Learning ‘before’ before ‘after’: modeling temporal connective acquisition with and without a language of thought

Neil Rathi (rathi@stanford.edu)
Department of Mathematics, Stanford University
Stanford, CA, 94305

Abstract

Language-of-thought (LOT) models have seen success in modeling meaning acquisition, particularly in the domain of quantifier acquisition. Here, I apply LOT models to the nuanced question of temporal connective acquisition, extending a model from (Gorenstein, Zhang, & Piantadosi, 2020). I show that when introduced to real-world learning constraints, the model fails to reflect human-like acquisition trajectories. I argue that this comes out of an artifact of the formulation of the posterior probability, but *not* the LOT prior. To this end, I provide evidence from a neural network language model, showing that this model—which does not explictly encode semantics to the extent that LOT models do—deviates significantly from patterns of human learning.

Keywords: statistical learning; language-of-thought; acquisition; semantics

Introduction

You read the abstract *before* the introduction. I worked on the paper *until* I submitted it. We regularly have to describe events that are ordered in time. Cross-linguistically, language users utilize **temporal connectives** like ‘before’ and ‘until’ to express relative temporal relations (von Fintel & Matthewson, 2008). Despite this importance, however, temporal connectives provide a unique challenge to language acquisition: they are learned late, and they are learned in a highly specific order (Clark, 1971; Feagans, 1980).

In particular, children tend to first acquire words communicating sequential order, like ‘before’ and ‘after,’ followed by words which communicate durational co-occurence like ‘while,’ and finally words like ‘since’ and ‘until’ which communicate *both* order and duration (Feagans, 1980). Indeed, Feagans (1980) found that children learn ‘before’ and ‘after’ by the age of three, but only learn ‘while’ at the age of seven—and continue to struggle with ‘since’ and ‘until’ well after. And this phenomenon is not relegated to English; Winskel (2003) found that similar patterns exist in Thai and Lisu.

This poses an interesting problem for modeling connective acquisition—what mechanisms lead to this differentiated learning rate? The theoretical consensus is that learning is influenced by the (somewhat vague) notion of **semantic complexity** (Feagans, 1980). Less complex connectives (e.g. ‘before’) are easier to acquire than connectives expressing more complex temporal relations (e.g. ‘since’).

Much work on function word acquisition—particularly in the domain of quantifier learning—has found success using **language-of-thought** (LOT) models, which focus on training models from some basic semantic primitives (S. T. Piantadosi, 2011; S. T. Piantadosi, Tenenbaum, & Goodman, 2013; S. T. Piantadosi & Jacobs, 2016). Not only are these models empirically successful, but they also remain somewhat agnostic towards actual implementation, making minimal assumptions about the algorithmic processes behind production and processing.

To this end, Gorenstein et al. (2020) present a computational LOT model of connective acquisition, formalizing the notion of semantic complexity through string representations of lambda expressions. Empirically, their model is able to capture the same patterns of learning that humans exhibit.

However, the implementation of their model relies on several assumptions which do not reflect real-world learning scenarios. Particularly, the model assumes that learners have a uniform distribution over connectives. However, it is well-attested that this is not the case—production is in large part determined by pragmatic biases towards informativity (Goodman & Stuhlmüller, 2013; Goodman & Frank, 2016; Grice, 1975). And beyond this, because they are less applicable, more ‘specific’ connectives are seen less frequently during development.

In this work I enhance the model to accommodate these real-world learning considerations. Ultimately, however, I find that this is unable to accurately model human learning patterns. I argue that this is an artifact of the formulation of posterior production probabilities used in the model.

I then consider a separate, but related question: do neural network language models—trained without any explicit biases towards semantic complexity—acquire temporal connectives in human-like order? I show that a relatively simplistic language model *does* learn connectives in a differentiated order, but that this order is not consistent with human language acquisition. This sheds some light on the benefits of using LOT models: the kind of semantic complexity that shapes connective learning cannot be induced simply through training on large amounts of data, but rather requires some kind of inductive bias.

