Distributional Semantics Still Can't Account for Affordances

Cameron R. Jones
Department of Cognitive Science
UC San Diego,
c8jones@ucsd.edu

Tyler A. Chang
Department of Cognitive Science
UC San Diego,
tachang@ucsd.edu

Seana Coulson
Department of Cognitive Science
UC San Diego,
scoulson@ucsd.edu

James A. Michaelov
Department of Cognitive Science
UC San Diego,
j1michae@ucsd.edu

Sean Trott
Department of Cognitive Science
UC San Diego,
strott@ucsd.edu

Benjamin K. Bergen
Department of Cognitive Science
UC San Diego,
bkbergen@ucsd.edu

Abstract

Can we know a word by the company it keeps? Aspects of meaning that concern physical interactions might be particularly difficult to learn from language alone. Glenberg and Robertson (2000) found that although human comprehenders were sensitive to the distinction between afforded and nonafforded actions, distributional semantic models were not. We tested whether technological advances have made distributional models more sensitive to affordances by replicating their experiment with modern Neural Language Models (NLMs). We found that only one NLM (GPT-3) was sensitive to the affordedness of actions. Moreover, GPT-3 accounted for only one third of the effect of affordedness on human sensibility judgements. These results imply that people use processes that go beyond distributional statistics to understand linguistic expressions, and that NLP systems may need to be augmented with such capabilities.

Keywords: neural language models; distributional semantics; affordances; embodied cognition

Introduction

A long-standing debate in cognitive science concerns the extent to which the meaning of a word can be learned from information about how it is distributed in language. The debate is important both theoretically, for explaining how human comprehenders understand language, and practically, for building computational systems that represent and respond to language in a human-like way. Previous work has shown that distributional theories fail to account for affordances—the actions that an agent can perform with an object—suggesting that information about how words are distributed is insufficient to explain what they mean (Glenberg & Robertson, 2000). We re-evaluate this claim with contemporary models, testing whether technological advances make it possible to extract affordance information from distributional statistics.

Distributional theories of meaning are based on the distributional hypothesis: words derive their meanings from the linguistic contexts in which they are used, i.e. the way they are *distributed* in language (Firth, 1957; Harris, 1954; Wittgenstein, 1953). These theories have been operationalized in computational models that learn, for instance, that *road* and *street* are similar, because the contexts in which they are used are similar. Recent methodological innovations have produced computational models that encode an impressive amount of linguistic knowledge (Rogers, Kovaleva, & Rumshisky, 2020; Tenney, Das, & Pavlick, 2019); and predict a number of behavioral measurements, including word relatedness (Trott & Bergen, 2021; Li & Joanisse, 2021), visual similarity ratings (Lewis, Zettersten, & Lupyan, 2019), category-membership judgements (Lenci, 2018), N400 amplitude (Michaelov, Coulson, & Bergen, 2021; Frank, Otten, Galli, & Vigliocco, 2015), and reading time (Shain, 2019; Goodkind & Bicknell, 2018). Schrimpf et al. (2021) find that transformer-based NLMs predict nearly 100% of explainable variance in neural responses to sentences (fMRI and ECoG) and suggest that "predictive ANNs serve as viable hypotheses for how predictive language processing is implemented in human neural tissue" (p.8). Critics of the distributional account, however, argue that trying to understand a word's meaning from its linguistic context is "like trying to learn a language by listening to the radio" (Elman, 1990).

The debate relates to a broader discussion about whether cognition is constituted by embodied perceptual and motor experiences of the world (Barsalou, 1999) or by formal operations on disembodied, amodal symbols (Mahon, 2015). A central critique of disembodied theories of cognition—including distributional theories of meaning—is that they do not provide a mechanism for meanings to be *grounded*. That is, the meanings of symbols in the system can only be defined with reference to other abstract symbols, and therefore do not make contact with the world (Harnad, 1990; Searle, 1980).

Glenberg and Robertson (2000) illustrated this critique by testing the sufficiency of distributional models to deal with an aspect of meaning which appears to rely on embodied experience: affordances. The concept of an affordance was introduced by Gibson (1979) to describe the set of actions that an environment makes possible for an animal. Affordances are co-determined by the environment and agent: a chair might afford sitting for a person, but not an elephant. Through our interaction with the environment, we learn about nonobvious affordances of objects that might never be described in language. For instance, though we may never have tried to chisel ice from a windshield with either a golf club or a ham sandwich, we have learned incidentally through our experience with these objects that the former would be more suitable than the latter. We might fail to pick up on such incidental properties through language experience alone.

