Humans vs. Machines: Comparing Adjective Learning Performance

It has been argued that children rely on syntactic bootstrapping during the word learning process,[4,8] in which they recruit "frames" (syntactic environments) to narrow down possible word meanings. Although the literature has predominately focused on verbs, recent research has examined the power of bootstrapping in the adjectival domain.[5] While adjectives are found in frames that are often un(der)informative for subcategorization— *single* frames are consistent with several subclasses[1,2]— tracking an adjective across environments is revealing.[5] However, the relative informativity of the individual frames has yet to be examined, leaving open the question of what exactly learners are using to "bootstrap" meaning. At the same time, interest in language models (LMs) has re-ignited a debate surrounding statistical learning and the potential of generalizing patterns based on probabilities.[6,10] Are learners simply tuned to regularities in the input (as with LMs), or are they sensitive to deeper properties of the grammar (as with bootstrapping)? In this research, we (i) assess the contribution of individual frames on adjective learning, and subsequently (ii) compare human performance with a language model trained specifically on CHILDES.

Experimental Study: Here we adopt a tradition inspired by the Human Simulation Paradigm[3] of assessing the cues that adults integrate to predict word meanings as a proof-of-concept for establishing the usefulness of such frames in the input to the child. We conducted a fill-in-the-blank task, in which participants (N=100) read sentences missing a single word, and were asked to provide a guess as to what that missing word was; target frames were compatible with adjectives belonging to one of five adjective classes (tough-adjectives, aesthetic, dimensional, and evaluative control adjectives, and predicates of personal taste), whereas control frames were compatible with either nouns or verbs. To further assess the contribution of the syntax alone, participants were randomly assigned to a bleached condition, in which semantic information was removed, or a "contentful" condition, in which it was included.

Model Analysis: Our experimental design allows for a simple comparison to model behavior. We evaluated BabyBERTa[5], which is a masked language model trained on CHILDES to predict missing words in a larger context (e.g., a whole sentence). We gathered the top 10 words that BabyBERTa predicted as most likely for each of the frames from the experimental stimuli. In addition to predictions, we calculated entropy – a measure of uncertainty with smaller values representing more constrained predictions.

Results: For human learners, we find that certain syntactic cues strongly favor particular adjective classes, regardless of the condition. However, semantic information "sways" participant responses for frames that are compatible with a larger number of adjective classes (Fig-1). BabyBERTa, however, often fails to predict adjectives at all, even when semantic information is provided (Fig-2). Overall, these results indicate that humans recruit syntactic information in a way that goes beyond simple co-occurrence patterns learned by an LM, and that certain cues bias learners toward particular categories—suggesting that although a given frame may be compatible with multiple classes, not all are considered equally likely.

Figure 1. Human Performance (Experimental Study). Cells represent proportion of responses coded by word-type (5 adjective classes, other adjectives, and other non-adjectives) with green representing the most produced category and grey representing other grammatical options for the target adjectives.

Syntactic Cue(s)	Condition	tough	evaluative	aesthetic	dimensional	PPTs	Other (ADJ)	Other (Non-ADJ)
expletive + INF	Vague	80%	3%	0%	1%	2%	6%	8%
	Contentful	72%	18%	0%	0%	1%	3%	6%
expletive + judge	Vague	76%	2%	0%	12%	4%	6%	0%
	Contentful	74%	0%	0%	16%	8%	2%	0%
INF + judge	Vague	42%	0%	0%	0%	6%	52%	0%
	Contentful	72%	0%	0%	0%	2%	10%	16%
gerund	Vague	92%	4%	0%	0%	4%	0%	0%
	Contentful	80%	2%	0%	0%	14%	4%	0%
comparison class	Vague	14%	4%	0%	64%	0%	14%	4%
	Contentful	0%	4%	0%	90%	2%	4%	0%
too ADJ	Vague	0%	0%	0%	60%	0%	6%	34%
	Contentful	6%	4%	0%	68%	0%	22%	0%
measure phrase	Vague	0%	0%	0%	60%	0%	6%	34%
	Contentful	0%	0%	0%	98%	0%	0%	2%
to-phrase	Vague	34%	8%	2%	30%	10%	12%	4%
	Contentful	10%	8%	6%	10%	54%	12%	0%
such ADJ	Vague	30%	4%	14%	22%	4%	26%	0%
	Contentful	2%	8%	4%	70%	2%	14%	0%
exclamative	Vague	48%	6%	12%	12%	16%	2%	4%
	Contentful	30%	16%	18%	10%	4%	22%	0%