In what remains, I first give an overview of Gorenstein et al. (2020)’s model of connective acquisition. I then detail how S. T. Piantadosi et al. (2013)’s model of pragmatic production can be used to augment the Gorenstein et al. (2020) model, and show that with these modifications, standard pat-

Word	Target Meaning
before	$a_1 < b_1$
after	$b_1 < a_2$
since	$(a_1 < t) \wedge (t \leq a_2) \wedge (a_1 \leq b_2) \wedge (b_1 < t)$
until	$(a_1 \leq t) \wedge (t < a_2) \wedge (b_1 \leq a_2) \wedge (t < b_2)$
while	$(b_1 < a_2) \wedge (a_1 < b_2)$

Table 1: Target meanings for each temporal connective, as defined by Gorenstein et al. (2020). **Notes:** 'before' and 'after' do not have reciprocal meaning (i.e. '*A* before *B*' $\not\Rightarrow$ '*B* after *A*'); this is consistent with classical semantic analyses (Anscombe, 1964; Beaver & Condoravdi, 2003). We analyze 'since' and 'until' as specific cases of 'before' and 'after,' respectively, as in Kamp (1968)'s well-accepted work on the subject. Moens (1987) argues that 'while' describes temporal *overlap* of events *A* and *B*, rather than *containment*, and our target meaning reflects this.

terns of learning break down. Finally, I illustrate an example of connective learning in a small neural language model.

Basic Modeling Principles

Here, I briefly describe the learning model of Gorenstein et al. (2020). This model builds on previous models of quanitifier acquisition (S. T. Piantadosi et al., 2013; S. T. Piantadosi, Tenenbaum, & Goodman, 2016).

Setup

At its core, the model learns a set of meanings m given utterances u and contexts c. Each of these meanings connects two events. We use a simplified world model where each event E is an interval over time, with start and end points e_1 and e_2. Here, time is discrete (i.e. integral) and linear, which allows for efficient computation in our learning model while being highly compatible with work on temporal perception (Ivry & Hazeltine, 1995). A **context** comprises of two events A and B, along with the time of utterance t.[1]

Meanings are defined on top of these contexts as lambda expressions with interval-based event representations (see Table 1). For example, 'before' is represented as

$$\lambda ABt\,.\,a_1 < b_1.$$

This would return true in any context $c = (a_1, a_2, b_1, b_2, t)$ where $a_1 < b_1$. When learning, we consider mapping from tuples (u_i, c_i) to meanings m_i. We aim to compute

$$p(m \mid u, c) \propto p(u \mid m, c) \cdot p(m), \qquad (1)$$

the probability of meaning set $m = \langle m_1, \ldots, m_5 \rangle$ given utterances $u = \langle \text{before}, \ldots \text{while} \rangle$ and observed contexts c.

Computing the Prior

At a very high level, the prior encodes beliefs about which meanings—that is, which lambda expressions—are more likely *a priori*. In particular, we would like to encode a **simplicity bias**: humans tend to learn simpler hypotheses when possible, provided they are consistent with observed data (Feldman, 2000; Chater & Vitányi, 2003).

To do so, we construct a probabilistic context-free grammar (PCFG) to generate meanings. The exact specification is provided in Gorenstein et al. (2020), but in brief, it provides expansion rules for primitive logical operators (*and, or, not*) and integer relations (less than, less than or equal to, and equal to). The logical connectives pass as inputs Boolean values, and the integer relations accept the integer values of the context, i.e. a_1, a_2, b_1, b_2, t. This grammar produces Boolean functions of the inputs.

This grammar is extremely expressive, producing $2^{2^{n+1}}$ possible expressions for every n logical operators. Indeed, it can create many expressions entirely unrelated to the meanings we consider here. Thus, it is nontrivial to correctly learn specific target hypotheses because it involves sorting over a very large space of possibilities.

We convert this grammar specification into a PCFG by assigning uniform weights to each non-terminal expansion rule. We define $p(m)$ as the probability distribution induced over the PCFG. In this way, longer—and thus, more complex—hypotheses have lower production probabilities, reflecting Feldman (2000)'s simplicity bias.

Computing the Posterior

The posterior probability $p(u \mid m, c)$ corresponds to the learner's model of production. That is, it represents the learner's *beliefs* regarding utterance production: given a speaker with a set of meanings m, which utterance would they choose to describe context c_i?

Utterance probabilities are independent of one another, such that the overall posterior probability can be decomposed into

$$p(u \mid m, c) = \prod_{\substack{u_i \in u, \\ c_i \in c}} p(u_i \mid m, c_i). \qquad (2)$$

Intuitively, this probability should be inversely proportional to the number of possible true utterances. If a context can be correctly described through just one possible utterance, that utterance will be chosen with much higher probability than if, say, three utterances were possible.