Glenberg and Robertson (2000) tested the hypothesis that human comprehenders would be sensitive to the distinction between afforded and nonafforded actions in a way that distributional models were not. They constructed scenarios in

In J. Culbertson, A. Perfors, H. Rabagliati & V. Ramenzoni (Eds.), *Proceedings of the 44th Annual Conference of the Cognitive Science Society.*

which characters used objects in ways that were either afforded or nonafforded while ensuring that a distributional model, Latent Semantic Analysis (LSA; Landauer & Dumais, 1997) , showed no effect of the afforded/nonafforded distinction. Human comprehenders rated the afforded scenarios as significantly more sensible than the nonafforded scenarios, implying that they were sensitive to the affordances of objects, and were using information not available to LSA when comprehending the sentence. The authors took this result as evidence for the insufficiency of distributional semantic methods in accounting for human language comprehension.

There are several reasons to revisit this study, 22 years later. First, the claim that distributional semantic methods in general are insufficient to account for the influence of affordances could be undermined by data from *any* distributional model. It is therefore sensible to test this claim using a variety of distributional models. Second, enormous progress in natural language understanding in the last two decades—catalyzed by increases in processing power, training data, and architecture improvements—make modern Neural Language Models (NLMs) far more sensitive to nuances of linguistic meaning than LSA was (Kocijan, Davis, Lukasiewicz, Marcus, & Morgenstern, 2022; Wang, Singh, et al., 2019; Wang, Pruksachatkun, et al., 2019). Forbes, Holtzman, and Choi (2019) found that BERT (an NLM) is competitive with humans at predicting affordances from objects. Finally, LSA's word representations are fixed and context-invariant. Glenberg & Robertson identified this as a crucial obstacle to understanding how objects could be used in novel ways. Modern NLMs account for the influence of context on a word's meaning, providing an additional reason to believe they will have a better handle on affordances.

However, there are reasons to temper optimism. The scenarios were creatively designed to ensure that participants were unlikely to have encountered the afforded objects being used in these ways. By extension, models are unlikely to have encountered these object affordances explicitly expressed in their training data. For example, although a warm thermos could be used to press wrinkles out of a skirt, it is unlikely that this specific use appears in the training data. In order for an NLM to make this connection, *warm thermos* would need to have a sufficiently similar distributional pattern to other lexical items for which this association already exists (e.g., *iron*), such that a statistical model could identify this regularity and use it for prediction. This lack of relevant training data is compounded by reporting bias: perceptually obvious features of objects are often not discussed explicitly, precisely because they are perceptually obvious (Gordon & Van Durme, 2013). Finally, these models make frequent commonsense errors, which indicate that they lack capabilities such as world knowledge and causal reasoning (Bender & Koller, 2020; Davis & Marcus, 2015).

In the present work, we first ask whether modern NLMs are *sensitive* to affordances, by testing for an effect of afforded vs nonafforded actions on several NLM measures. Secondly, we ask whether NLMs can *account for the influence* of affordances on human judgements. We test whether the Affordedness condition (Afforded vs Nonafforded) explains marginal variance in human sensibility judgements when controlling for NLM measures. If condition explains variance on top of NLM measures, it would indicate that affordedness influences humans in a way that NLM measures do not capture.

Study 1: NLM Analysis

In our first study, we asked whether NLMs were sensitive to the distinction between afforded and nonafforded actions. Glenberg and Robertson (2000) designed their stimuli to ensure that LSA showed no difference between these conditions. Therefore, an effect of Affordededness on NLM measures would suggest that technological advances in language modelling have allowed models to extract sufficient information from distributional statistics of language to make a distinction between afforded and nonafforded actions, even when both actions were rated as equally unrelated by LSA.

Method

Materials The stimuli, from Glenberg and Robertson (2000) Experiment 1, comprised 18 scenarios. Each scenario contained a setting sentence (1) and a critical sentence (2) that described a character using an object to perform an action.

(1) After wading barefoot in the lake, Erik needed something to get dry.

(2) a. He used his *shirt* to **dry** his feet. [Afforded]
b. He used his *glasses* to **dry** his feet. [Nonafforded]
c. He used his *towel* to **dry** his feet. [Related]

There were three versions of each critical sentence, corresponding to the three conditions in the experiment. The versions differed only in the objects used by the character, referred to as the distinguishing concepts (italicized). Each critical sentence also featured a central concept (boldface), which was the same across conditions and was intended to conceptually capture the use to which the object was being put. Objects in the Afforded condition (2-a) afforded the character's intended action, while objects in the Nonafforded condition (2-b) did not. It is easy to imagine how a shirt could be used to dry wet feet; the same is not true for glasses.

To control for distributional information, the authors found the cosine angle between the LSA representations of i) the setting and critical sentences, and ii) the central and distinguishing concepts. The stimuli were designed so that there was no difference on either LSA measure between the Afforded and Nonafforded conditions. That is, to the extent that LSA measures relatedness, the distinguishing concepts in the Afforded and Nonafforded conditions were equally unrelated to the rest of the scenario. In the third, Related, condition (2-c), the object also afforded the character's intended action (as in the Afforded condition). However, the object in the Related condition was more closely related to the rest of the

scenario compared to the Afforded and Nonafforded conditions, as measured by LSA. Both towels and shirts could be used to dry wet feet, but towels are more strongly associated with drying than shirts are. These distinctions allowed us to separately test for an effect of affordedness (*shirt* vs *glasses*) and relatedness (*towel* vs *shirt*) on our NLM measures.

