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

Youn-Gyu Park (The University of Texas at Austin) youngyu.park@utexas.edu

Introduction

English pseudoclefts refer to sentences in which 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 which case it takes two possible forms (den Dikken 2006).

(2) a. **What** I want to do **is eat** an apple. [*bare*-inf pivot]
b. **What** I want to do **is to eat** an apple. [*to*-inf pivot]

Interestingly, the distribution of *bare*-infinitives (*bare*-inf) is **more restrictive** than that of *to*-infinitives (*to*-inf); Unlike *to*-infs, a *bare*-inf is ungrammatical following a copula in most cases (Biber et al. 2002; Huddleston and Pullum 2002), giving *to*-infs an advantage 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 were trained on a dataset that lacks *bare*-inf pseudoclefts?
In other words, under the expected frequency effect, do LMs require **direct evidence** in order to learn idiosyncratic constructions?

Linguistic backgrounds

bare-infs vs. *to*-infs

In English, *bare*-infs can be licensed in restricted syntactic environments where *to*-ints are typically allowed to appear (Biber et al. 2002; Huddleston and Pullum 2002).

(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. Even in these cases, it can be substituted by its *to*-inf counterpart (Flickinger and Wasow 2013).

(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 (e.g,. *all*, *the thing*, ...) can be used instead. Constructions of this type are called semi-pseudoclefts. In (semi-)pseudoclefts, the pivot can be realized in various categories (cf., den Dikken 2006; 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

As a case study for understanding how LMs acquire language, this study conducted a pre-training experiment using the OPT-125M architecture on the BabyLM dataset (Warstadt et al. 2023).

Hypotheses

Given these linguistic properties, this study formulates the following hypotheses:

- **H_0**: LMs will fail to acquire *bare*-inf pseudoclefts if the training set lacks direct evidence.
- **H_1**: LMs can still acquire *bare*-inf pseudoclefts even if the training set lacks direct evidence.

Methodology

Dataset modification

Using the `re` and `spaCy` packages, the `vanilla` BabyLM dataset was modified by removing: (i) all *to*-inf pseudocleft data, (ii) all *bare*-inf pseudocleft data, and (iii) both *to*- and *bare*-inf pseudocleft data.

Pre-training and evaluation

OPT-125M architecture was pre-trained on four different datasets for 10 epochs. Two evaluation tasks were then conducted using a test dataset consisting of 50 grammatical tokens of *bare*- and *to*-inf pseudocleft minimal pairs.

Task 1: Within-model surprisal comparison between *to*- and *bare*-inf pseudoclefts

- A random mixed-effects regression analysis was conducted for each `model` (see Fig. 1)
- Variables: `v_type` (*to*-inf/*bare*-inf), and epoch (1-10 epochs)

Task 2: Across-model comparison for the surprisal difference between *to*- and *bare*-inf pseudoclefts
For each models on the two types of VPs ($S_M(VP)$):

- **Comparison 1**. $S_{NTB}(bare) - S_{NTB}(to)$ vs. $S_V(bare) - S_V(to)$:
  - If $[S_{NTB}(bare) - S_{NTB}(to) > S_V(bare) - S_V(to)]$, then models are more surprised by bare-inf pseudoclefts than by to-inf pseudoclefts, even though both lack direct evidence.
  - This indicates that models require **more direct evidence** to acquire idiosyncratic constructions.
- **Comparison 2**. $S_{NB}(bare) - S_V(bare)$ vs. $S_{NT}(to) - S_V(to)$:
  - If $[S_{NB}(bare) - S_V(bare) > S_{NT}(to) - S_V(to)]$, then models are more surprised to see *bare*-inf pseudoclefts without direct evidence than *to*-inf pseudoclefts.
  - This suggests that the acquisition of *bare*-inf pseudoclefts requires more direct evidence than that of *to*-inf pseudoclefts, demonstrating **the frequency effect**.

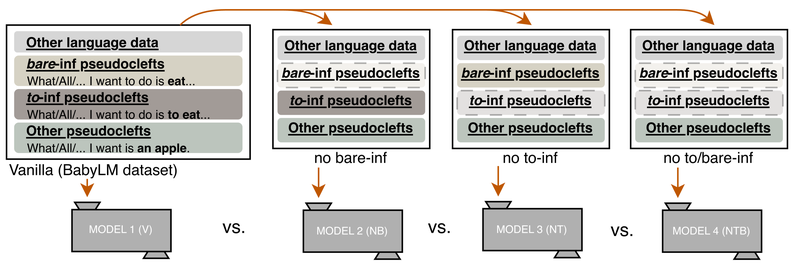

Figure 1: Data modification using BabyLM dataset

Results

Task 1: Within-model comparison

The results revealed a significant interaction between `type` and epoch ($p < .001$). Although performance improved across epochs in both conditions, the `to_inf` condition consistently exhibited lower *z*-scores than the `bare_inf` condition at every epoch, except for the 10th epoch in the `no_to_inf` model. Importantly, this difference decreased over time, indicating partial convergence between the conditions, but it did not disappear entirely by the final epoch.

