and Wasow 2013; Park and Kim 2023).

- (6) a. **All** I ate for dinner **was a salad**. (Tellings 2020: 1)
b. **All** I did **was sit** around. (Flickinger and Wasow 2013: 11)
c. **All** that one has to do **is to start** training earlier. (Kay 2013: 34)
d. **All** I can remember **is my parents had nothing**. (Park and Kim 2023: 577)

An experiment

Overview

As a case study to understand how LMs acquire language, this study performed a pre-training experiment using OPT-125M architecture and BabyLM dataset.

Hypotheses

Given the linguistic properties, this study sets the following hypotheses:

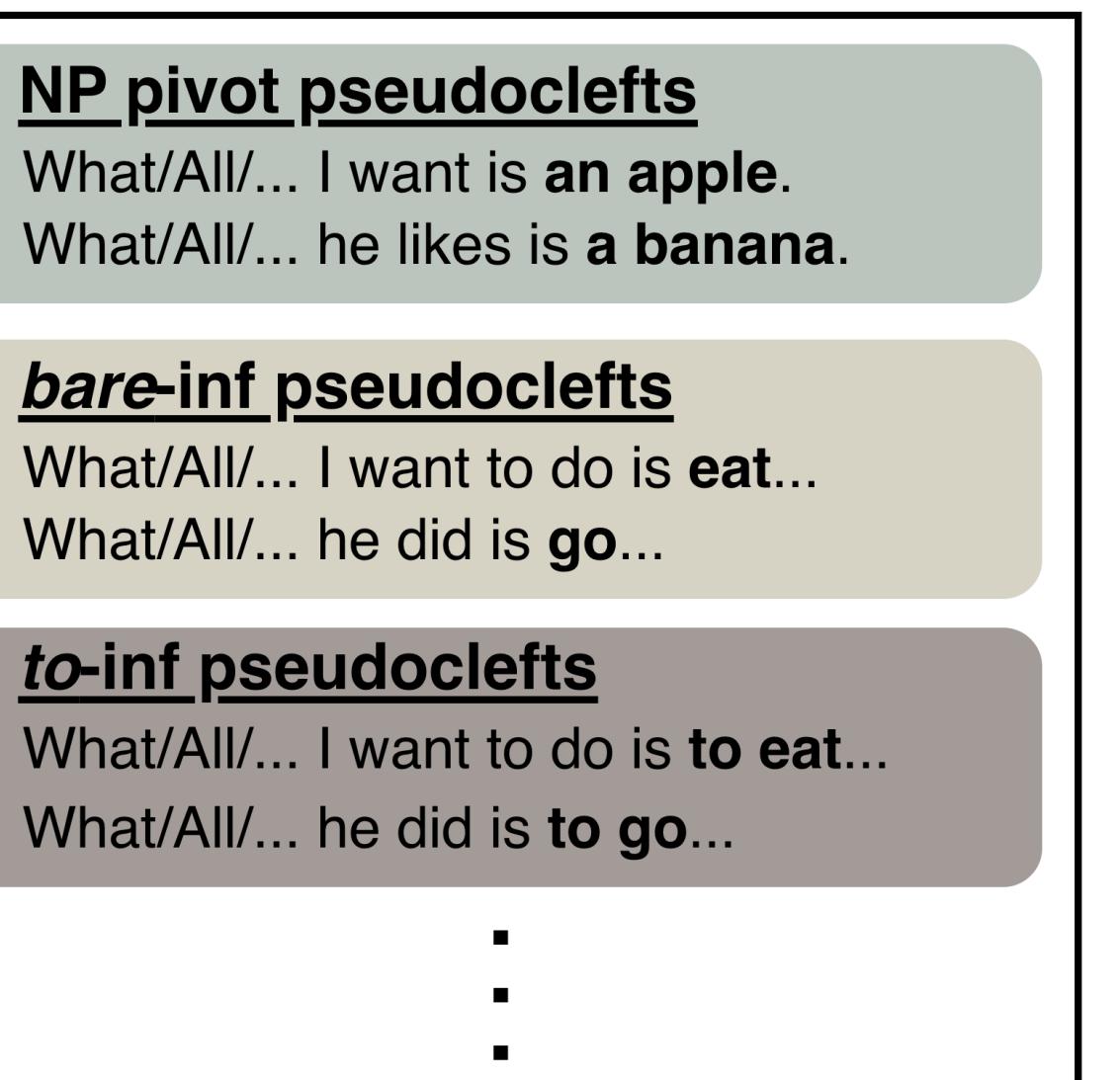
- H₀: LMs will fail to acquire *bare*-inf pseudoclefts if the training set lacks the direct evidence.
- H₁: LMs still acquire *bare*-inf pseudoclefts although the training set lacks the direct evidence.