Formally, for any context, we define two subsets of w. $w^m_{\text{true}}(c_i)$ consists of those utterances that are true given the set of meanings m and the context c_i, and $w^m_{\text{false}}(c_i)$ consists of all other possible utterances. We assume that the data is somewhat noisy—that is, speakers are not guaranteed to produce accurate utterances. We model the probability that speakers are 'genuine,' i.e. that they generate true utterances, as $\alpha = 0.95$.

[1] In the implementation of our model, we restrict $t \in [-100, 100]$, which ensures that inference over contexts is tractable. Without this restriction, the model would have access to infinite possible events and utterance times. This modeling assumption reflects the fact that speakers are unlikely to speak about the relative ordering of two events in the far future or distant past.

Then, we can define the probabilities in Equation 2 as

$$p(u_i \mid m, c_i) = \begin{cases} \dfrac{\alpha}{|w^m_{\text{true}}(c_i)|} + \dfrac{1-\alpha}{|w|} & \text{if } u_i \in w^m_{\text{true}}, \\ \dfrac{1-\alpha}{|w|} & \text{if } u_i \in w^m_{\text{false}}. \end{cases} \tag{3}$$

If an utterance is true, and the speaker acts genuinely, we *uniformly* sample from possible true utterances in w^m_{true}; if the speaker does not act genuinely, we uniformly sample from all possible utterances. If an utterance is false, we randomly select from all possible words if the speaker is not genuine (if the speaker is genuine, the probability is 0).

This formulation obeys Tenenbaum (1999)'s size principle, such that learners are biased towards more specific meanings. Thus, it overcomes the 'subset problem' where a learner does not acquire distinct meanings for more specific utterances (e.g. 'until' relative to 'before').

Augmenting the Model

Gorenstein et al. (2020) show that the above formulation very accurately models when connectives are learned (see Figure **??**A). However, some of the assumptions made do not accurately simulate human learning. Particularly, in modeling the posterior, Gorenstein et al. (2020) assume that speakers will select among all true utterances with uniform probability. This, however, is not necessarily true.

Speakers are generally biased towards producing more **informative** utterances whenever possible (Grice, 1975). This is a well-documented fact which has seen much success in both theoretical accounts of pragmatic production as well as modern computational models (Goodman & Frank, 2016; Degen, 2023). Thus, a speaker would be more likely to produce 'since' over 'before' in a context where both are true, because 'since' is more informative.

In the following, I formally describe this idea of informativity through a framework used in S. T. Piantadosi et al. (2013). I then show that this enhanced model does *not* exhibit human-like learning tendencies.

Informativity

As in S. T. Piantadosi et al. (2013), we model informativity by weighting production probabilities such that true words are more likely to produced if they are more informative on average. It is important to know that in this sense, we are not measuring how informative a given utterance is about the specific context at hand; instead, we are measuring how informative the utterance is in an *average* context.

If an utterance tends to be true less often, we say that it is more informative (Frank & Goodman, 2012). Thus, we write the weight q of u_i

$$q(u_i) = \frac{1}{\nu + p_{\text{true}}(u_i)}, \tag{4}$$

where $p_{\text{true}}(u_i)$ is the probability that u_i is true in a random context, and $\nu = 0.2$ is a smoothing term which ensures that the weights do not blow up. Thus, if an utterance has a high probability of being true in an average context, it is less likely to be uttered compared to a highly specific utterance. We estimate $p_{\text{true}}(u_i)$ over a large (100,000 data points) simulated dataset of randomly generated contexts (see Table 2).

Word	$p_{\text{true}}(w)$	**Weight**
before	0.498	0.861
after	0.829	0.593
since	0.199	1.422
until	0.198	1.421
while	0.662	0.703

Table 2: Empirical informativity weights as formulated by Equation 4, with $\nu = 0.25$. We stochastically generate a dataset of 100k contexts. For each of these contexts, we apply the target meaning functions of Table 1, and return the frequency at which word w is true. These weights are then normalized to sum to 5.