Post-hoc analyses

Vanilla condition: Both factors, `v_type` ($p < .01$) and epoch ($p < .001$), were statistically significant, but their interaction was not ($p = .34$).

No *to*-inf condition: The `v_type` factor was not significant ($p = .348$), although the epoch factor and its interaction with `v_type` were significant ($p < .001$).

No *bare*-inf condition: Both factors, `v_type` and epoch, as well as their interaction, were statistically significant ($p < .001$).

No *to*/*bare*-inf condition: Both factors, `v_type` ($p < .01$) and epoch, along with their interaction, were statistically significant ($p < .001$).

Task 2: Across-model comparison

Comparison 1: The results exhibit a mixed pattern, consistent with the *z*-scored mean surprisal results shown in Fig. 2.

Comparison 2: Although both S_{NB} and S_{NT} lacked direct evidence for bare- and to-inf pseudoclefts, respectively, the results indicate that bare-inf pseudoclefts elicit higher surprisal.

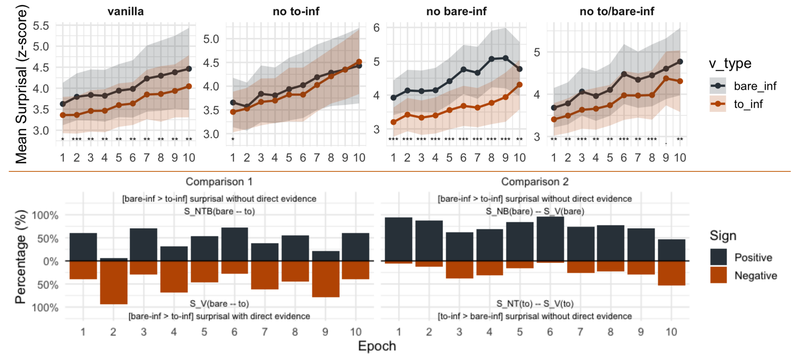

Figure 2: Results: *z*-scored mean surprisal (Task 1, top) and *z*-scored surprisal comparison (Task 2, bottom)

Discussion & Implication

Direct evidence and the Frequency effect

The results suggest that learning an idiosyncratic syntactic construction is difficult for a model in the absence of direct evidence. This difficulty can be understood as **a frequency effect** (cf., Goldberg 2006).

Although the `vanilla` model was trained on the dataset containing direct evidence for *bare*-inf pseudoclefts as well, only the `no_to-inf` model was able to learn this idiosyncratic syntactic form properly.

The reason is that it was the only model exposed only (or at least more often) to direct evidence of *bare*-inf pseudoclefts than *to*-inf pseudoclefts, unlike the other models, where a set of relativizers (see (6)) are followed by idiosyncratic *do-be*-VP[*bare*] string.

A Construction Grammar perspective

This study argues that the results can be explained from a Construction Grammar (CxG) perspective (Goldberg 2006, *inter alia*).

In English, *bare*-infs occur in a restricted set of syntactic environments (e.g., imperatives, some (di-)transitives), particularly with the canonical argument structure (ARG-ST) of the copula, whereas *to*-infs appear more frequently in the same positions (see (5)). Such constraints on the ARG-ST of the English copula influence overall language acquisition.

The ARG-ST of canonical copula constructions and the higher collocation frequency of *to*-infs thus play a role in the acquisition process, indicating that the syntactic information of related constructions matters for the language acquisition of LMs.

Nonetheless, when certain priming occurs (e.g., relativizers), the probability of encountering the idiosyncratic ARG-ST increases, facilitating the LM's acquisition of syntactic constructions with special ARG-STs.

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. Specificational 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.
Tellings, Jos. 2020. An analysis of *all*-clefts. *Glossa: A journal of general linguistics* 5(1): 1–25.
Warstadt, Alex, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang and Samuel R. Bowman. 2023. The BabyLM challenge: Sample-efficient language model pre-training guided by developmental science. In Richard Diehl Martinez, Hope McGovern, Zebulon Goriely, Christopher Davis, Andrew Caines, Paula Buttery and Lisa Beinborn (eds.), *Proceedings of the babylm challenge at the 27th conference on computational natural language learning*, 1–14. Singapore: Association for Computational Linguistics. URL aclanthology.org.