# Language Learning, Representation, and Processing in Humans and Machines: Introduction to the Special Issue

Marianna Apidianaki
University of Pennsylvania
Department of Computer and
Information Science
marapi@seas.upenn.edu

Abdellah Fourtassi
Aix Marseille University
CNRS, LIS
abdellah.fourtassi@gmail.com

Sebastian Padó
University of Stuttgart
Institute for Natural Language
Processing (IMS)
pado@ims.uni-stuttgart.de



*Large Language Models (LLMs) and humans acquire knowledge about language without direct supervision. LLMs do so by means of specific training objectives, while humans rely on sensory experience and social interaction. This parallelism has created a feeling in NLP and cognitive science that a systematic understanding of how LLMs acquire and use the encoded knowledge could provide useful insights for studying human cognition. Conversely, methods and findings from the field of cognitive science have occasionally inspired language model development. Yet, the differences in the way that language is processed by machines and humans—in terms of learning mechanisms, amounts of data used, grounding and access to different modalities—make a direct translation of insights challenging. The aim of this edited volume has been to create a forum of exchange and debate along this line of research, inviting contributions that further elucidate similarities and differences between humans and LLMs.*

## 1. Introduction

Large Language Models (LLMs) have come to dominate the field of computational linguistics. One reason is their ability to acquire rich information regarding linguistic structure and world knowledge (Tenney et al. 2019; Hewitt and Manning 2019; Petroni et al. 2019; Mahowald et al. 2024; Chang and Bergen 2024). A rather surprising aspect of

https://doi.org/10.1162/coli_e_00539

LLMs is that they demonstrate this ability despite using typically very simple training objectives, learning how to sensibly continue (or fill gaps in) text without the need for explicit supervision (Bengio, Ducharme, and Vincent 2000; Goldberg 2017; Devlin et al. 2019). In this broad sense, LLMs appear to work analogously to how humans develop most of their knowledge about language structure, meaning, and use—that is, spontaneously and without direct supervision.

Since the introduction of LLMs, there has been a feeling among some communities in both NLP and cognitive science that a systematic understanding of how these models work and how they use the knowledge they encode could help to shed light on the way humans acquire, represent, and process this same knowledge (Dupoux 2018; Cichy and Kaiser 2019; Caucheteux and King 2022; Goldstein et al. 2022). Conversely, findings from the areas of psycholinguistics and cognitive science have already inspired language model development, since these models are expected to exhibit human-like behavior in language use. For instance, datasets developed in—or inspired by—the field of cognitive science, including eye tracking and brain imaging datasets, have been used to evaluate models' behavior and to provide additional training signal (Ettinger 2020; Dunbar et al. 2017; Binz and Schulz 2023; Bingel, Barrett, and Søgaard 2016; Hollenstein et al. 2021).

Yet, there are also unmistakable differences between machines and humans, which put the direct translation of insights into question. Chief among them is the difference in learning mechanisms, as a consequence of which the size of data required to train LLMs far exceeds—by orders of magnitude—what humans need to acquire sophisticated conceptual structures and meanings (Frank 2023; Warstadt et al. 2023). Furthermore, human language is inherently grounded and multi-modal. In particular, children acquire world knowledge not only via exposure to language, but also via sensory experience and social interaction (Clark 2003; Tomasello 2009; Vigliocco, Perniss, and Vinson 2014). While some LLMs do learn from modalities other than text or speech, the degree to which this learning mirrors that of humans is far from obvious.

The aim of this special issue is to consolidate this exciting line of research, inviting contributions that further elucidate similarities and differences in the study of humans and LLMs, broadening the research scope to a range of linguistic levels and methodologies. The main questions we encouraged researchers to engage with are whether and how methods used in psycholinguistics (and cognitive science more generally) for studying the mechanisms of language processing and acquisition can be applied to the study of LLMs; and, conversely, whether the study of linguistic phenomena using LLMs, the investigation of the conceptual and world knowledge they encode, and the learning and processing principles they employ, can provide useful insights for studying human cognition.

We received a large number of interesting and engaging submissions, from which ten papers were accepted after two rounds of reviewing. In this preface, we provide an overview of these articles and discuss their contributions in terms of the most important lessons and challenges going forward.

## 2. Overview of the Articles in this Special Issue

The papers in this special issue can be grouped in terms of the topic addressed.

### 2.1 Meaning and Pragmatics

**Ohmer, Bruni, and Hupkes (2024)** explore LLMs' language understanding abilities. They prompt a model (GPT-3.5) with linguistic expressions that have the same underlying meaning and evaluate its consistency across different tasks. These might be expressions with the same reference in the real world (such as "morning star" and "evening star" for Venus), paraphrases or translations ("the sum of two and two", "two plus two", "zwei plus zwei"). The proposed experiments involve tasks of increasing complexity, from basic truth-conditional statements to more complex tasks, such as paraphrase detection and NLI. High consistency would suggest the LLM might be linking the expressions to their common underlying meaning. The results and follow-up analyses demonstrate that the model's meaning representations are strongly tied to form, and its understanding is still quite far from being consistent and human-like. The authors provide an interesting discussion on the consequences of these findings for the role of LLMs as explanatory models of semantic understanding in humans.