NLM Measures We elicited responses to the stimuli from three transformer-based NLMs: BERT (large, cased) (Devlin, Chang, Lee, & Toutanova, 2019), RoBERTa (large) (Liu et al., 2019), and GPT-3 (davinci) (Brown et al., 2020). These models learn to encode representations of language by predicting sequences of tokens (word-parts) on the basis of the surrounding linguistic context. GPT-3 is unidirectional: its representations are conditioned only on the preceding tokens in a sequence (the left context). BERT and RoBERTa are both bi-directional: their representations are conditioned on both left and right context.

We selected BERT because it is probably the most widely studied transformer NLM, having become "a ubiquitous baseline in NLP experiments" (Rogers et al., 2020); RoBERTa because it uses a BERT-like architecture, but with more extensive pre-training, which substantially improved its performance on several benchmarks (Liu et al., 2019); and GPT-3 because it set SOTA performance on several benchmarks without any task-specific fine-tuning. GPT-3 is much larger (175B parameters) and has been trained on much more data (300B tokens) than either BERT (340m parameters, 3.3B words) or RoBERTa (340m parameters, ∼30B words). We accessed BERT and RoBERTa through the *Transformers* Python package (Wolf et al., 2020), and accessed GPT-3 through the OpenAI API.

We elicited six measures from these NLMs. The first two measures closely paralleled the Glenberg and Robertson (2000) analysis. We found the cosine distance between the mean BERT embeddings for the setting and critical sentences (**BERT Cosine S-C**), and between the central and distinguishing concepts (**BERT Cosine C-D**). We used the second-to-last layer of BERT because it was found to perform better than any other single-layer representation on a named entity recognition task (Devlin et al., 2019). Larger cosine distances indicate that the distinguishing concepts are more dissimilar from their contexts, and would therefore be expected for less sensible objects.

In addition to providing representations of tokens in a sentence, NLMs also generate predictions for observing specific tokens given their surrounding context. We took advantage of this by comparing the surprisal ($-\log_2 p(token)$) of the tokens in the distinguishing concepts of each scenario version. Larger surprisal indicates a lower probability of observing the distinguishing concepts, and would therefore be expected for less sensible objects. We elicited **BERT Surprisal** and **RoBERTa Surprisal** by masking the distinguishing concept tokens and finding the mean surprisal of the masked tokens. Because GPT-3 is unidirectional, and important information appears in the right context of the distinguishing concepts, we used two different measures of GPT-3 surprisal. In the first, **GPT-3 Surprisal (dc)**, we measured the sum surprisal of the tokens in the distinguishing concept, conditioned on the left context. In the second, **GPT-3 Surprisal (dc+rc)**, we measured the mean surprisal of the tokens in the distinguishing concept and their right context. We used mean surprisal to control for variation in the length of the right context.

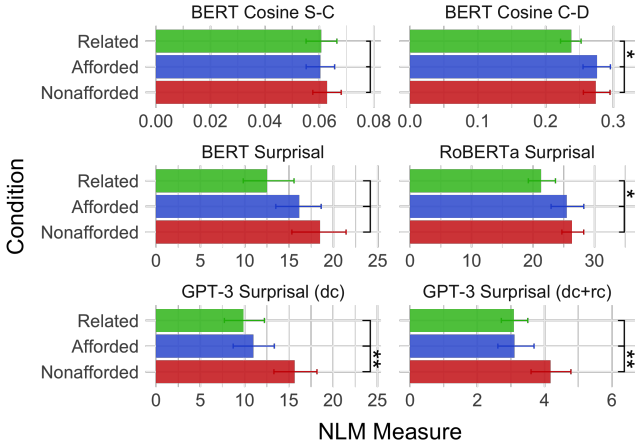

Figure 1: Only GPT-3 is surprised by nonafforded actions. GPT-3 Surprisal is significantly higher for Nonafforded actions than Afforded ones: both the total surprisal of the distinguishing concept (dc; $\chi^2(1) = 9.125$, $p = 0.003$), and the mean surprisal of the distinguishing concept plus right context (dc+rc; $\chi^2(1) = 6.617$, $p = 0.010$). RoBERTa Surprisal and the cosine distance between the central and distinguishing concepts (BERT Cosine C-D) show an effect of Related vs Afforded, but not Afforded vs Nonafforded. The BERT Cosine between setting and critical sentences (BERT Cosine S-C) and BERT Surprisal show no effects of either comparison.

Results

We created two subsets of the scenarios to separately test whether models were sensitive to affordedness (Afforded vs Nonafforded) and relatedness (Afforded vs Related). We constructed linear mixed effects models that predicted each NLM measure on the basis of condition, controlling for the log frequency of the distinguishing concept, and with random intercepts by scenario. We used likelihood ratio tests to assess whether condition improved model fit for each comparison.