With these weights, we redefine the posterior probabilities of Equation 3 as

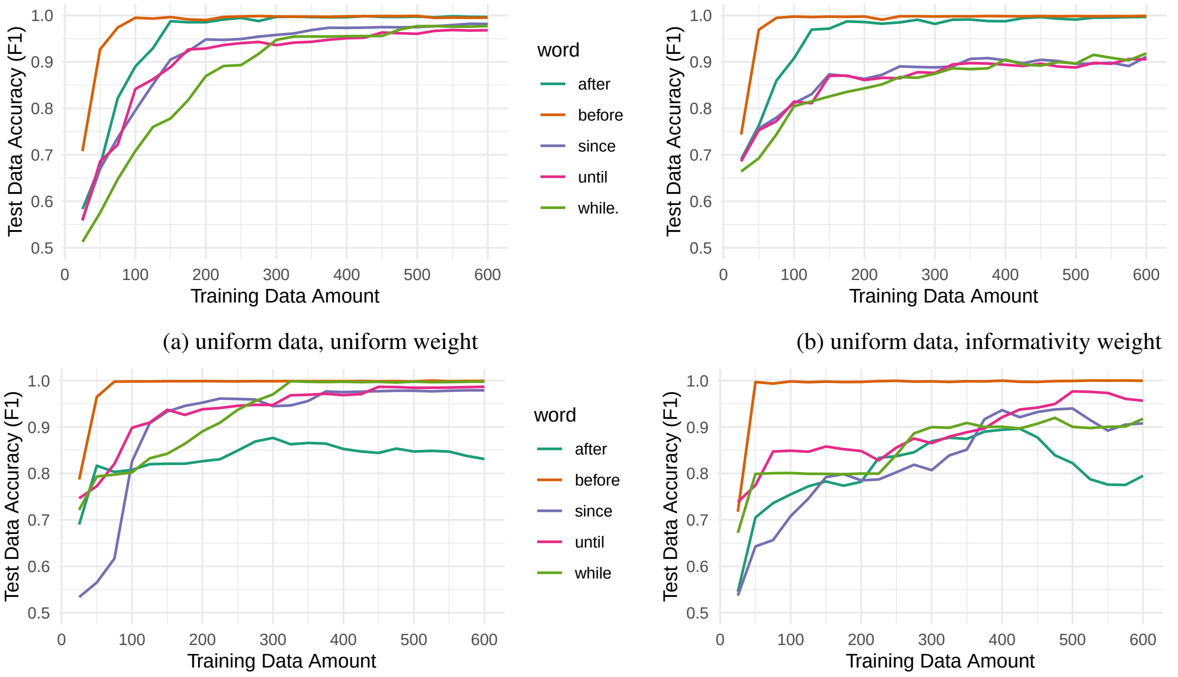
$$p(u_i) = \begin{cases} \dfrac{\alpha \cdot q(u_i)}{|w^m_{\text{true}}(c_i)|} + \dfrac{(1-\alpha) \cdot q(u_i)}{|w|} & \text{if } u_i \in w^m_{\text{true}}, \\ \dfrac{(1-\alpha) \cdot q(u_i)}{|w|} & \text{if } u_i \in w^m_{\text{false}}. \end{cases} \tag{5}$$

Inference

Many of the hypotheses generated by the grammar will have either low posterior probabilities (because of an inability to explain the data) or low prior probabilities (because of length). Thus, to approximate the posterior, we rely on Markov chain Monte Carlo (MCMC) methods. To construct appropriate Markov chains, we use Metropolis-Hastings with 50,000 steps.

We incrementally sample in 24 steps of 25 data points each. At each increment, we randomly generate contexts over a time interval $[-100, 100]$ and use the posterior of Equation 5 to generate utterances. We then store the top ten sets of meanings with the greatest posterior probabilities from the Metropolis-Hastings algorithm at each step of training.

Unlike Gorenstein et al. (2020), who assume that listeners are exposed to all connectives with uniform frequency, we also use approximated frequencies from the CHILDES corpus of child-directed speech (MacWhinney, 2000). This more accurately reflects the variety of data that learners are exposed to and has a relatively large effect on learning rate. In particular, I find that 'while' makes up around 32% of all connectives, 'before' makes up around 28%, and 'until' makes up around 20%, while 'after' forms around 12% of the data and 'since' forms only around 8%. This disparity—particularly between 'before'/'after' and 'since'/'until'—is a key aspect of the data that language learners have access to.



(a) uniform data, uniform weight

(b) uniform data, informativity weight

(c) CHILDES data, uniform weight

(d) CHILDES data, informativity weight

Figure 1: Results from model evaluation. At each increment of twenty-five data points, we evaluate and average over the top ten highest likelihood meaning-set hypotheses on 10,000 automatically generated test contexts. Random chance performance would then be 50%.

