Failures and successes to learn a core conceptual distinction from the statistics of language

Zhimin Hu[1], Jeroen van Paridon[1], and Gary Lupyan[*1]

[*]corresponding author
[1]Department of Psychology, University of Wisconsin-Madison, Madison, USA

Abstract

Generic statements like "tigers are striped" and "cars have radios" communicate information that is, in general, true. However, while the first statement is true *in principle*, the second is true only statistically. People are exquisitely sensitive to this principled-vs-statistical distinction. It has been argued that this ability to distinguish between something being true by virtue of it being a category member versus being true because of mere statistical regularity, is a general property of people's conceptual machinery and cannot itself be learned. We investigate whether the distinction between principled and statistical properties can be learned from language itself. If so, it raises the possibility that language experience can bootstrap core conceptual distinctions and that it is possible to learn sophisticated causal models directly from language. We find that language models are all sensitive to statistical prevalence, but struggle with representing the principled-vs-statistical distinction controlling for prevalence. Until GPT-4, which succeeds.

Keywords: distributional semantics; generics; world models

1. Introduction

People interpret generic statements such as *airplanes have wings*, and *dogs bark* to mean that the named property is, in general, true of the category (Hollander et al., 2009). Other statements of this form, however, such as *airplanes carry passengers* and *dogs wear collars*, while also being judged as generally true, have a decidedly different quality. In a series of papers, Prasada and colleagues (Prasada & Dillingham, 2006; Prasada, 2016; Prasada et al., 2013) drew a distinction between generics that express *principled* properties and generics that express merely *statistical* properties. A statement expressing a principled property such as *airplanes have wings* retains its truthfulness when asked whether it is true because of (or by virtue of) being that thing. For example, in the experiments we describe below, on a scale of -3 = completely false to +3 = completely true, people judged the statement *airplanes have wings* with mean of 2.9. This declines only slightly if asked whether it is true that airplanes have wings *because they are airplanes*

(M=2.6). A statement like *airplanes have passengers* is judged to also be mostly true (M=1.8), but if asked whether airplanes have passengers *because* they are airplanes, the truth estimate drops (M=0.6). Importantly, this key result remains when one controls for confounds such as prevalence and cue-validity, showing that it is not simply an artifact of principled connections being more common or it being harder to come up with counter-examples.

Results like these have been used to argue that people's ability to distinguish between principled and statistical generics requires an *a priori* sensitivity to a distinction between statistical vs. "in-principle" properties. Because there are no structural differences between generics that could inform this distinction, it is thought that the distinction cannot be learned through associations (see Prasada et al., 2013; Haward, Wagner, Carey, & Prasada, 2018), and perhaps cannot even be represented by an associative mechanism (Prasada, 2021).

However, even though generic statements do not encode the principled/statistical distinction in their structure, the distinction might still be captured in the distributional structure of language itself. In this study, we investigated whether the statistical/generic distinction is recoverable from the statistics of language. We did this by predicting human judgments of generic statements from judgments derived from distributional language models. Finding that this distinction can be learned by an associative mechanism from language alone is important for two main reasons. First, it shows that it is *in principle* possible to learn a formal conceptual distinction argued to be unlearnable (and even unrepresentable) by an associative mechanism. Second, it opens the door to asking questions of key interest to the study of language evolution: (1) Are languages structured to facilitate extracting principled item-property relationships, (2) Where in language is such information represented? (3) Are languages not only a *source* of generic information (Rhodes, Leslie, & Tworek, 2012) but do they help structure the very core of our conceptual system?

To anticipate our results, we find that language models are all sensitive to item prevalence. Statements probing frequent item-property combinations like *orange grow on trees* and *kangaroos have pouches* are judged by models as more true than statements probing rarer item-property combinations such as *professors are absent-minded* and *birds are kept in cages*. However, a distinction in truth judgments between principled and statistical relations when controlling for prevalence and cue-validity only appeared for the largest language models we tested.

2. Human ratings

We began by constructing a corpus of 208 generic statements and having them rated on several scales using a procedure adapted from Prasada et al., 2013.

2.1. *Participants*

We recruited 91 native speakers of English residing in the United States through Amazon Mechanical Turk in exchange for a $2 payment. Seven participants were rejected for failing basic attention checks, leaving 84 participants.