Two of the NLM measures showed no differences for any of the condition analyses (BERT Cosine S-C, and BERT Surprisal). RoBERTa Surprisal and BERT Cosine C-D showed a significant difference for the Related/Afforded distinction. Both GPT-3 Surprisal (dc) ($\chi^2(1) = 9.125$, $p = 0.003$) and GPT-3 Surprisal (dc+rc) ($\chi^2(1) = 6.617$, $p = 0.010$) showed a significant effect of the Afforded/Nonafforded comparison, but no effect of Afforded/Related. Surprisal was higher for Nonafforded scenarios than for Afforded ones (see Figure 1).

Discussion

None of the BERT or RoBERTa measures showed an effect of Affordedness. In contrast, GPT-3 surprisal was significantly larger for Nonafforded vs Afforded scenarios. It is possible that BERT surprisal would show an effect on a larger sample (n=18), however, the GPT-3 effect shows that power is in principle sufficient. GPT-3's sensitivity to the Affordedness distinction might be used to support claims regarding tacit knowledge available in the model, i.e., that the model understands a shirt can be used to dry one's feet but glasses cannot.

It seems likely that GPT-3 performed better than the other models due to the relationship between the amount of compute used to train a model and its language modelling performance Brown et al. (2020). It is striking that neither of the GPT-3 measures show the effect of relatedness that was seen for the *central-to-distinguishing* LSA cosine distances in the original study, and the similar BERT Cosine C-D measure. This might indicate that GPT-3 is exploiting deeper contextual cues beyond the superficial co-occurrence statistics that characterize relatedness.

Study 2: Comparison of Human and NLM responses

The results of the NLM re-analysis provide evidence that at least one NLM (GPT-3) is sensitive to the distinction between afforded and nonafforded actions. In order to test the stronger claim that distributional information is sufficient to explain the interpretation of language by human comprehenders, we asked whether NLM measures can account for the effect of affordedness on human judgements. We replicated Experiment 1 from Glenberg and Robertson (2000), which asks human participants to rate scenarios based on how sensible they are. We then tested whether condition (Afforded vs Nonafforded) explains marginal variance in human sensibility ratings when controlling for the effect of NLM measures. A marginal effect of condition would imply that affordedness is influencing human comprehension in a way that NLM measures cannot account for.

Method

Participants All research was approved by the University's Institutional Review Board. We recruited 142 undergraduate students from the Psychology Department Subject pool, who provided informed consent using a button press and received course credit as compensation for their time. We excluded 2 participants who indicated they were not native English speakers; 7 participants who took over 1 hour to complete the experiment; 10 participants who failed $> 1/3$ attention checks; and 1 participant who had $> 20\%$ of their trials excluded. We excluded 6 trials where the response time was $> 120s$ (indicating inattention), and 66 trials where the response time was $\pm 2.5SD$ from the participant mean. We retained 2142 trials from 123 participants (90 female, 30 male, 2 non-binary, 1 prefer not to say; mean age = 20.6, $\sigma = 2.97$). The study lasted 17.6 mins on average ($\sigma = 6.37$).

Procedure We used the same stimuli as described in the NLM re-analysis section above. The procedure was similar to the method outlined in Glenberg and Robertson (2000), in that participants were asked to read the scenarios and rate the sensibility of the sentences, on a scale from 1 (virtual nonsense) to 7 (completely sensible). In the original study, participants also rated sentences for how easy they were to envision, but we only elicited sensibility judgements because the results from both ratings were very similar ($r \geq 0.9$) and because eliciting envisioning ratings might artificially induce participants to recruit perceptual experience. Moreover, in the original experiment, each participant rated all versions of each scenario. We decided to present only one version of each scenario to each participant, to prevent them from implicitly comparing different versions of the scenario. The experiment was designed using jsPsych (De Leeuw, 2015) and hosted online. Participants saw one scenario at a time and rated it on a seven point scale by clicking on a rating. Each participant saw 18 scenarios. The version of the scenario (condition) was randomized, so that all participants saw scenarios from all three conditions, but no two versions of the same scenario. The order of the items was randomized. Each participant also saw 3 attention checks that asked them to select a specific rating.

Results

We replicated the original effect of condition on human sensibility judgements in both the Afforded vs Nonafforded ($\chi^2(1) = 45.7$, $p < 0.001$) and Afforded vs Related ($\chi^2(1) = 10.6$, $p = 0.001$) comparisons (see Figure 2, left).

For the Afforded vs Nonafforded comparison, three NLM measures significantly improved the fit of a base model that predicted human sensibility judgements with a fixed effect of the log frequency of the distinguishing concept, random intercepts by item and participant, and a random effect of condition by participant (see Table 1, Afforded vs Nonafforded, NLM vs Base). For all NLM measures, a significant improvement in model fit was produced by including condition as an additional variable (see Table 1, Afforded vs Nonafforded, Full vs NLM). While GPT-3 Surprisal (dc) explained the highest proportion of variance in human sensibility judgements of any of the NLM measures (R^2 = 0.13), condition explained a much larger proportion of variance (R^2 = 0.34).