Results

In sum, we study the learning rate of connectives when alternating for data uniformity (uniform and CHILDES) and for informativity. For each condition, we train 30 models. With each model, we evaluate the ten top-performing hypothesis sets on a 10,000 example test set. This test set consists of 5,000 true and 5,000 false context-utterance pairs generated using the meanings in Table 1; chance performance falls at 50%. We test the accuracy of the learned model hypotheses as a classification task.

In Figure 1a, we replicate Gorenstein et al. (2020)'s results with uniform data and uniform weights, finding that the order of learning (i.e. when the model achieves 0.9 accuracy) is 'before' $\prec$ 'after' $\prec$ 'since' $\sim$ 'until' $\preceq$ 'while.' When we introduce the CHILDES data, however, which significantly decreases the ratio of 'after' and 'since,' we see the learning rate of 'after' degrade significantly (Figure 1c). In particular, our model never fully acquires the semantics of 'after.'

When we introduce informativity weights to the posterior production probability, this learning rate breaks down further. In the uniform data case (Figure 1b), 'before' is learned before 'after,' but 'since,' 'until,' and 'while' are learned roughly simultaneously with no real distinction. Incorporating the CHILDES data ratios (Figure 1d) adds further noise to the data: 'before' is still consistently learned first, but the rest of the connectives experience roughly simultaneous trajectories.

Analysis

The common thread between all four of these conditions is that 'before' is consistently learned before all other connectives, with very high accuracy. When we introduce informativity weights into the calculation of the posterior, the learning trajectories of 'since,' 'until,' and 'while' become less distinct. In large part, this is due to the fact that 'while' tends to be relatively uninformative, while 'since' and 'until' are highly informative in contexts where they are true (see Table 2).

This tendency for 'while,' 'until,' and 'since,' to be learned simultaneously persists in the CHILDES data models. The notable difference here is that 'after' is consistently learned much later than all other connectives, and never achieves very high accuracy despite not being semantically complex. In the CHILDES data, 'after' is the second least common connective after 'since,' which leads to it being significantly underrepresented in the data. Additionally, while 'since' is generally highly informative and true in a very small number of contexts, 'after' is the exact opposite. Thus, the data poverty of 'since' is mitigated by production constraints, but this is not the case for 'after.'

Indeed, the uniform weight CHILDES model (i.e. Figure 1c) exhibits very similar learning trajectories to the uniform data model (Figure 1a) for all connectives after the first few iterations, *except* for 'after.'

So then, Gorenstein et al. (2020)'s model does relatively well even after incorporating real-world considerations of pragmatic production and data poverty in every case besides ‘after.’ What makes ‘after’ so difficult to learn for the model in a way that does not exist in human data?

Contexts that are true for our formulation of ‘after’ are more likely to be compatible—on average—with other connectives as well, moreso than for any other connective in our data set. In other words, if after $\in w^m_{\text{true}}(c_i)$,

$$\mathop{\mathbb{E}}_{c_i}[|w^m_{\text{true}}(c_i)|],$$

the size of $|w^m_{\text{true}}|$ averaging over contexts, is very high. Indeed, both ‘while’ and ‘until’ necessarily entail ‘after.’ Thus the posterior probability of ‘after’ drops as the model gets closer to the true meaning. Because ‘after’ is significantly underrepresented the CHILDES data, it is difficult for the model to acquire its correct meaning. Part of this is induced by the fact the CHILDES data is already reflective of speaker production probabilities. That is, if a word has a low posterior, it is unlikely to be represented in the data, leading to a compounding effect.

It is similarly worthwhile to ask if this issue in learning comes out of the formulation of the prior probability; i.e. is a language-of-thought model even the right model for this learning task? As positive evidence for LOT models, the successes of the models trained above generally arose from principles of semantic complexity which are encoded by the prior. For example, ‘before’ is less semantically complex than ‘since’ and ‘until,’ despite not having a high posterior probability.[2] Thus, it has a very high prior $p(m)$ which allows it to be learned very quickly.

Learning Without a Language-of-Thought

As negative evidence for this question, it is interesting to consider how a model without an explicit language-of-thought would respond to this task. Here, I examine the acquisition of temporal connectives in a long short-term memory recurrent neural network (LSTM RNN) trained only on English data. This model does *not* have explicit access to any notion of semantic complexity, which allows us to determine whether complexity must necessarily be ‘hard-wired’ into a learning model for it to exhibit human-like behavior.