**Allaway et al. (2024)** investigate how LLMs reason about generalizations using generics, a particular type of statement that is fundamental to human reasoning but challenging to analyze semantically. Generics express generalizations (e.g., birds can fly) without explicit quantification; notably, they generalize over their instantiations (sparrows can fly) yet hold true even in the presence of exceptions (penguins do not). The authors use a framework grounded in pragmatics to automatically generate a large-scale dataset of generics, including both instantiations and exceptions. With this dataset, they probe whether LLMs exhibit similar behavior to humans in terms of quantification and property inheritance. LLMs show evidence of overgeneralization, similar to humans, but sometimes struggle to reason about exceptions. They are also found to exhibit similar non-logical behavior to humans when considering quantification and property inheritance.

**de Varda et al. (2024)** address the question of pseudoword meaning interpretation. Pseudowords are letter strings that are consistent with the orthotactical rules of a language, but do not appear in its lexicon and are traditionally considered to be meaningless (e.g., "knackets" or "spechy"). Previous studies that demonstrated humans' ability to make sense of pseudowords were limited by their focus on specific features (e.g., the emotional values of words). de Varda et al. instead analyze speakers' free definitions for pseudowords. They also show that LLMs compute embeddings for pseudowords which resemble the definitions given by study participants. This study confirms previous findings that pseudowords have semantic content. It shows a flexible form-to-meaning mapping that is useful to speakers when they encounter novel lexical entries.

### 2.2 Syntax and Grammar Induction

**Lampinen (2024)** highlights challenges in the comparison of LLMs and humans, taking the processing of recursively nested grammatical structures as a case study. While previous work found that language models cannot compete with humans, Lampinen argues that these studies disadvantaged language models by providing them with less task-related information than human participants. The study shows that simple prompting yields performance comparable—or even superior—to human results. The paper demonstrates the importance of methodological care and the difficulty in establishing comparability between humans and language models.

**Jon-And and Michaud (2024)** focus on the mechanisms of grammar induction and, more specifically, on simple cognitive principles that would support this

learning in humans, such as sequence memory. Their model uses Reinforcement Learning to identify sentences in a stream of words where cues to sentence borders (such as punctuation and capitalization) have been removed, and reuses these chunks in the learning process. They test the model on artificial languages—instantiating grammars at various complexities—showing that it succeeds in inducing parsimonious tree structures. Their study showcases how simple cognitive mechanisms like sequence memory and chunking can be effective in grammar induction.

### 2.3 Situational Grounding

**Jones, Bergen, and Trott (2024)** address the symbol grounding problem in language models and explore whether Multimodal LLMs (MLLMs) provide a plausible solution to this challenge. They investigate the degree to which MLLMs integrate modalities and if the way they do so mirrors the mechanisms believed to underpin grounding in humans, especially embodied simulation. Across a series of experiments, they ask whether MLLMs are, like humans, sensitive to sensorimotor features that are implied, but not explicit, in descriptions of an event. They find similarities and differences with human behavior, revealing strengths and limitations in the ability of current MLLMs to integrate language with other modalities.

**Beuls and Van Eecke (2024)** model the way language could emerge in a situated communicative context, similar to how humans acquire their native language. The authors conduct experiments where an artificial agent learns linguistic constructions, relying on situation-based intention reading and syntactico-semantic pattern identification. They argue that a situated and communicative learning context is essential to modeling human-like language acquisition. This goal contrasts with typical learning in LLMs where the input is predominantly text-based, and where the distribution of words serves as a basis for modeling meaning.

### 2.4 Phonology

**Georges et al. (2024)** investigate developing neural models for a challenging task in language acquisition, namely, learning how to form sounds with one's vocal tract (i.e., perform discrete articulatory gestures) from raw acoustic data, a very much underspecified problem. The article presents a series of modeling ideas, including the use of physiologically motivated inductive biases to regularize the learning problem. A series of careful evaluations demonstrates partial success—the model is able to learn interpretable gestures that lead to comprehensible speech—but still struggles with broader ecological validity (e.g., generalization between speakers). In this sense, the article presents a case study of the perspectives and pitfalls in modeling complex mechanisms of language acquisition.

**Pouw et al. (2024)** elucidate the extent to which neural models of Automatic Speech Recognition (ASR) detect phonological changes. More specifically, they consider the case of assimilation, where sounds change according to their context. Psycholinguistic studies have shown that human speakers can relate the phonology of a word and its phonetic realization, which raises the questions of where the representations learned by ASR models stand in this regard and what context cues they pay attention to. The authors carry out innovative intervention experiments on the Wav2Vec2 model, establishing strong evidence that the model focuses on local phonological context and that phonological "normalization" takes place in the final layers of the model. The resulting ASR model shows a substantial degree of match to human behavior, but is still limited in terms of time course and phenomena it can account for.