2.2. *Procedure*

Participants were asked to judge four different aspects of generic statements, sentences describing properties of objects, people, and animals: (1) *Bare generic truth judgment:* "How true is the following statement: *Airplanes have seatbelts.*"; (2) *By-virtue-of truth judgment*: "How true is the following statement: *Because they are airplanes, airplanes have seatbelts.*"; (3) *Prevalence rating*: "Think of airplanes, how likely are they to have seatbelts?"; (4): *Cue validity rating*: "You learn that [unknown things/people] have wings, how likely is it [are they] to be airplanes?" Each participant was presented with 26 statements of each type. The statements were counterbalanced across participants and rating questions so that no participant rated a given generic more than once.

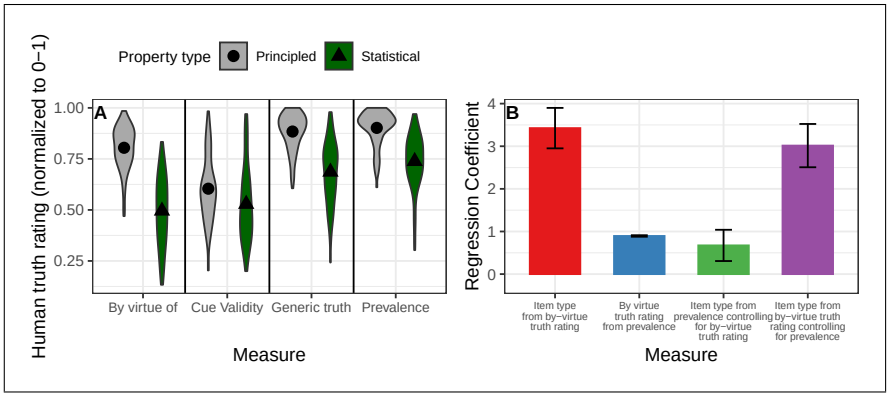

Figure 1. A. Mean human truth ratings for each sentence frame, comparing principled and statistical relationships. B. Regression coefficients (with SEs) showing key relationships between truth ratings, property type, and prevalence (see text.)

2.3. *Results*

Our results, shown in Fig. 1, closely replicate the findings of Prasada et al., 2013. Bare generics ("Airplanes have wings") expressing principled relationships are rated as more true than statements expressing statistical relationships ("Airplanes have passengers") and the same goes for by-virtue-of judgments (but more so). By-virtue truth ratings *are* affected by prevalence (the blue bar in Fig. 1B) and prevalence predicts item-type (principled vs. statistical) when controlling for the

by-truth rating (green bar). Importantly, the ability of by-virtue judgments to predict property type (red bar) is nearly undiminished when we control for prevalence (cf. red and purple bars). We will be comparing the model results to this U-shaped pattern of coefficients shown in Fig. 1B.

3. Can the principled/statistical distinction be learned from language itself?

To determine whether distributional models can differentiate between statistical and principled generics, we predicted property type (statistical vs. principled) from the cosine similarity between the target-word and the property.

3.1. *Models*

We tested the language models listed Table 1 using the Huggingface implementations of BERT (Devlin et al., 2018), ALBERT (Lan et al., 2019), DistilBERT (Sanh et al., 2019), RoBERTa (Liu et al., 2019), GPT (Radford et al., 2018), and GPT-2 (Radford et al., 2019). We used the OpenAI APIs for GPT-3.5 and GPT-4.

Table 1. Overview of the models we tested

Model name	Training sources	Size of training corpus	# Number of parameters
BERT (base)	Wiki, books	3.3B tokens (13 GB data)	110M
ALBERT (base-v1)	Wiki, books	3.3B tokens (13 GB data)	11M
Distilbert (base)	Wiki, books	3.3B tokens (13 GB data)	66M
RoBERTa (base)	Wiki, books, web crawl	161 GB data	125M
GPT	Web crawl	800M tokens	110M
GPT-2 (base)	Web crawl, Reddit,	8M documents (40 GB data)	117M
GPT-3.5	Unknown superset of GPT-2	Unknown	Unknown
GPT-4	Unknown superset of GPT-3.5	Unknown	Unknown