Only BERT Surprisal improved model fit over a base model in the Afforded vs Related dataset. Again, the addition of condition in the full model improved model fit over all NLM measures (see Table 1, Afforded vs Related).

Discussion

Although three of the NLM measures showed a significant effect on human sensibility judgements, no measure was able to account for all of the variance explained by Affordedness condition. This is the crucial test. If NLMs were sufficiently sensitive to affordance information to explain the effects of affordances on human language comprehension, then there would be no residual variance in responses remaining that would be explained by condition after the effect of NLM mea-

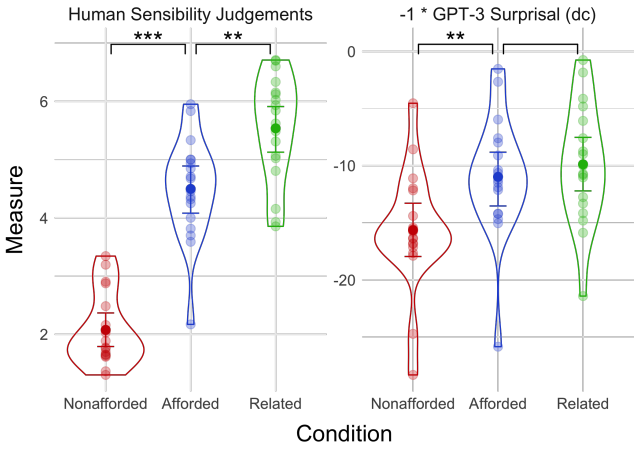

Figure 2: Human raters made more fine grained distinctions between experimental stimuli than did NLMs. **Left:** On a scale of 1 (nonsense) to 7 (sensible), participants rated Related actions as more sensible than Afforded but unrelated actions ($\chi^2(1) = 10.6$, $p = 0.001$), and Afforded actions as more sensible than Nonafforded ones ($\chi^2(1) = 45.7$, $p < 0.001$). **Right:** GPT-3 Surprisal for the distinguishing concept (dc) tokens, inverted to facilitate comparison, showed no effect of Afforded vs Related and a less pronounced effect of Afforded vs Nonafforded ($\chi^2(1) = 9.125$, $p = 0.003$).

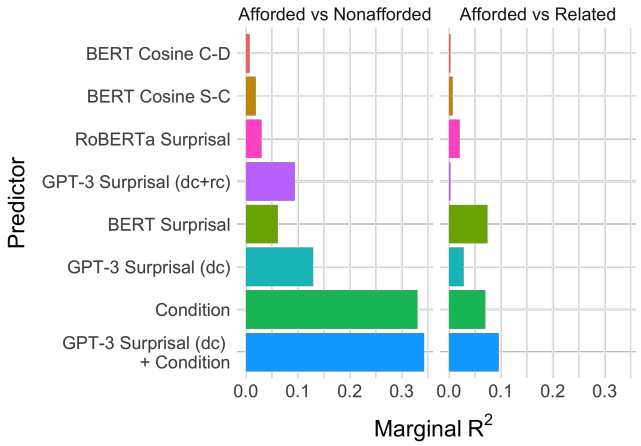

Figure 3: Affordedness explained more variance than any of the NLM measures. **Left**: For the Afforded vs Nonafforded comparison, GPT-3 Surprisal of the distinguishing concept tokens (GPT-3 Surprisal, dc) explained more variance in human sensibility judgments than any other NLM measure (R^2 = 0.13, see §NLM Measures), but only a fraction of the variance explained by condition (R^2 = 0.34). **Right:** NLMs are much closer to explaining the effects of relatedness on human judgements.

sures on human sensibility ratings had been accounted for. In fact, the best performing measure—GPT-3 Surprisal (dc)—explains only around a third of the variance explained by Affordedness. This implies that human sensibility judgements are being influenced by affordances in ways that are not being captured by even state of the art NLMs.

General Discussion

Glenberg and Robertson (2000) found that distributional techniques of the time were unable to account for the influence which affordances have on human language comprehension. They took this as evidence that human comprehenders draw on experience and processes that go beyond distributional linguistic information in order to understand language. Techniques in distributional semantics have progressed enormously in the last 22 years. We therefore asked whether modern NLMs can account for the influence of affordances, which would undermine the claim that distributional information is insufficient for learning these relations.

On a weaker interpretation of this question—*are NLMs sensitive to the distinction between afforded and nonafforded actions?*—we see progress. Specifically, GPT-3 surprisal is higher for Nonafforded vs Afforded uses of objects. This is noteworthy for several reasons. First, it contrasts with the result obtained from LSA in Glenberg and Robertson (2000), and undermines the conclusion that distributional information is insufficient to capture affordance information. Second, it is striking that only GPT-3 shows this sensitivity to affordances. All of the BERT and RoBERTa measures fail to show a significant difference between Afforded and Nonafforded conditions. This implies that affordance-sensitivity is a non-trivial property of GPT-3, specifically: perhaps a result of its larger number of parameters and training data.