Methodology

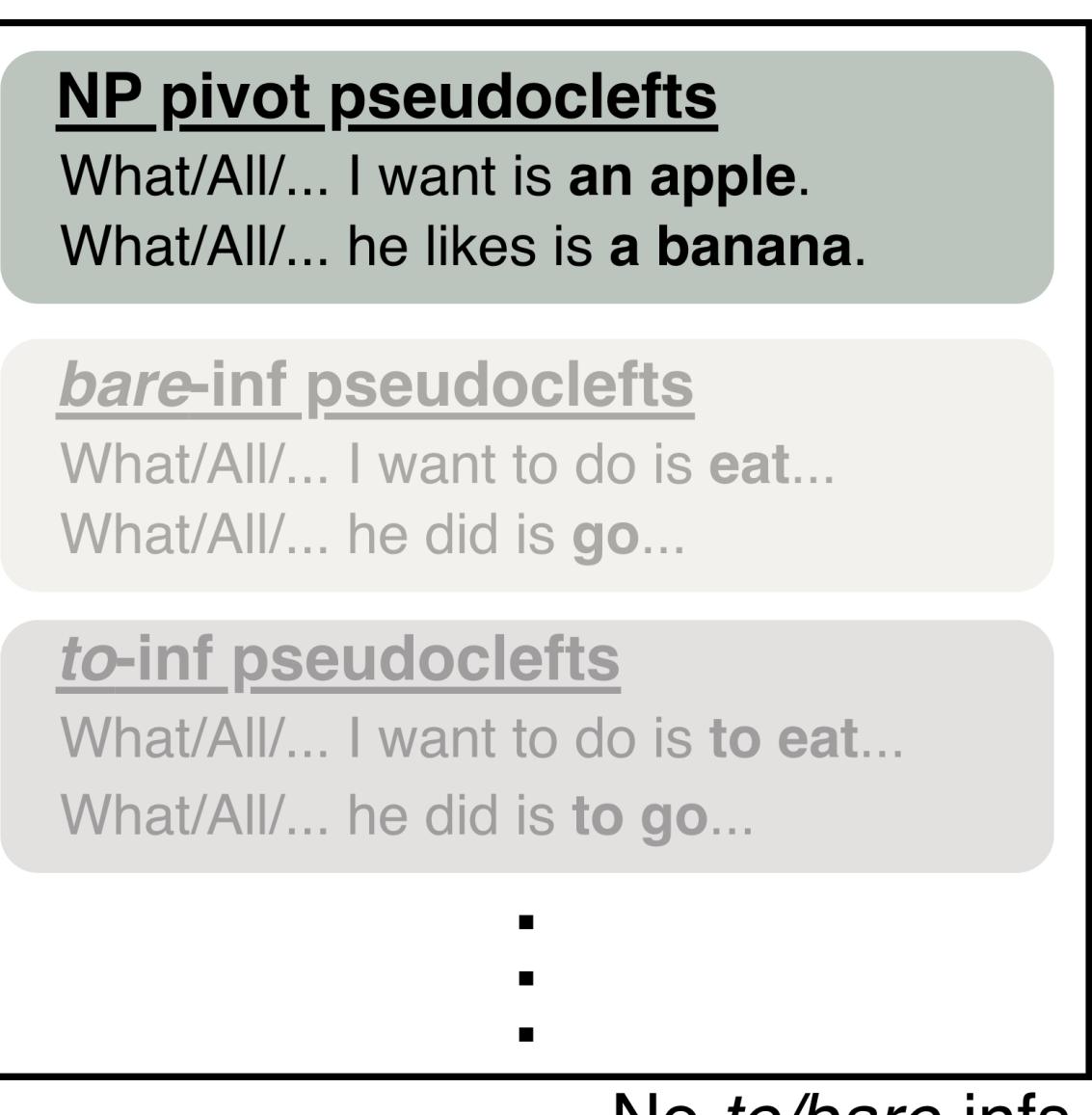