3.2. *Methods*

To measure the represented similarity between the target words and their properties, we first needed to obtain their model embeddings. Because the transformer models only generate contextual embeddings, we simulated a decontextualized context by using the "all but the top" method proposed by (Mu & Viswanath, 2018). This method removes the top k principal components (here, k=7) as computed by sampling additional corpuses of text from the NLI dataset (Bowman et al., 2015) and wiki-103 (Merity et al., 2016). It ensures the resulting embeddings reflect a more contrastive meaning of a given phrase. The models' truth judgment was then operationalized as the cosine similarity between the target-word (e.g., "airplanes") and the property ("have wings").

Because GPT-3.5 and 4 are fine-tuned for question-answering, it was possible to probe their 'knowledge' more directly by having them rate the generics using

the same prompt as human participants. The models received the following type prompt: *return only one integer between -3 and 3 where -3 means the sentence is definitely false and 3 means the sentence is definitely true : Because they are airplanes, airplanes have wings*. We tested each of the 208 generics 15 times and averaged the ratings. The variance of this average was less than 0.01.

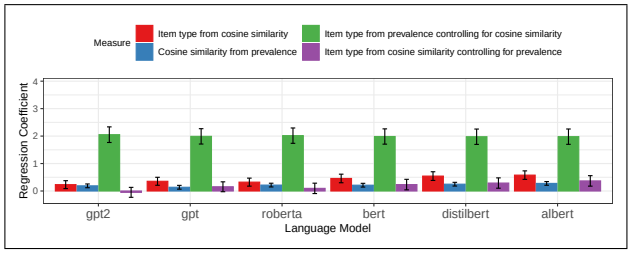

Figure 2. Regression coefficients (with SEs) indicating relationships between item-property cosine-similarity, property-type, and prevalence using the analogous models used in Fig. 2.

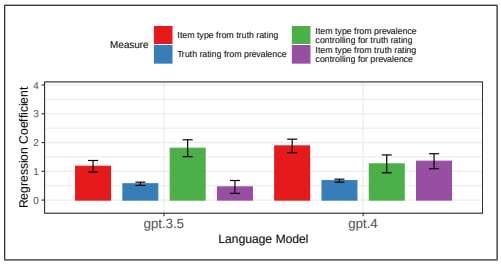

Figure 3. Regression coefficients (with SEs) indicating relationships between model-generated by-virtue-of truth ratings, property-type and human-ratings of prevalence.

4. Results

The basic pattern of results from the cosine similarity analyses is shown in Fig. 2. Across all six models we see the same qualitative pattern. The models distinguish between principled and statistical connections: the similarity between the target word like 'airplane' and a principled property like 'wings' is greater than a statistical property like 'passengers (red bars). However, when we control for prevalence, this association largely disappears (purple bars; it is only marginally above 0 in ALBERT). Human truth ratings (especially by-virtue ratings) are much better predictors of property-type than the prevalence ratings. For the models, this is not the case as indicated by the large green bar in comparison to 1B.

Turning to our experiments of GPT-3.5 and GPT-4 in which we were able to directly query their truth judgments, we find a rather different result (Fig. 3).

In GPT-3.5, truth ratings only are barely predictive of property-type when controlling for prevalence (green bar; t=2.05, p=.04), while for GPT-4 they remain strongly predictive, t=5.18, p <.00001). As a complementary analysis, we examined by-item relationships. For each item (e.g., dogs, trampolines, trumpets), we can compare the by-human virtue-of truth judgment for the principled vs. statistical statement, and compare it to the cosine-similarity-based measure for the BERT-type models and to the truth-judgments for the GPT models. We find correlations ranging from .21 for BERT to .28 for DistilBERT. These increase to .49 for GPT-3.5 and to .61 for GPT-4.