On a stronger interpretation of the question—*are NLMs capable of accounting for the influence of affordances on human comprehension?*—distributional methods continue to fall short. None of the NLM measures were able to account for the effect which Affordedness had on human sensibility judgements. That is, even after controlling for the influence of the NLM measures, a large amount of variance in human responses could be explained by including condition as a predictor. This suggests that the fact of an action being afforded has a consistent influence on human sensibility judgements that is not captured by the distributional information learned by any of these NLMs. This is the crucial sense in which distributional semantics still can't account for affordances.

The evidence of progress in Study 1 invites one interpretation of these results: distributional semantic techniques are improving and with sufficient data, parameters, or architectural improvements, models will become capable of extracting any relevant information from the distributional signal needed to explain influences on human language comprehension. This interpretation is consistent with research on scaling laws: model performance increases in a law-like manner with increased computational resources (Kaplan et al., 2020).

Alternatively, one might interpret the failure of all of the NLMs to account for the influence of affordances on human

	LRT χ^2 (p-value)			
NLM Measure	Afforded vs Nonafforded		Afforded vs Related	
	NLM vs Base	Full vs NLM	NLM vs Base	Full vs NLM
BERT Cosine C-D	0.00 (0.993)	45.7 (<**0.001**) ***	0.226 (0.634)	11.7 (<**0.001**) ***
BERT Cosine S-C	0.946 (0.331)	45.3 (<**0.001**) ***	0.742 (0.389)	10.8 (<**0.001**) ***
BERT Surprisal	4.74 (**0.029**) *	44.5 (<**0.001**) ***	10.6 (**0.001**) **	7.10 (<**0.001**) ***
RoBERTa Surprisal	1.77 (0.184)	45.9 (<**0.001**) ***	2.54 (0.111)	8.37 (<**0.001**) ***
GPT-3 Surprisal (dc)	11.1 (**0.001**) **	36.8 (<**0.001**) ***	3.59 (0.058) .	10.2 (<**0.001**) ***
GPT-3 Surprisal (dc+rc)	7.48 (**0.006**) **	39.4 (<**0.001**) ***	0.192 (0.661)	10.7 (<**0.001**) ***

Table 1: Results of Likelihood Ratio Tests comparing Base models (random effects + log frequency), NLM models (Base + NLM Measure), and Full models (Base + NLM + Condition) in predicting human sensibility judgements. For the Afforded vs Nonafforded comparison, three NLM measures significantly predicted human sensibility ratings: BERT Surprisal, GPT-3 Surprisal (dc) and GPT-3 Surprisal (dc+rc), see §NLM Measures. The full model, with condition as an additional predictor, explained significant variance on top of all NLM measures, indicating that none of the NLM measures accounted for all of the variance caused by Affordedness. Only one NLM measure (BERT Surprisal) improved fit over the Base model in the Afforded vs Related comparison. Again the inclusion of condition improved model fit over each NLM-only model.

judgements in Study 2 as evidence for a limit on the sufficiency of distributional information to explain word meaning. This interpretation is consistent with the proposal in Glenberg and Robertson (2000) that human comprehenders draw on their embodied experience of the world to simulate how the affordances of objects will mesh with novel actions. The NLMs assessed in this work clearly lack this bodily experience, which might explain their inability to account for the influence of affordances. The results are equally consistent, however, with proposals that NLMs lack other crucial capacities: innate knowledge about concepts (Fodor, 1975), a deliberative reasoning faculty (Russin, O'Reilly, & Bengio, 2020), or an internal workspace to store and retrieve intermediate products of their cognition (VanRullen & Kanai, 2021). In order to test whether humans owe the affordance-sensitivity highlighted by these results to their embodiment, more evidence is needed. Suitable experiments might try to interfere with participants' sensitivity to affordances, by limiting their embodied experience with relevant objects, or burdening non-linguistic systems that are theorized to play a role in the deployment of embodied information during language comprehension (Ostarek & Bottini, 2021).