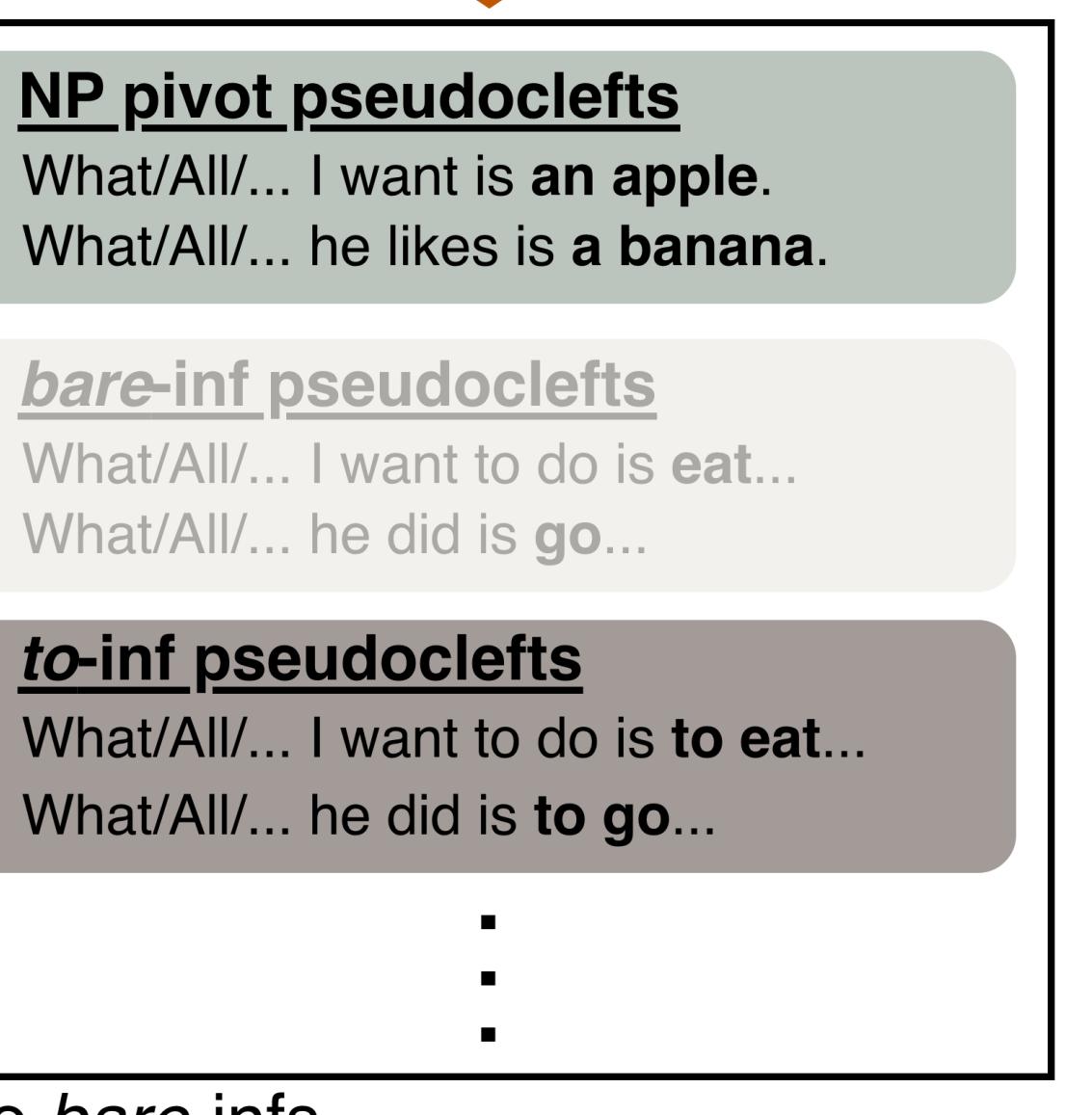
