aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Current Directions in Psychological Science
2024, Vol. 33(3) 181–189

DOI: 10.1177/09637214241242746
www.psychologicalscience.org/CDPS
Sage

What Can Language Models Tell Us About Human Cognition?

Louise Connell and Dermot Lynott
Department of Psychology, Maynooth University

Abstract

Language models are a rapidly developing field of artificial intelligence with enormous potential to improve our understanding of human cognition. However, many popular language models are cognitively implausible on multiple fronts. For language models to offer plausible insights into human cognitive processing, they should implement a transparent and cognitively plausible learning mechanism, train on a quantity of text that is achievable in a human's lifetime of language exposure, and not assume to represent all of word meaning. When care is taken to create plausible language models within these constraints, they can be a powerful tool in uncovering the nature and scope of how language shapes semantic knowledge. The distributional relationships between words, which humans represent in memory as linguistic distributional knowledge, allow people to represent and process semantic information flexibly, robustly, and efficiently.

Keywords

language models, linguistic distributional knowledge, semantics, cognitive plausibility

Imagine you overhear someone talking about their holiday plans and catch the words "beach," "ice cream," and "swimming." It is not difficult to realize that this person is most likely talking about a summer rather than winter break. But how can you reach this conclusion? One way is to consider each activity in turn—going to the beach, eating ice cream, and swimming—and conclude that they are all far more likely to occur in the summertime than in wintertime. Alternatively, one might realize that the words you overheard all tend to occur in the context of summer far more often than the context of winter. Rather than needing to engage in a process of deep reasoning to figure out the timing of holiday plans, it might simply be enough for the word "summer" to come to mind.

What this example shows is that the distributional relationships between words are useful. Language is full of statistical patterns of how words appear in relation to one other, which people can learn passively from cumulative language exposure (e.g., Savic et al., 2022) and represent in memory as linguistic distributional knowledge (Wingfield & Connell, 2022). To a large extent, the distributional relationships in language reflect the structure of the world (Louwerse, 2011; Riordan & Jones, 2011). For example, people often go to the beach in summer and therefore often talk about the beach and summer in the same context. As a result, language can capture a wide variety of semantic relations that are useful in cognitive processing (Brown et al., 2023; Wingfield & Connell, 2022; see Table 1). For instance, *syntagmatic relations* are learned when two words occupy complementary syntactic positions in the same sentence (e.g., "she has brown eyes" links brown and eyes syntagmatically). *Paradigmatic relations* are learned when two words occupy the same syntactic position across different sentences (e.g., "she has brown eyes" and "he has blue eyes" link brown and blue paradigmatically). Last, bag-of-words relations are learned across high-level situations or themes outside syntactic roles (e.g., the tale of Newton's discovery of universal gravitation links apple and gravity in a bag-of-words relation).

Types of Language Model

Computational language models readily capture these distributional relations between words when they are trained on a large corpus (i.e., collection of texts). Mostly emerging from artificial intelligence (AI) and

Corresponding Author:
Louise Connell, Department of Psychology, Maynooth University
Email: louise.connell@mu.ie

Table 1. Types of Distributional Relationship Between Words and Examples of Some of the Semantic Relations They Capture, as Outlined by Wingfield and Connell (2022)

Distributional relation	Learned from	Semantic relation	Examples
Syntagmatic	Complementary syntactic positions in same sentence	Concept-property	eyes-brown, childhood-happy
		Constituent (part-of)	dog-tail, car-wheels
		Compositional	river-water, cake-chocolate
		Locative	egg-nest, boat-lake
		Temporal	beach-summer, breakfast-morning
		Functional (instrumental)	ball-throw, chair-sit
		Agent-patient	dog-ball, chef-meal
		Agent-action	cat-meow, customer-pay
Paradigmatic	Same syntactic position in different sentences	Synonym	blue-azure, run-sprint
		Antonym	hot-cold, rise-fall
		Shared category	cat-dog, happy-angry
		Taxonomic class	dog-animal, chair-furniture
Bag of words	High-level situation (not governed by syntax)	Abstracted-thematic	apple-gravity, physics-proton, castle-money, computer-internet