If the performance limitations of NLMs are indeed a result of their lack of embodied experience, how might we augment models to endow them with the relevant capabilities? In their initial presentation of LSA, Landauer and Dumais (1997) argued that distributional linguistic representations can be grounded in perceptual experience of the world by training models on multimodal data: "Indeed, if one judiciously added numerous pictures of scenes with and without rabbits... LSA could easily learn that the words rabbit and hare go with pictures containing rabbits" (p. 227). While this task may not be as easy as Landauer and Dumais had originally envisioned, there are a number of promising avenues for improving the sensitivity of these models to the physical affordances of objects. One approach is to enrich distributional representations with perceptual norms generated by human participants (Andrews, Vigliocco, & Vinson, 2009; Davis & Yee, 2021; Hoffman, McClelland, & Lambon Ralph, 2018). Johns and Jones (2012) applied this technique to the Glenberg and Robertson (2000) stimuli used here and reproduced the pattern observed in human data. Similarity between verbs and objects in the critical sentence was greatest for the Related, intermediate for Afforded, and smallest for the Nonafforded scenarios. This result provides compelling evidence that the information which LSA lacked in order to make the Afforded/Nonafforded distinction could have been perceptual in nature. This result also highlights that there are multiple ways to acquire the same knowledge: evidence that distributional information is sufficient for a task does not imply that humans use it. Other promising approaches include having models learn joint representations over linguistic and perceptual input, for instance by learning to match video frames to an audio transcript (Zellers, Lu, et al., 2021) or having language agents interact with real or simulated environments, on the grounds that the sensorimotor contingencies that humans learn through their interaction with the world form the basis for the meanings they assign (Bisk et al., 2020). A final possibility is that models lack—beyond relevant data—a capacity to simulate novel events. Future work should explore training models to generate dynamic simulations of described events, inspired by evidence for similar capacities in humans (Battaglia, Hamrick, & Tenenbaum, 2013; Zellers, Holtzman, et al., 2021)

The results presented here show that distributional methods have progressed substantially in the last two decades at exploiting diffuse linguistic cues to learn nonobvious relationships between agents, objects, and actions. However, these models are still far from being able to explain the rich influence of these subtleties on human comprehenders. This gap will continue to be closed by more powerful models. However, these results also encourage us to look elsewhere—to our embodiment and the world—to explain human language comprehension, and to engineer machines that think like us.

Acknowledgements

We thank Arthur Glenberg for sharing the stimuli for the original study and providing feedback on the research, the UC San Diego Brain and Cognition Lab for an insightful discussion about this work, and four anonymous reviewers for their thoughtful comments.

References

Andrews, M., Vigliocco, G., & Vinson, D. (2009). Integrating experiential and distributional data to learn semantic representations. *Psychological review*, *116*(3), 463. doi: 10.1037/a0016261

Barsalou, L. W. (1999). Perceptions of perceptual symbols. *Behavioral and Brain Sciences*, *22*(4), 637–660.

Battaglia, P. W., Hamrick, J. B., & Tenenbaum, J. B. (2013). Simulation as an engine of physical scene understanding. *Proceedings of the National Academy of Sciences*, *110*(45), 18327–18332. doi: 10.1073/pnas.1306572110

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198). doi: 10.18653/v1/2020.acl-main.463

Bisk, Y., Holtzman, A., Thomason, J., Andreas, J., Bengio, Y., Chai, J., ... Nisnevich, A. (2020). Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8718–8735). doi: 10.18653/v1/2020.emnlp-main.703

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... Amodei, D. (2020). Language Models are Few-Shot Learners. In *Advances in Neural Information Processing Systems* (Vol. 33, pp. 1877–1901). Curran Associates, Inc.

Davis, C. P., & Yee, E. (2021). Building semantic memory from embodied and distributional language experience. *Wiley Interdisciplinary Reviews: Cognitive Science*, *12*(5), e1555. doi: 10.1002/wcs.1555

Davis, E., & Marcus, G. (2015). Commonsense reasoning and commonsense knowledge in artificial intelligence. *Communications of the ACM*, *58*(9), 92–103. doi: 10.1145/2701413

De Leeuw, J. R. (2015). jsPsych: A JavaScript library for creating behavioral experiments in a Web browser. *Behavior research methods*, *47*(1), 1–12.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. 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. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics.

Elman, J. L. (1990). Finding structure in time. *Cognitive science*, *14*(2), 179–211. doi: 10.1207/s15516709cog1402_1

Firth, J. (1957). *A synopsis of linguistic theory 1930-1955 in studies in linguistic analysis, Philological Society.* Oxford.

Fodor, J. A. (1975). *The language of thought* (Vol. 5). Harvard university press.

Forbes, M., Holtzman, A., & Choi, Y. (2019). Do Neural Language Representations Learn Physical Commonsense? In *The 41st Annual Meeting of the Cognitive Science Society.* Montreal, Quebec, Canada.

Frank, S. L., Otten, L. J., Galli, G., & Vigliocco, G. (2015). The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, *140*, 1–11. doi: 10.1016/j.bandl.2014.10.006

Gibson, J. J. (1979). The ecological approach to visual perception.