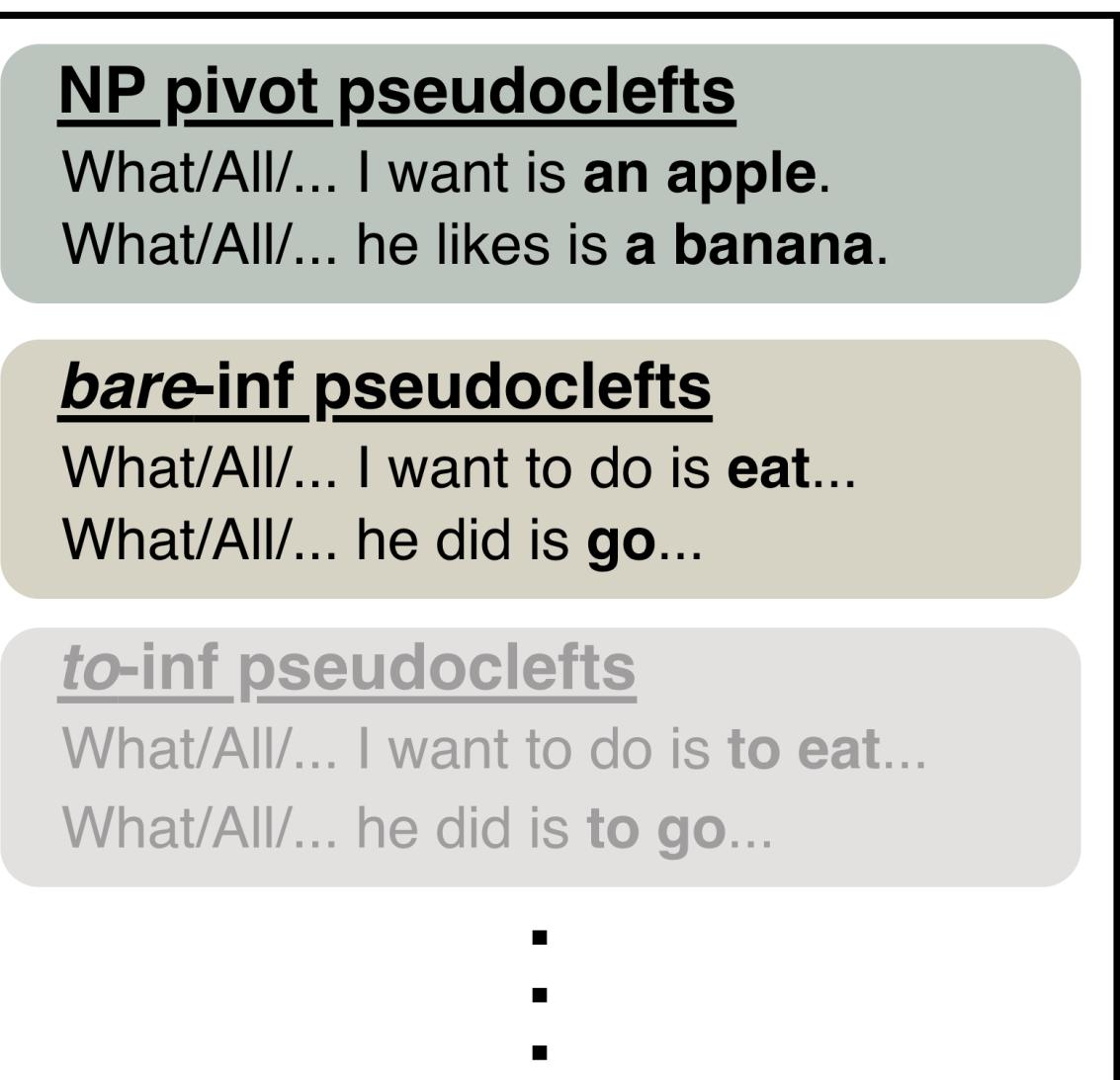
Dataset modification