Figure 2. Model Performance (BabyBERTa). Cells represent proportion of responses coded by word-type (5 adjective classes, other adjectives, and other non-adjectives) with green representing the most produced category, grey representing other grammatical options, and red representing cases where the most likely category differs from humans.

Syntactic Cue(s)	Condition	tough	evaluative	aesthetic	dimensional	PPTs	Other (ADJ)	Other (Non-ADJ)
expletive + INF	Vague	50%	10%	0%	0%	0%	10%	30%
	Contentful	40%	10%	0%	0%	0%	0%	50%
expletive + judge	Vague	30%	10%	0%	50%	0%	0%	10%
	Contentful	30%	10%	0%	50%	0%	0%	10%
INF + judge	Vague	10%	0%	0%	0%	0%	10%	80%
	Contentful	10%	0%	0%	0%	0%	10%	80%
gerund	Vague	30%	20%	10%	20%	20%	0%	0%
	Contentful	30%	20%	10%	30%	10%	0%	0%
comparison class	Vague	20%	10%	0%	60%	0%	0%	10%
	Contentful	20%	10%	0%	40%	0%	10%	20%
too ADJ	Vague	20%	0%	0%	80%	0%	0%	0%
	Contentful	20%	10%	0%	40%	0%	10%	20%
measure phrase	Vague	0%	0%	0%	0%	0%	10%	90%
	Contentful	10%	10%	0%	0%	0%	20%	60%
to-phrase	Vague	40%	20%	0%	20%	10%	10%	0%
	Contentful	20%	20%	0%	20%	20%	10%	10%
such ADJ	Vague	10%	10%	0%	20%	0%	0%	60%
	Contentful	30%	30%	10%	20%	10%	0%	0%
exclamative	Vague	10%	10%	10%	20%	0%	20%	30%
	Contentful	0%	0%	0%	20%	10%	70%	0%

Selected References

[1] Bylinina, L. (2014). *The grammar of standards: Judge-dependence, purpose-relativity, and comparison classes in degree constructions*. PhD Thesis. **[2]** Davies, C., Lingwood, J., & Arunachalam, S. (2020). Adjective forms and functions in British English child-directed speech. *Journal of child language*, *47*(1). **[3]** Gillette, J., Gleitman, H., Gleitman, L., & Lederer, A. (1999). Human simulations of vocabulary learning. *Cognition*, *73*(2), 135-176.**[4]** Gleitman, L. (1990). The structural sources of verb meanings. *Language acquisition*, *1*(1). **[5]** Gotowski, M. (2022). *Syntactic bootstrapping in the adjectival domain: learning subjective adjectives*. PhD Thesis. **[6]** Huebner, P., E. Sulem, F. Cynthia, & D. Roth. (2021). BabyBERTa: Learning More Grammar With Small-Scale Child-Directed Language. In *Proceedings of the 25th Conference on Computational Natural Language Learning,* 624–46. **[7]** Katzir, R. (2023). Why large language models are poor theories of human linguistic cognition. A reply to Piantadosi. To appear in Biolinguistics. **[8]** Landau, B., & Gleitman, L. R. (1985). Language and experience: Evidence from the blind child. **[9]** Lidz, J. (2020). Learning, memory, and syntactic bootstrapping: A meditation. *Topics in Cognitive Science*, *12*(1). **[10]** Piantadosi, S.T. (2024). Modern language models refute Chomsky's approach to language. in From fieldwork to linguistic theory: A tribute to Dan Everett. E. Gibson, M. Poliak, (Eds). Language Science Press.