computational linguistics research, there are several different families of language models (see Table 2). *N*-gram models simply count how often various words co-occur within a certain window around the target word. Given two words, they score how frequently those words appear in the same context across language. Examples include Google's Web 1T 5-gram corpus (Brants & Franz, 2006), but they can be created easily from any corpus of text. Count-vector models accumulate co-occurrences in a similar way to *n*-gram models. Given two words, they use vector geometry to score how close (similar) the contexts of those words are. Examples include latent semantic analysis (Landauer & Dumais, 1997) and bound encoding of the aggregate language environment, or BEAGLE (Jones & Mewhort, 2007), but like *n*-gram models they can be created easily from any corpus. Predict models operate quite differently by training a neural network to predict a target word from a provided context (or the reverse). Given two words, they use vector geometry to score how similar the network's representations of those words are. Popular examples include word2vec (Mikolov et al., 2013) and GloVe (which combines aspects of count-vector and predict models; Pennington et al., 2014), both of which can be used in off-the-shelf (pretrained) versions or trained on another corpus. Finally, transformer models train a very large neural network to predict a target word (or group of words) from a given context using a complex array of settings to prioritize more important words and then typically retrain the model to fine-tune its application to a particular task. Rather than creating a representation for each word encountered, they instead treat a word separately per context it appears in. Given two words, different transformer models have different methods of collapsing their contextual representations to score them. Examples include bidirectional encoder representations from transformers, or BERT (Devlin et al., 2019), and OpenAI's generative pretrained transformer (GPT; Radford et al., 2018). Their size and complexity (i.e., millions to trillions of parameters) makes them expensive and time-consuming to train, and so they are most often used in off-the-shelf versions.

Language Models as Plausible Cognitive Models

The large differences between various forms of language models raises the question: Which of these models do what humans do? Typically, cognitive scientists study language models by comparing their performance on particular cognitive tasks with human performance on the same tasks. For example, a model's scores between word pairs (e.g., "beach" and "summer") might provide a good fit to human similarity ratings for the words, or to the cognitive-processing effort involved in reading the words in sequence (i.e., semantic priming). If a language model is to offer any insights into human cognition, then it must be cognitively plausible. In other words, the way in which it learns, processes, and represents information in cognitive tasks—that is, what Marr (1982) called the algorithmic level of analysis—must plausibly correspond to what humans can do. By meeting the constraints of cognitive plausibility, a language model can thereby act as a cognitive model of how humans learn, process, and represent information in a given task. Researchers have considered a number of different criteria when considering the cognitive plausibility of language models. We concentrate here on learning mechanisms, corpus size, and grounding.

Table 2. Families of Language Models With Sample Models Per Family and Comparison to Humans in Terms of Their Cognitive Plausibility and Performance in Cognitive Tasks

Model family	Sample model	Similarities to humans	Differences from humans
N-gram (also known as direct co-occurrence or first-order distributional)	Web 1T 5-gram (Brants & Franz, 2006)	Plausible learning (Hebbian); performs well on tasks such as semantic relatedness, spatial iconicity (Louwerse & Jeuniaux, 2010), categorization, and spatial relatedness (Louwerse, 2011)	Implausible corpus size far exceeds human experience; not grounded
	PPMI 6-gram (Wingfield & Connell, 2022)	Plausible learning (Hebbian); plausible corpus size; performs well on semantic relatedness, thematic relatedness, and semantic priming and moderately well on semantic similarity; grounded in sensorimotor experience in later implementation, in which it performs well on category production (Banks et al., 2021)	Systematic analysis of ungrounded model shows poor performance on synonym judgment and concrete-abstract semantic decisions (Wingfield & Connell, 2022)
Count vector (also known as indirect co-occurrence or second-order distributional)	BEAGLE (Jones & Mewhort, 2007)	Plausible learning (Hebbian); plausible corpus size in some implementations (Hills et al., 2012; Johns et al., 2019), in which it performs well on synonym judgment, category production, semantic similarity, and semantic priming	Not grounded (but see Johns & Jones, 2012)
	PPMI count vector (e.g., Bullinaria & Levy, 2007)	Plausible learning (Hebbian); plausible corpus size in some implementations (Brown et al., 2023; Mandera et al., 2017; Wingfield & Connell, 2022), in which it performs well on synonym judgment, semantic similarity, semantic priming, and concrete-abstract semantic decisions and moderately well on semantic relatedness, thematic relatedness, categorization, and free association; grounded in vision in some variants (e.g., Bruni et al., 2014), in which it performs well on semantic relatedness and categorization	Implausible corpus size and/or model not grounded in a given implementation; systematic analysis of ungrounded model shows weak performance on verbal analogies (Lenci et al., 2022)
Predict (also known as word embedding)	GloVe (Pennington et al., 2014)[a]	Some plausible learning (Hebbian); plausible corpus size in some implementations (e.g., Brown et al., 2023), in which it performs well on semantic relatedness and thematic relatedness and moderately well at semantic similarity; grounded in vision in some implementations (e.g., Derby et al., 2018; Shahmohammadi et al., 2023), in which it performs well on semantic similarity and property verification	Some controversial learning (not backpropagation but still error-driven; see Kumar, 2021); implausible corpus size and/or model not grounded in a given implementation; systematic analysis of ungrounded model shows weak performance on verbal analogies (Lenci et al., 2022)