Setup

I study the behavior of the LSTM architecture from Gulordava, Bojanowski, Grave, Linzen, and Baroni (2018), trained on data from English Wikipedia.[3] We normalize the data to uniformly represent each connective, which results in a total corpus of 30 million words.

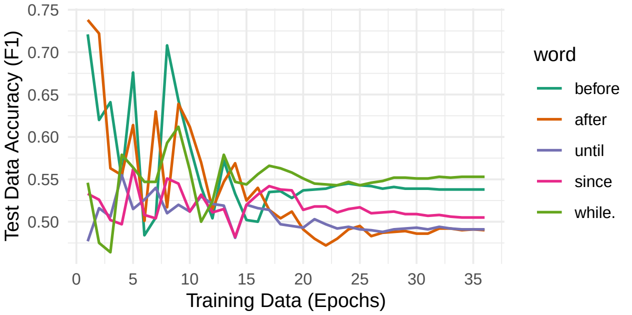


Figure 2: Results from language model evaluation. At each epoch, we evaluate the model on 50,000 data points and train a logistic classifier on the model’s normalized softmax activation for each connective. We plot the accuracy of this classifier on the y-axis.

Evaluation

After each pass through the training data (i.e. each epoch), I evaluated the language model on a test set of 10,000 contexts per connective. Each context consisted of two temporal events described in natural language (e.g. ‘Alice started swimming at a_1 and stopped swimming at a_2’), as well as an utterance time. I then extracted the model surprisal—calculated from normalized softmax activation—-for the connective of interest (e.g. $-\log p(\text{before})$ in ‘Alice swam *before* Bob cooked’). This practice is standard in work evaluating linguistic knowledge encoded in language models, as it allows us to quantitatively extract something analogous to human acceptability judgements (Futrell et al., 2019).

If the model correctly acquired the meaning of a connective, we would expect low surprisals in true contexts and high surprisals in false contexts. To this end, for each connective at each epoch, we train a simple logistic classifier to map from surprisal values to true/false judgements (Papadimitriou, Chi, Futrell, & Mahowald, 2021). We then return the accuracy (F1) of this classifier. If the classifier is highly accurate, this means that the model surprisal is a good predictor of truth value, thus quantifying the degree to which the model has learned connective meaning.

Results

We find that overall, the LSTM does not exhibit human-like learning trajectories for temporal connectives (Figure 2). The model struggles to acquire all connectives, but does best on ‘while’ and ‘before,’ ultimately not resembling human acquisition trends (e.g. Figure 1a).

This is likely in large part due to the fact that the language models do not represent semantic complexity in an explicit way. Thus, their computation of $p(m \mid u, c)$ does not encode biases towards simpler expressions, which leads to learning behavior which does not penalize complexity.

[2] Similarly to ‘after,’ ‘before’ is compatible with very ‘general’ contexts, which leads to a low production probability.

[3] The model has two 650 unit hidden layers and was trained with a batch size of 128, a dropout rate of 0.2, and an initial learning rate of 20.

Conclusions

In this work, I expanded upon a model of temporal connective acquisition by introducing conditions that more closely reflect real-world learning situations. I found that the model does well when introduced to a more realistic production setup, using evidence from pragmatic language modeling. However, I showed that with more realistic training data, the model fails to follow human learning patterns.

It is likely that this breakdown emerges out of the formulation of the posterior production probability $p(u \mid m, c_i)$, rather than out of the prior probability $p(m)$ which encodes semantic complexity. To this end, I provide both positive evidence from the language-of-thought model and supporting negative evidence from an LSTM language model. Future work should focus on both (1) better modeling the posterior and (2) more rigorously studying the learning behaviour of (larger) language models. To this second point, it is hard to make claims about the capabilities of language models purely based on the work done here, as the models were trained on a rather small amount of data.

On a broader scale, this work joins a wider discussion about how meaning is represented in human language acquisition. Much recent work has argued that in light of the capabilities of large language models, 'classical' approaches to meaning representation fail (S. Piantadosi, 2023; Katzir, 2023; Wilcox, Futrell, & Levy, 2022). Here, I show that language models trained without explicit meaning representations generally do not follow human-like acquisition trajectories, while language-of-thought models *do*.