5. General Discussion

People know that airplanes have wings and carry passengers, and simultaneously know that the former but not the latter property is part of what *it means* to be an airplane. Since this distinction is not marked in language, it has been thought that it must come from elsewhere such as an innate generative type-token mechanisms (Prasada, 2016). We show here that it is, in principle, possible to learn this distinction from the statistics of language, but it is far from trivial, emerging most clearly only in GPT-4. All tested transformer models trained on English text were sensitive to prevalence as shown by significant associations between prevalence and cosine similarity/model truth judgments. People's judgments too show sensitivity to prevalence which makes sense since it is often a good proxy for whether a relationship is principled or statistical: that *principled* relationships have, on average, considerably higher prevalence than merely *statistical* ones). But human judgments continue to strongly distinguish principled and statistical relationships when prevalence is partialled out–consistent with the view that people base their judgments on causal models, presumably learned from rich multimodal experience (see e.g. Prasada & Dillingham, 2006; Prasada et al., 2013). The failure of language models to distinguish statistical from principled properties once prevalence is partialled out indicates that the models are basing their 'judgments' on statistical co-occurrence. And yet, when we test more recent models such as GPT-3.5 and especially GPT-4, the picture starts to shift consistent with the possibility that lowering next-token prediction error at scale can lead to the models inducing more sophisticated world models (e.g., Li et al., 2022; Mirchandani et al., 2023; Michaelov et al., 2023; Li et al., 2021). Although it is unknown at present what allows GPT-4 to succeed, our experiments provide an in-principle proof that it is possible to induce sophisticated causal models of item-property relations from language alone. Although it is rather unlikely that people learn the distinction between principled and statistical properties from language in the same way, our results hint that input from language may be more instrumental for laying down core conceptual distinctions than previously thought.

References

Bowman, S. R., Angeli, G., Potts, C., & Manning, C. D. (2015). A large annotated corpus for learning natural language inference. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 632–642). Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Haward, P., Wagner, L., Carey, S., & Prasada, S. (2018). The development of principled connections and kind representations. *Cognition*, *176*, 255–268.

Hollander, M. A., Gelman, S., & Raman, L. (2009). Generic language and judgements about category membership: Can generics highlight properties as central? *Language and cognitive processes*, *24*(4), 481–505.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., & Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations. In *International conference on learning representations.*

Li, B. Z., Nye, M., & Andreas, J. (2021). Implicit representations of meaning in neural language models. In *Proceedings of the 59th annual meeting of the association for computational linguistics and the 11th international joint conference on natural language processing (volume 1: Long papers)* (pp. 1813–1827).

Li, K., Hopkins, A. K., Bau, D., Viégas, F., Pfister, H., & Wattenberg, M. (2022). Emergent world representations: Exploring a sequence model trained on a synthetic task. In *The eleventh international conference on learning representations.*

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Merity, S., Xiong, C., Bradbury, J., & Socher, R. (2016). Pointer sentinel mixture models. In *International conference on learning representations.*

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2023). Can peanuts fall in love with distributional semantics? *arXiv preprint arXiv:2301.08731*.

Mirchandani, S., Xia, F., Florence, P., Driess, D., Arenas, M. G., Rao, K., Sadigh, D., Zeng, A., et al.. (2023). Large language models as general pattern machines. In *7th annual conference on robot learning.*

Mu, J., & Viswanath, P. (2018). All-but-the-top: Simple and effective postprocessing for word representations. In *International conference on learning representations.*

Prasada, S. (2016). Mechanisms for thinking about kinds, instances of kinds, and kinds of kinds. In *Core knowledge and conceptual change* (pp. 209–224). New York, NY, US: Oxford University Press.

Prasada, S. (2021). The physical basis of conceptual representation - An addendum to. *Cognition*, *214*, 104751.

Prasada, S., & Dillingham, E. M. (2006). Principled and statistical connections in common sense conception. *Cognition*, *99*(1), 73–112.

Prasada, S., Khemlani, S., Leslie, S.-J., & Glucksberg, S. (2013). Conceptual distinctions amongst generics. *Cognition*, *126*(3), 405–422.

Radford, A., Narasimhan, K., Salimans, T., Sutskever, I., et al.. (2018). Improving language understanding by generative pre-training. *OpenAI blog*.

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al.. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9.

Rhodes, M., Leslie, S.-J., & Tworek, C. M. (2012). Cultural transmission of social essentialism. *Proceedings of the National Academy of Sciences*, *109*(34), 13526–13531.

Sanh, V., Debut, L., Chaumond, J., & Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.