(continued)

Table 2. *(continued)*

Model family	Sample model	Similarities to humans	Differences from humans
	word2vec (Mikolov et al., 2013)	Plausible corpus size in some implementations (e.g., Brown et al., 2023; Mandera et al., 2017; Wingfield & Connell, 2022), in which it performs well on synonym judgment, semantic similarity, semantic relatedness, thematic relatedness, and free association and moderately well on semantic priming; grounded in vision in some implementations (e.g., Derby et al., 2018), in which it performs well on semantic similarity and property verification	Controversial learning via backpropagation; implausible corpus size and/or model not grounded in a given implementation; systematic analysis of ungrounded model shows weak performance on concrete-abstract semantic decisions (Wingfield & Connell, 2022) and verbal analogies (Lenci et al., 2022)
Transformer (also known as deep learning or large language models)	GPT (Radford et al., 2018)[b]	Grounded in vision in most recent implementation (GPT-4), in which it performs well on metaphor interpretation and verbal analogies and moderately well on semantic classification and semantic odd-one-out tasks (Loconte et al., 2023)	Implausible black-box architecture; controversial learning via backpropagation; implausible learning via human feedback; implausible corpus size far exceeds human experience; systematic analysis shows poor performance on semantic absurdity detection, social-norm violation, and spatial planning (Loconte et al., 2023)
	BERT (Devlin et al., 2019)	Performs well on semantic similarity and semantic relatedness and moderately well on synonym judgment and categorization (Lenci et al., 2022); plausible corpus size for children (not adults) in some implementations (Warstadt et al., 2023), in which it performs well on semantic similarity and semantic relatedness; grounded in vision in some implementations (e.g., Shahmohammadi et al., 2023), in which it performs well on semantic similarity and semantic relatedness	Implausible black-box architecture; controversial learning via backpropagation; implausible corpus size and/or model not grounded in a given implementation; systematic analysis of ungrounded model shows poor performance on verbal analogies (Lenci et al., 2022)

Note: The tasks listed for a given model are cognitive tasks that involve semantic processing; natural language processing model evaluation tasks such as sentiment analysis or grammaticality are not included. PPMI = positive pointwise mutual information; BEAGLE = bound encoding of the aggregate language environment; GPT = generative pretrained transformer; BERT = bidirectional encoder representations from transformers. [a]GloVe combines aspects of count-vector and predict models. [b]Operational details (e.g., algorithm, code, training data) of this proprietary model have been increasingly withheld in later versions.

Learning mechanisms

One key criterion concerns the mechanisms of learning and operation. Broadly, both *n*-gram and count-vector models are plausible in their use of error-free, Hebbian learning, in which information gradually accumulates to strengthen associative connections between words (Davis & Yee, 2021; Kumar, 2021). *N*-gram models, despite their simplicity, perform well in tasks that rely predominantly on syntagmatic relations (e.g., thematic relatedness, spatial relatedness, word association; see Table 2) because such words, by definition, appear in the same context. They also do well on some tasks that rely wholly or predominantly on paradigmatic relations (e.g., semantic similarity, semantic relatedness, category production) because these words often appear together in context (e.g., "beach in summer"). However, their poor performance in concrete/abstract semantic decisions (e.g., determining whether "beach" is a concrete or abstract word) suggests they are relatively insensitive

to bag-of-words relations (Wingfield & Connell, 2022). Count-vector models, on the other hand, do well on a broader range of cognitive tasks, including paradigmatic tasks (e.g., synonym judgment, semantic similarity, categorization), bag-of-words tasks (e.g., concrete/abstract semantic decisions), and tasks that typically rely on a mix of distributional relations (e.g., semantic priming). They can learn such relations because the words share many similar contexts (e.g., "beach" and "mountain" both appear in discussions of vacations and landscapes). They also perform reasonably well—if not quite as strongly—on syntagmatic tasks (e.g., thematic relatedness, property verification, word association) because syntagmatically related words often share similar contexts (e.g., "beach" and "summer" both appear in discussions of vacations).