References

Anscombe, G. E. M. (1964). Before and after. *The Philosophical Review*, *73*(1), 3–24.

Beaver, D., & Condoravdi, C. (2003). A uniform analysis of 'before' and 'after'. In *Semantics and linguistic theory* (Vol. 13, pp. 37–54).

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19–22.

Clark, E. V. (1971). On the acquisition of the meaning of before and after. *Journal of Verbal Learning and Verbal Behavior*, *10*(3), 266–275.

Degen, J. (2023). The rational speech act framework. *Annual Review of Linguistics*, *9*, 519–540.

Feagans, L. (1980). Children's understanding of some temporal terms denoting order, duration, and simultaneity. *Journal of Psycholinguistic Research*, *9*(1), 41.

Feldman, J. (2000). Minimization of boolean complexity in human concept learning. *Nature*, *407*(6804), 630–633.

Frank, M. C., & Goodman, N. D. (2012). Predicting pragmatic reasoning in language games. *Science*, *336*(6084), 998–998.

Futrell, R., Wilcox, E., Morita, T., Qian, P., Ballesteros, M., & Levy, R. (2019, June). Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 32–42). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N19-1004` doi: 10.18653/v1/N19-1004

Goodman, N. D., & Frank, M. C. (2016). Pragmatic language interpretation as probabilistic inference. *Trends in cognitive sciences*, *20*(11), 818–829.

Goodman, N. D., & Stuhlmüller, A. (2013). Knowledge and implicature: Modeling language understanding as social cognition. *Topics in cognitive science*, *5*(1), 173–184.

Gorenstein, M., Zhang, C., & Piantadosi, S. T. (2020). A model of temporal connective acquisition. In *Proceedings of the 41st annual conference of the cognitive science society.*

Grice, H. P. (1975). Logic and conversation. In *Speech acts* (pp. 41–58). Brill.

Gulordava, K., Bojanowski, P., Grave, E., Linzen, T., & Baroni, M. (2018, June). Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long papers)* (pp. 1195–1205). New Orleans, Louisiana: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/N18-1108` doi: 10.18653/v1/N18-1108

Ivry, R. B., & Hazeltine, R. E. (1995). Perception and production of temporal intervals across a range of durations: evidence for a common timing mechanism. *Journal of Experimental Psychology: Human Perception and Performance*, *21*(1), 3.

Kamp, J. A. W. (1968). *Tense logic and the theory of linear order*. University of California, Los Angeles.

Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition. a reply to piantadosi (2023). *Ms., Tel Aviv University*.

MacWhinney, B. (2000). *The childes project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.

Moens, M. (1987). Tense, aspect and temporal reference.

Papadimitriou, I., Chi, E. A., Futrell, R., & Mahowald, K. (2021, February). Multilingual BERT, ergativity, and grammatical subjecthood. In *Proceedings of the society for computation in linguistics 2021* (pp. 425–426). Online: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/2021.scil-1.50`

Piantadosi, S. (2023). Modern language models refute chomsky's approach to language. *Lingbuzz Preprint, lingbuzz, 7180*.

Piantadosi, S. T. (2011). *Learning and the language of thought*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

Piantadosi, S. T., & Jacobs, R. A. (2016). Four problems solved by the probabilistic language of thought. *Current Directions in Psychological Science*, *25*(1), 54–59.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2013). Modeling the acquisition of quantifier semantics: a case study in function word learnability. *Unpublished Manuscript*.

Piantadosi, S. T., Tenenbaum, J. B., & Goodman, N. D. (2016). The logical primitives of thought: Empirical foundations for compositional cognitive models. *Psychological review*, *123*(4), 392.

Tenenbaum, J. B. (1999). *A bayesian framework for concept learning*. Unpublished doctoral dissertation, Massachusetts Institute of Technology.

von Fintel, K., & Matthewson, L. (2008). Universals in semantics. , *25*(1-2), 139–201. Retrieved 2023-06-07, from https://doi.org/10.1515/TLIR.2008.004 doi: doi:10.1515/TLIR.2008.004

Wilcox, E. G., Futrell, R., & Levy, R. (2022). Using computational models to test syntactic learnability. *Linguistic Inquiry*, 1–88.

Winskel, H. (2003). The acquisition of temporal event sequencing: A cross-linguistic study using an elicited imitation task. *First Language*, *23*(1), 65–95.