Glenberg, A. M., & Robertson, D. A. (2000). Symbol Grounding and Meaning: A Comparison of High-Dimensional and Embodied Theories of Meaning. *Journal of Memory and Language*, *43*(3), 379–401. doi: 10.1006/jmla.2000.2714

Goodkind, A., & Bicknell, K. (2018). Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th workshop on cognitive modeling and computational linguistics (CMCL 2018)* (pp. 10–18). doi: 10.18653/v1/W18-0102

Gordon, J., & Van Durme, B. (2013). Reporting bias and knowledge acquisition. In *Proceedings of the 2013 workshop on Automated knowledge base construction* (pp. 25–30). doi: 10.1145/2509558.2509563

Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena*, *42*(1-3), 335–346. doi: 10.1016/0167-2789(90)90087-6

Harris, Z. S. (1954). Distributional structure. *Word*, *10*(2-3), 146–162. doi: 10.1080/00437956.1954.11659520

Hoffman, P., McClelland, J. L., & Lambon Ralph, M. A. (2018). Concepts, control, and context: A connectionist account of normal and disordered semantic cognition. *Psychological review*, *125*(3), 293. doi: 10.1037/rev0000094

Johns, B. T., & Jones, M. N. (2012). Perceptual Inference Through Global Lexical Similarity: Topics in Cognitive Science. *Topics in Cognitive Science*, *4*(1), 103–120. doi: 10.1111/j.1756-8765.2011.01176.x

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., ... Amodei, D. (2020). Scaling Laws for Neural Language Models. *arXiv:2001.08361 [cs, stat]*.

Kocijan, V., Davis, E., Lukasiewicz, T., Marcus, G., & Morgenstern, L. (2022). The Defeat of the Winograd Schema Challenge. *arXiv:2201.02387 [cs]*.

Landauer, T. K., & Dumais, S. T. (1997). A solution to Plato's problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological review*, *104*(2), 211. doi: 10.1037/0033-295X.104.2.211

Lenci, A. (2018). Distributional Models of Word Meaning. *Annual Review of Linguistics*, *4*(1), 151–171. doi: 10.1146/annurev-linguistics-030514-125254

Lewis, M., Zettersten, M., & Lupyan, G. (2019). Distributional semantics as a source of visual knowledge. *Proceedings of the National Academy of Sciences*, *116*(39), 19237–19238. doi: 10.1073/pnas.1910148116

Li, J., & Joanisse, M. F. (2021). Word senses as clusters of meaning modulations: A computational model of polysemy. *Cognitive Science*, *45*(4), e12955. doi: 10.1111/cogs.12955

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

Mahon, B. Z. (2015). What is embodied about cognition? *Language, cognition and neuroscience*, *30*(4), 420–429. doi: 10.1080/23273798.2014.987791

Michaelov, J. A., Coulson, S., & Bergen, B. K. (2021). So Cloze yet so Far: N400 Amplitude is Better Predicted by Distributional Information than Human Predictability Judgements. *arXiv preprint arXiv:2109.01226*.

Ostarek, M., & Bottini, R. (2021). Towards Strong Inference in Research on Embodiment – Possibilities and Limitations of Causal Paradigms. *Journal of Cognition*, *4*(1), 5. doi: 10.5334/joc.139

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A Primer in BERTology: What We Know About How BERT Works. *Transactions of the Association for Computational Linguistics*, *8*, 842–866. doi: 10.1162/tacl_a_00349

Russin, J., O'Reilly, R. C., & Bengio, Y. (2020). Deep learning needs a prefrontal cortex. *Work Bridging AI Cogn Sci*, *107*, 603–616.

Schrimpf, M., Blank, I. A., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., ... Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*(45). doi: 10.1073/pnas.2105646118

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and brain sciences*, *3*(3), 417–424. doi: 10.1017/S0140525X00005756

Shain, C. (2019). A large-scale study of the effects of word frequency and predictability in naturalistic reading. 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)*. doi: 10.18653/v1/N19-1413

Tenney, I., Das, D., & Pavlick, E. (2019). BERT Rediscovers the Classical NLP Pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics* (pp. 4593–4601). Florence, Italy: Association for Computational Linguistics. doi: 10.18653/v1/P19-1452

Trott, S., & Bergen, B. (2021). RAW-C: Relatedness of Ambiguous Words in Context (A New Lexical Resource for English). 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. 7077–7087). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.550

VanRullen, R., & Kanai, R. (2021). Deep learning and the Global Workspace Theory. *Trends in Neurosciences*, *44*(9), 692–704. doi: 10.1016/j.tins.2021.04.005

Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., ... Bowman, S. (2019). SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 3266–3280). Curran Associates, Inc.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., & Bowman, S. R. (2019). GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. In *International Conference on Learning Representations*.

Wittgenstein, L. (1953). *Philosophical Investigations*. Oxford: Blackwell.

Wolf, T., Chaumond, J., Debut, L., Sanh, V., Delangue, C., Moi, A., ... Shleifer, S. (2020). Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations* (pp. 38–45). doi: 10.18653/v1/2020.emnlp-demos.6

Zellers, R., Holtzman, A., Peters, M., Mottaghi, R., Kembhavi, A., Farhadi, A., & Choi, Y. (2021). PIGLeT: Language Grounding Through Neuro-Symbolic Interaction in a 3D World. 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. 2040–2050). Online: Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.159

Zellers, R., Lu, X., Hessel, J., Yu, Y., Park, J. S., Cao, J., ... Choi, Y. (2021). MERLOT: Multimodal Neural Script Knowledge Models. In *Advances in Neural Information Processing Systems* (Vol. 34, pp. 23634–23651). Curran Associates, Inc.