Predict models are more controversial in their use of error-driven, supervised learning, particularly in the form of *backpropagation*, in which the difference between the model output and the correct answer is fed back into the network until it gradually learns the desired patterns of associations between words. Although some researchers consider it plausible for a model to learn from its errors because it is consistent with the principles of reinforcement learning (Mandera et al., 2017), others disagree and note that backpropagation leads a model to catastrophically "forget" old information in a way that is implausible for learning language and semantics (Mannering & Jones, 2021). In performance terms, predict models tend to do well on many of the same tasks as count-vector models (see Table 2), offering a better fit to human performance on paradigmatic tasks but a worse fit on bag-of-words tasks. They can learn such relations by abstracting across similar contexts (much like count-vector models), but their optimization for paradigmatic relations appears to come at the cost of more general bag-of-words relations. Indeed, predict models are not overall superior to count-vector models when a wide range of tasks are considered (Lenci et al., 2022; Wingfield & Connell, 2022; see also Brown et al., 2023), which suggests the controversy of their learning mechanisms can be avoided if researchers so wish. Such variable performance across model families with (arguably) plausible learning mechanisms suggests that humans represent multiple forms of linguistic distributional knowledge in semantic memory, or use multiple mechanisms to flexibly access such knowledge, to process semantic information according to the cognitive requirements of a task (Wingfield & Connell, 2022; see also Kumar, 2021).

Transformer models, however, are the least plausible on the issue of learning mechanisms because—in addition to learning via backpropagation and in some cases via human feedback during fine-tuning (e.g., GPT)—they have the additional problem of being a black box in processing terms. That is, although their output appears to emulate human behavior in some cognitive tasks, their high complexity (i.e., millions to trillions of parameter settings) means how and why they work as they do remain opaque. Indeed, so uncertain is the current understanding of the latest generation of transformer models that recent work has begun to use cognitive tests to try to determine what is going on inside them (e.g., Binz & Schulz, 2023). However, others have criticized such approaches because they move the goalposts from trying to understand how humans operate to trying to understand why a model behaves like a human (Johns et al., 2023). More generally, any post hoc explanations of black-box models are likely to be inadequate at best and misleading at worst (Rudin, 2019). Both these criticisms call into question the suitability of using black-box transformer models for generating theories about human cognition. As a result, although transformer models can perform at human levels in complex tasks such as metaphor comprehension and verbal analogies (Loconte et al., 2023; see also Table 2), it is unclear what can reasonably be concluded about human cognition from such reports.

Corpus size

A further criterion of cognitive plausibility relates to the size of the corpus used to train the model. With enormous quantities of text available on the Internet, it is very easy to allow language models to learn distributional relationships across billions or trillions of words. However, if a model can approximate human behavior using only a corpus that is many times larger—even orders of magnitude larger—than that accumulated in a human's lifetime of language experience, then it is not a plausible model of how linguistic distributional knowledge works in the human mind (Johns et al., 2023; Warstadt et al., 2023; Wingfield & Connell, 2022). Although people gain a lot of their language experience through spoken language, both in terms of conversation and media consumption (e.g., watching television and movies), the fastest way to accumulate language experience is through reading written texts. For instance, a full day of social interaction accumulates around 32,000 words, whereas a typical literate adult could accumulate the same number of words in a couple of hours of reading (see Brysbaert et al., 2016). On this basis, Wingfield and Connell (2022) proposed that the typical bounds of language experience for an adult speaker of English are approximately 200 million words (for a 20-year-old who rarely reads) to 2 billion words (for an avid reader over 60 years old). By these standards,