Using re and spaCy packages, the vanilla BabyLM dataset was modified by getting rid of i) all *to*-inf pseudocleft data, ii) *bare*-inf pseudocleft data, and iii) both *to*- and *bare*-inf pseudocleft data.

Vanilla



No to-infs



No bare-infs

No to/bare-infs

Figure 1: Data modification using BabyLM dataset

Pre-training and evaluation

OPT-125M architecture was pre-trained on four different datasets for 10 epochs. Two evaluation tasks were conducted: One is the forced-choice task based on the perplexity value between pseudocleft minimal pairs (Task 1), and the other is based on the average perplexity value on the test dataset, using minicons package (Task 2). The test dataset contains 50 grammatical tokens of *bare*- and *to*-inf pseudocleft minimal pairs.

Variables and analysis

This experiment sets the factors train data (see Fig. 1), v_type (*to*-inf/*bare*-inf), and epoch (1-10 epochs) as its variables.

A series of random mixed-effect regression analyses for the v_type and epoch factors with respect to the train data factor was conducted in order to see the statistical significance.

Results

Task 1: Forced-choice

The results indicate that *to*-infs are generally more preferred across all the four models. Nonetheless, in the no *to*-inf model, the chance of *bare*-inf selection becomes higher at the later stage of training.

Task 2: Perplexity

A mixed-effects model with random intercept model revealed a significant interaction between type ($p < .001$) and epoch (epoch = 2, $p < .01$; epoch > 3 , $p < .001$). Although performance improved across epochs in both conditions, the *to* *inf* condition consistently showed lower z-scores than the *bare* *inf* condition at every epoch. Importantly, this difference decreased over time, indicating partial convergence between conditions, but did not disappear entirely by the final epoch.

Post-hoc analyses

Vanilla and No bare-inf: The two factors v_type and epoch, and their interaction are statistically significant ($p < .001$).

No to-inf: The two factors v_type and epoch are statistically significant ($p < .001$) as well as their interaction ($p < .01$), except for the main effect of epoch at the second epoch ($p > .01$). The main effect of epoch at the third epoch is slightly weaker ($p < .01$) than at the other epochs.

No to/bare-inf: The two factors v_type and epoch are statistically significant ($p < .001$) as well as their interaction ($p < .01$), except for the main effect of epoch at the second epoch ($p > .01$).