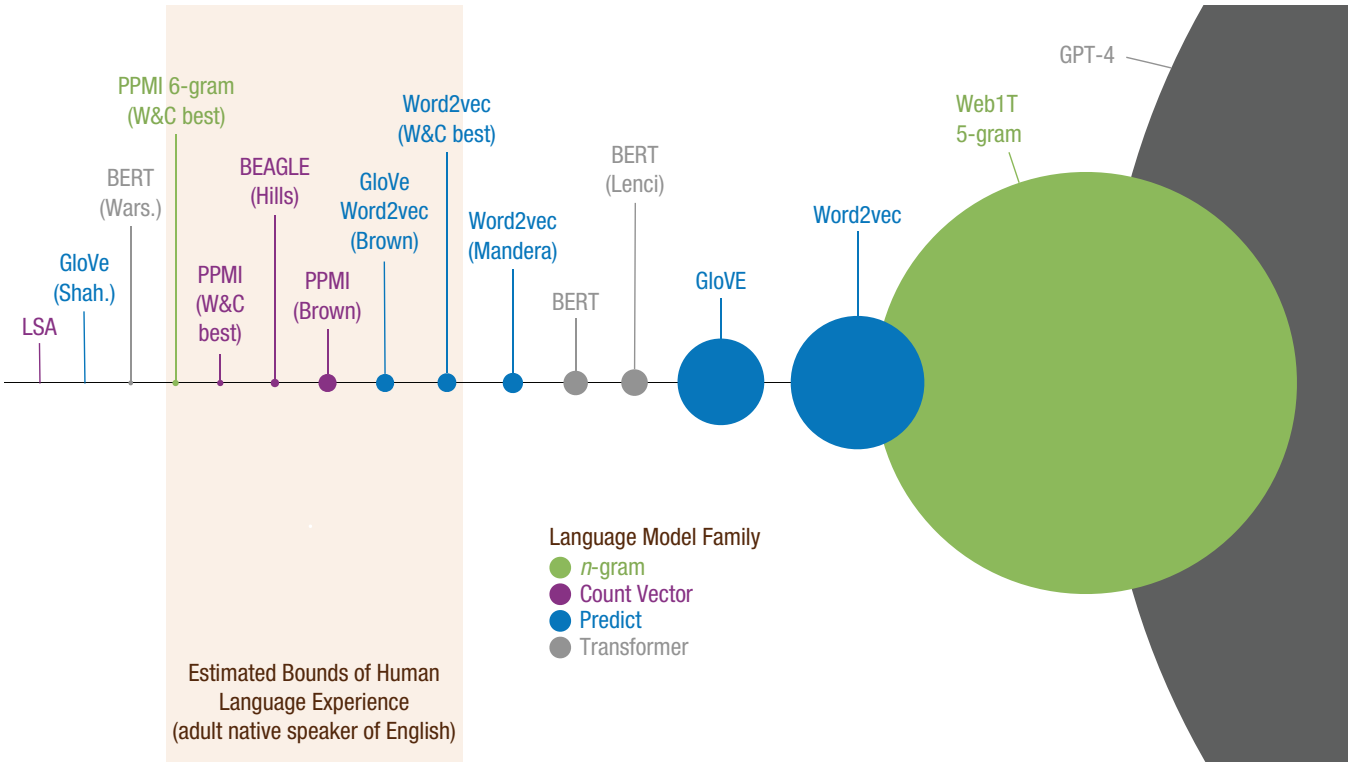

Fig. 1. Visualization of corpus size used to train a sample of language models, with reference to the cognitively plausible range of adult language experience. The corpus size is represented as the area of the circle. Models are listed by the English-language corpus size in the default (original) instantiation unless otherwise specified. Brown = Brown et al. (2023); Hills = Hills et al. (2012); Lenci = Lenci et al. (2022); Mandera = Mandera et al. (2017); Shah. = Shahmohammadi et al. (2023); W&C = Wingfield and Connell (2022); Wars. = Warstadt et al. (2023). This figure and its underlying calculations are available under a CC-BY 4.0 licence at https://osf.io/kgwsc/.

many popular language models are based on implausibly large quantities of text (see Fig. 1).

However, some language models have constrained their training corpus to cognitively plausible sizes. *N*-gram and count-vector models have demonstrated some of their best performance within these constraints across a wide range of cognitive tasks (see Table 2), although there is some evidence that using a high-quality corpus (i.e., few typos, representative content) is also important to maintaining strong performance (Johns et al., 2019; Wingfield & Connell, 2022). Predict models perform better with corpus sizes that exceed the extent of human language experience (e.g., Pennington et al., 2014; Shahmohammadi et al., 2023) but still do well within plausible limits, particularly for paradigmatic tasks (Wingfield & Connell, 2022). The vast majority of work using transformer models is based on training data that far exceed human language experience, which—coupled with their implausible architecture—makes it difficult to draw sound conclusions about human cognition from their successes and failures on cognitive tasks (see Table 2). Recent efforts have focused on approximating the language experience of children by limiting the training corpus to 100 million words (for review, see Warstadt et al., 2023), which produces generally good performance for semantic similarity (i.e., a paradigmatic task) but has not been subject to more systematic testing of cognitive tasks that span multiple distributional relations.

Nonetheless, there is good evidence that language models (if not necessarily transformer models) perform strongly at cognitive tasks when constrained to the language experience of an adult human, which suggests that vast quantities of text are not necessary to learn the semantic information of linguistic distributional knowledge. Investigating how much language experience is required for language models to capture different kinds of distributional relation could offer valuable insights into the developmental trajectory of linguistic distributional knowledge.

Grounding

Finally, one of the most important criteria in model plausibility relates to the representation of meaning (i.e., conceptual or semantic representations). Language models are obviously based on language, in which the meaning of a given word is effectively represented in terms of other words. It has long been recognized in the cognitive sciences that such circular definitions of meaning are deficient. Words cannot derive their meaning solely

via associations with one another in a self-contained system but must instead connect to their counterparts in the physical world (i.e., the symbol grounding problem; Harnad, 1990). As humans, we do not have a symbol grounding problem because the meaning of words is not only based on language but also on our experience of perceiving and interacting with the world around us (Barsalou et al., 2008; Connell & Lynott, 2014). The implication is that language models cannot be expected to account for all meaning representation, and therefore the information they capture is—at best—a partial implementation of how humans represent and process semantics. If a language model aims to account for all of the semantic processing in a task, then it is not a plausible account of how humans perform that task. However, if a language model is instead assumed to implement an essential component of word meaning, which is ultimately grounded in a complementary component of sensorimotor and other experiences, then it offers a cognitively plausible account (Davis & Yee, 2021; Johns et al., 2023; Wingfield & Connell, 2022).

Indeed, several approaches to cognitive modeling seek to address this issue directly by combining language models with some form of sensorimotor grounding, showing that this combination performs better than a stand-alone language or sensorimotor model (e.g., Banks et al., 2021; Bruni et al., 2014; Johns & Jones, 2012; for examples, see Table 2). Multidimensional profiles of sensorimotor experience have been used to ground *n*-gram models, which enhances performance in category production (e.g., how many types of "animal" one can name in 60 s; Banks et al., 2021). Count-vector, predict, and transformer models have all been grounded in visual information to similar effect in tasks such as semantic similarity or property verification (e.g., Derby et al., 2018; Shahmohammadi et al., 2023). Although grounding in vision alone is less cognitively plausible than incorporating multiple sensorimotor sources, it is arguably sufficient to avoid the grounding problem because word-to-word connections mean that not every word needs to be individually grounded (Louwerse, 2011). Overall, when retrieving or comparing concepts, people appear to rely both on activating a concept name via distributional relationships and activating a detailed representation of a concept based on sensorimotor experience, and having both forms of semantic information available leads to better performance.

Such findings support linguistic-sensorimotor (also known as linguistic-embodied) theories of cognition that propose conceptual processing is based on two components: linguistic distributional knowledge and grounded knowledge of the physical world (Barsalou et al., 2008; Connell & Lynott, 2014; Louwerse, 2011). Because the distributional relationships in language largely (although not perfectly; see Connell, 2019) reflect the structure of the world, it means that semantic information in a particular task can often come from either linguistic distributional knowledge or sensorimotor experience. For example, when asked to list different types of animals, one could start naming "dog," "cat," and so on either because these words are close in distributional terms (i.e., "animal" and "dog" are paradigmatically related) or because their sensorimotor representations are close to that of "animal" (e.g., animals and dogs are experienced in similar ways via perception and action). According to linguistic-sensorimotor theories, when both linguistic distributional knowledge and sensorimotor experience can provide similar information in a task, the former offers a semantic heuristic (i.e., adequate means of achieving a goal) that makes cognitive processing more efficient. Such redundancy between linguistic distributional knowledge and sensorimotor experience also allows language models to predict congenitally blind participants' judgments about object color and thus may provide a means for people with congenital sensory impairments to acquire semantic knowledge about things they cannot perceive (for review, see Campbell & Bergelson, 2022).

Finally, recent work has found that visual grounding improves the performance of predict models only if they are trained on a very small corpus, whereas it makes little difference to models with training data that far exceed human language experience (Shahmohammadi et al., 2023). This finding suggests that the implausibly large corpus sizes found in many language models may be unnecessary if they instead were grounded in sensorimotor experience. It also leads us to speculate that linguistic distributional knowledge and sensorimotor knowledge might interact over the life span, in which sensorimotor knowledge compensates for limited language experience in the early years but linguistic distributional knowledge plays an increasingly large role as language experience accumulates with age. However, the predict models in this study were tested only on paradigmatic tasks (i.e., semantic similarity and semantic relatedness), and because predict models tend to perform poorly at tasks using bag-of-words relations (Wingfield & Connell, 2022), one cannot yet assume that the same findings would apply to all cognitive tasks. Future work should examine more closely how corpus size and grounding interact in language models to drive performance across a comprehensive suite of cognitive tasks.

Summary

Language models are a rapidly developing field of AI with enormous potential to improve our understanding of human cognition, but they must not be used blindly. Many language models that have recently captured public imagination, such as ChatGPT, are cognitively

implausible on multiple fronts, which should come as no surprise because they are not designed to be cognitive models. Disregarding cognitive plausibility is a legitimate approach in pure AI research that is interested only in improving language-model performance to engineer a better tool (e.g., a better chatbot, a better text classifier, a better multilingual translator). However, cognitive modeling research is interested in replicating human cognitive processing with all its limitations and errors and therefore must—by definition—attend to the cognitive plausibility of the model in question to draw meaningful conclusions.