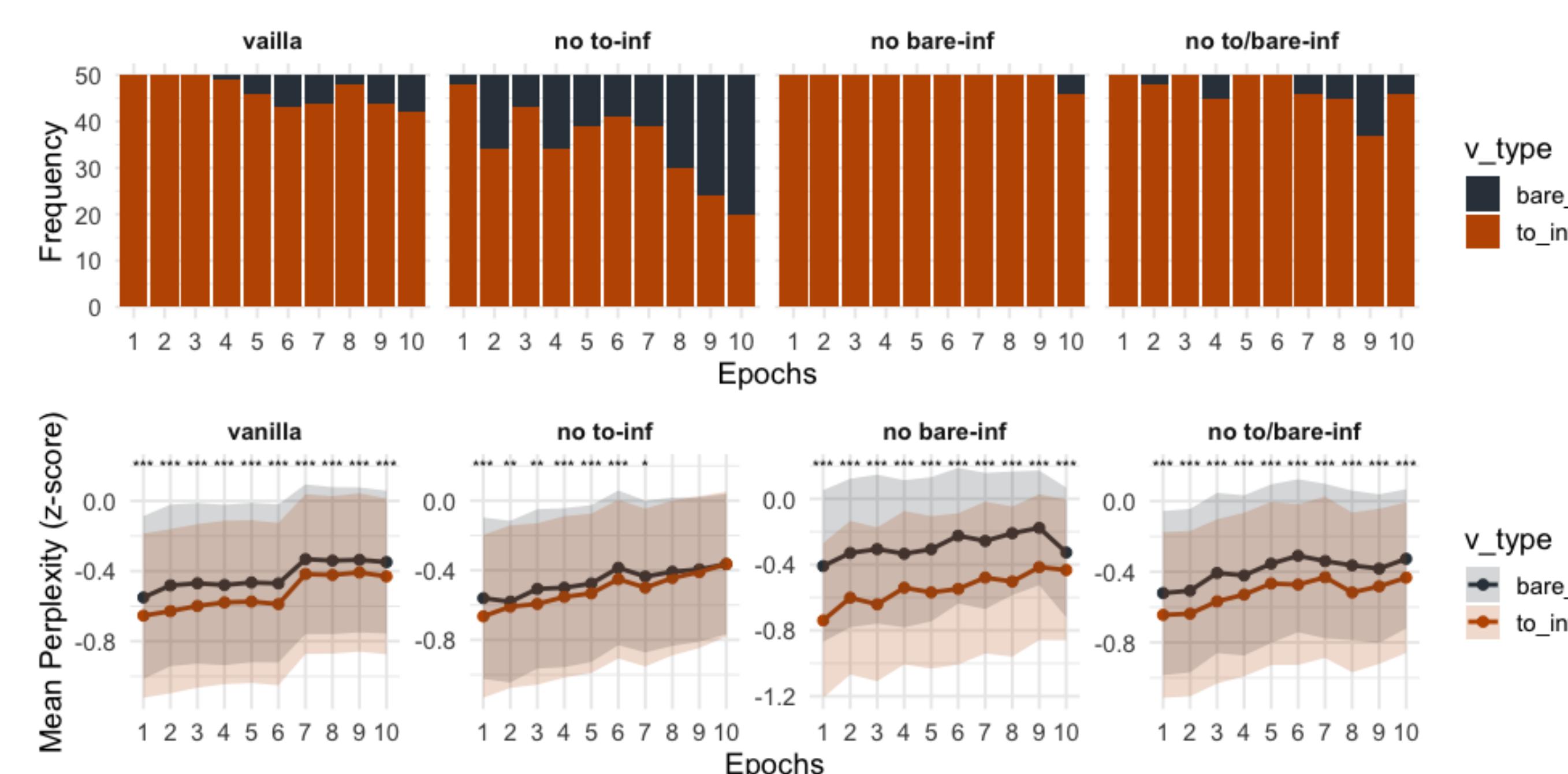


Figure 2: Results: Forced-choice (top) and mean perplexity (bottom)

Discussion & Implication

Direct evidence

The results imply that it is not easy for a model to learn an idiosyncratic syntactic construction under the lack of direct evidence. This can be seen as the effect of frequency effect, in which *bare*-infs appear in a limited syntactic environment (e.g., imperatives, some (di-)transitives), especially with the canonical argument structure (ARG-ST) of copula (see (5)).

Frequency effect

Let alone the other three models, even in the no *to*-inf model, which only saw the direct evidence for *bare*-inf pseudoclefts, the perplexity for *to*-infs were significantly lower than for *bare*-infs until the mid-point of pre-training.

This study claims that it is because of the syntactic ARG-ST of English copula, taking *to*-infs in most cases instead of *bare*-infs; The ARG-ST of canonical copula constructions do concern the acquisition process, showing that syntactic information of related constructions does matter in language acquisition of LMs.

Selected references

- Biber, Douglas, Susan Conrad and Geoffrey Leech. 2002. *Longman Student Grammar of Spoken and Written English*. Essex: Longman.
Collins, Peter. 1991. *Cleft and Pseudo-Cleft Constructions in English*. New York: Routledge.
den Dikken, Marcel. 2006. Specifical copular sentences and pseudoclefts. In Martin Everaert and Henk van Riemsdijk (eds.), *The Blackwell Companion to Syntax*, Malden: Blackwell.
Flickinger, Dan and Thomas Wasow. 2013. *The Core and the Periphery: Standing on the Shoulders of Ivan A. Sag* chap. A corpus-driven analysis of the *do-be* constructions. Cambridge: CSLI publications.
Goldberg, Adele. 2006. *Constructions at Work: The Nature of Generalization in Language*. New York: Oxford University Press.
Huddleston, Rodney and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.
Kay, Paul. 2013. The limit of (Construction) Grammar. In Thomas Hoffmann and Graeme Trousdale (eds.), *The Oxford Handbook of Construction Grammar*, 32-48. Oxford: University Press.
Park, Youn-Gyu and Jong-Bok Kim. 2023. All-cleft constructions in English: A corpus-based approach. *Korean Journal of English Language and Linguistics* 23: 571-586.
Telling, Jos. 2020. An analysis of all-clefts. *Glossa: A journal of general linguistics* 5(1): 1-25.