When care is taken to create language models that plausibly approximate human constraints of learning and representation, they can be a powerful tool in understanding the nature and scope of how language shapes semantic knowledge. Findings from such language models suggest that linguistic distributional knowledge enhances the robustness of learning, representing, and processing semantic information. Future use of plausible language models can help researchers to determine the full extent—and the limitations—of linguistic distributional knowledge in cognition.

Recommended Reading

Davis, C. P., & Yee, E. (2021). (See References). Overview of how semantic memory arises from both distributional language experience and sensorimotor experience.

Landauer, T. K., & Dumais, S. T. (1997). (See References). Seminal article that outlines how rich semantic information can emerge from latent structure in language experience.

Savic, O., Unger, L., & Sloutsky, V. M. (2022). (See References). Empirical demonstration of people learning syntagmatic and paradigmatic relations between words via implicit statistical learning.

Wingfield, C., & Connell, L. (2022). (See References). Review and comparison of language models and their cognitive plausibility and utility across different cognitive tasks.

Transparency

Action Editor: Robert L. Goldstone
Editor: Robert L. Goldstone
Declaration of Conflicting Interests
The author(s) declared that there were no conflicts of interest with respect to the authorship or the publication of this article.
Funding
This work was supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement 682848 to L. Connell).

ORCID iD

Louise Connell https://orcid.org/0000-0002-5291-5267

References

Banks, B., Wingfield, C., & Connell, L. (2021). Linguistic distributional knowledge and sensorimotor grounding both contribute to semantic category production. *Cognitive Science*, *45*(10), Article e13055. https://doi.org/10.1111/cogs.13055

Barsalou, L. W., Santos, A., Simmons, W. K., & Wilson, C. D. (2008). Language and simulation in conceptual processing. In M. De Vega, A. M. Glenberg, & A. C. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition* (pp. 245–283). Oxford University Press.

Binz, M., & Schulz, E. (2023). Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences, USA*, *120*(6), Article e2218523120. https://doi.org/10.1073/pnas.2218523120

Brants, T., & Franz, A. (2006). *Web 1T 5-gram* (LDC2006T13; Version 1) [Data set]. Linguistic Data Consortium. https://doi.org/10.35111/cqpa-a498

Brown, K. S., Yee, E., Joergensen, G., Troyer, M., Saltzman, E., Rueckl, J., Magnuson, J. S., & McRae, K. (2023). Investigating the extent to which distributional semantic models capture a broad range of semantic relations. *Cognitive Science*, *47*(5), Article e13291. https://doi.org/10.1111/cogs.13291

Bruni, E., Tran, N.-K., & Baroni, M. (2014). Multimodal distributional semantics. *Journal of Artificial Intelligence Research*, *49*, 1–47. https://doi.org/10.1613/jair.4135

Brysbaert, M., Stevens, M., Mandera, P., & Keuleers, E. (2016). How many words do we know? Practical estimates of vocabulary size dependent on word definition, the degree of language input and the participant's age. *Frontiers in Psychology*, *7*, Article 1116. https://doi.org/10.3389/fpsyg.2016.01116

Bullinaria, J. A., & Levy, J. P. (2007). Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, *39*, 510–526. https://doi.org/10.3758/BF03193020

Campbell, E. E., & Bergelson, E. (2022). Making sense of sensory language: Acquisition of sensory knowledge by individuals with congenital sensory impairments. *Neuropsychologia*, *174*, Article 108320. https://doi.org/10.1016/j.neuropsychologia.2022.108320

Connell, L. (2019). What have labels ever done for us? The linguistic shortcut in conceptual processing. *Language, Cognition and Neuroscience*, *34*(10), 1308–1318. https://doi.org/10.1080/23273798.2018.1471512

Connell, L., & Lynott, D. (2014). Principles of representation: Why you can't represent the same concept twice. *Topics in Cognitive Science*, *6*(3), 390–406. https://doi.org/10.1111/tops.12097

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

Derby, S., Miller, P., Murphy, B., & Devereux, B. (2018). Using sparse semantic embeddings learned from multimodal text and image data to model human conceptual knowledge. In A. Korhonen & I. Titov (Eds.), *Proceedings of the 22nd Conference on Computational Natural Language*

Learning (pp. 260–270). Association for Computational Linguistics. https://doi.org/10.18653/v1/K18-1026

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, & T. Solorio (Eds.), *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Vol. 1, pp. 4171–4186). Association for Computational Linguistics. https://doi.org/10.18653/v1/N19-1423

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

Hills, T. T., Jones, M. N., & Todd, P. M. (2012). Optimal foraging in semantic memory. *Psychological Review*, *119*(2), 431–440. https://doi.org/10.1037/a0027373

Johns, B. T., Jamieson, R. K., & Jones, M. N. (2023). Scalable cognitive modelling: Putting Simon's (1969) ant back on the beach. *Canadian Journal of Experimental Psychology*, *77*(3), 185–201. https://doi.org/10.1037/cep0000306

Johns, B. T., & Jones, M. N. (2012). Perceptual inference through global lexical similarity. *Topics in Cognitive Science*, *4*(1), 103–120. https://doi.org/10.1111/j.1756-8765.2011.01176.x

Johns, B. T., Jones, M. N., & Mewhort, D. J. K. (2019). Using experiential optimization to build lexical representations. *Psychonomic Bulletin & Review*, *26*, 103–126. https://doi.org/10.3758/s13423-018-1501-2

Jones, M. N., & Mewhort, D. J. K. (2007). Representing word meaning and order information in a composite holographic lexicon. *Psychological Review*, *114*(1), 1–37. https://doi.org/10.1037/0033-295X.114.1.1

Kumar, A. A. (2021). Semantic memory: A review of methods, models, and current challenges. *Psychonomic Bulletin & Review*, *28*(1), 40–80. https://doi.org/10.3758/s13423-020-01792-x

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), 211–240. https://doi.org/10.1037/0033-295X.104.2.211

Lenci, A., Sahlgren, M., Jeuniaux, P., Cuba Gyllensten, A., & Miliani, M. (2022). A comparative evaluation and analysis of three generations of Distributional Semantic Models. *Language Resources and Evaluation*, *56*(4), 1269–1313. https://doi.org/10.1007/s10579-021-09575-z

Loconte, R., Orrù, G., Tribastone, M., Pietrini, P., & Sartori, G. (2023). *Challenging ChatGPT 'intelligence' with human tools: A neuropsychological investigation on prefrontal functioning of a large language model*. SSRN. https://doi.org/10.2139/ssrn.4377371

Louwerse, M. M. (2011). Symbol interdependency in symbolic and embodied cognition. *Topics in Cognitive Science*, *3*(2), 273–302. https://doi.org/10.1111/j.1756-8765.2010.01106.x

Louwerse, M. M., & Jeuniaux, P. (2010). The linguistic and embodied nature of conceptual processing. *Cognition*, *114*(1), 96–104. https://doi.org/10.1016/j.cognition.2009.09.002

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001

Mannering, W. M., & Jones, M. N. (2021). Catastrophic interference in predictive neural network models of distributional semantics. *Computational Brain & Behavior*, *4*, 18–33. https://doi.org/10.1007/s42113-020-00089-5

Marr, D. (1982). *Vision*. W. H. Freeman & Company.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). *Efficient estimation of word representations in vector space*. arXiv. https://doi.org/10.48550/arXiv.1301.3781

Pennington, J., Socher, R., & Manning, C. (2014). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Radford, A., Narasimhan, K., Salimans, T., & Sutskever, I. (2018). *Improving language understanding by generative pre-training*. OpenAI. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

Riordan, B., & Jones, M. N. (2011). Redundancy in perceptual and linguistic experience: Comparing feature-based and distributional models of semantic representation. *Topics in Cognitive Science*, *3*(2), 303–345. https://doi.org/10.1111/j.1756-8765.2010.01111.x

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, *1*(5), 206–215. https://doi.org/10.1038/s42256-019-0048-x

Savic, O., Unger, L., & Sloutsky, V. M. (2022). Exposure to co-occurrence regularities in language drives semantic integration of new words. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *48*(7), 1064–1081. https://doi.org/10.1037/xlm0001122

Shahmohammadi, H., Heitmeier, M., Shafaei-Bajestan, E., Lensch, H. P., & Baayen, R. H. (2023). Language with vision: A study on grounded word and sentence embeddings. *Behavior Research Methods*. Advance online publication. https://doi.org/10.3758/s13428-023-02294-z

Warstadt, A., Mueller, A., Choshen, L., Wilcox, E., Zhuang, C., Ciro, J., Mosquera, R., Paranjabe, B., Williams, A., Linzen, T., & Cotterell, R. (Eds.). (2023). Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In *Proceedings of the BabyLM challenge at the 27th Conference on Computational Natural Language Learning* (pp. 1–34). Association for Computational Linguistics. https://doi.org/10.18653/v1/2023.conll-babylm.1

Wingfield, C., & Connell, L. (2022). Understanding the role of linguistic distributional knowledge in cognition. *Language, Cognition and Neuroscience*, *37*(10), 1220–1270. https://doi.org/10.1080/23273798.2022.2069278