### 2.5 Imaging for Discourse Processing

**Ling, Murphy, and Fyshe (2024)** address the question of comparing language processing in LMs and the human brain. They conduct a decoding analysis, using Multi-timescale LSTM (or MT-LSTM), a model with temporally tuned parameters to induce sensitivity to different timescales of language processing (i.e., related to near/distant words). Such a model is particularly suited for high-temporal resolution brain data like electroencephalography (EEG). They study the extent to which EEG signals predict MT-LSTM embeddings on various timescales. This innovative study, combining MT-LSTM with EEG data, complements previous research that has primarily focused on fMRI to study the representation of linguistic timescales in the brain.

Together, the papers cover various levels of linguistic knowledge, including phonology, syntax, semantics, pragmatics, and discourse.

A few papers are concerned with the mechanisms of language learning, particularly in syntax (Jon-And and Michaud 2024; Beuls and Van Eecke 2024) and phonology (Georges et al. 2024). These papers generally highlight mechanisms that are not typically part of the LLM's training pipeline but crucial to children's linguistic development, such as learning from a situated, communicative context (Beuls and Van Eecke 2024), using simple cognitive principles like chunking and sequence memory (Jon-And and Michaud 2024), and leveraging the motor system (Georges et al. 2024). This highlights once more the role of extratextual signals in human language learning, that many current LLMs do not have access to.

Most papers, however, focus on mechanisms of processing (Pouw et al. 2024; Lampinen 2024; de Varda et al. 2024; Ohmer, Bruni, and Hupkes 2024; Jones, Bergen, and Trott 2024; Ling, Murphy, and Fyshe 2024; Allaway et al. 2024). These papers generally compare LLMs to human performance on the same or similar tasks and obtain mixed results. Some papers find that LLMs perform similarly to humans. In particular, Chinchilla is able to process recursively nested grammatical structures (Lampinen 2024). Wav2Vec2 can recognize the underlying phonological form in challenging contexts (e.g., place assimilation [Pouw et al. 2024]). Finally, LLMs like GPT-2 and RoBERTa can generate human-like meaning definitions for novel words (de Varda et al. 2024). However, other papers find LLMs lacking when compared to human behavior or brain data. For instance, multimodal LLMs (like CLIP) are not adequately sensitive to sensorimotor features implicit in language (Jones, Bergen, and Trott 2024). GPT-3.5 struggles to represent meaning consistently across different linguistic forms (Ohmer, Bruni, and Hupkes 2024). Some LLMs (like GPT-4 and LLAMA-2) conflate universally quantified statements with generics, though humans also exhibit similar non-logical behavior (Allaway et al. 2024). Finally, the embeddings from Multi-timescale LSTMs can be decoded from EEG data for short, but not as well for medium or long timescales (Ling, Murphy, and Fyshe 2024).

## 3. Challenges and Directions for Future Research

While the studies reviewed above significantly advance our understanding of the research questions posed in this special issue, they represent a small sample of the studies needed to address the complex relationship between language learning and processing in humans and LLMs. The different fields involved in this research (in our case, cognitive science and computational linguistics) often operate under different principled and pragmatic assumptions about language, allowing them to focus on what they deem most relevant. However, this diversity also creates difficulties in transferring insights

across disciplines and in building a cumulative body of research. In the following, we highlight some recurring challenges and perspectives that can inform future work in this direction.

One important challenge arises from the fact that *LLMs are computational artifacts*. Psycholinguistics fundamentally assumes that its participants can be drawn randomly from a population of speakers that is relatively stable (at least over short time spans). In contrast, for LLMs, new models are being proposed almost every few months, and there is no guarantee that analyses of LLMs' cognitive capabilities carry over between model families. Furthermore, the choice of models is not random: A researcher has to grapple with the dilemma of testing open-source models, which provide a more reliable environment for reproducible research, and commercial models, which typically provide higher performance at the expense of transparency. The studies in this special issue use a diversity of models and evaluation methods, with a pattern emerging that researchers report results with smaller open-source models if these are sufficient to exhibit the target knowledge or mechanisms, and larger commercial models when the target knowledge is more challenging. While this diversity is vital in an early, exploratory phase of this research program, a more systematic approach appears essential for a more cumulative phase.

An important distinction made in psycholinguistics is between the *intrinsic difficulty of a task* and *auxiliary task demands*, for example, memory consumption or complex instructions. Participants can show bad performance at a task if they do not meet the task demands, irrespective of their ability to carry out the core task. The same is true for LLMs: If they fare poorly in a language processing task, it is not always easy to differentiate when poor performance is due to inadequate linguistic knowledge and when it is due to other types of demands made by the task—for example, in terms of working memory (Hu and Frank 2024). Thus, it is crucial to examine the model's understanding of a task and its ability to perform it with available resources before one can draw conclusions on the model's knowledge. As a case in point, Lampinen (2024) observes that LLMs can be disadvantaged when given less task-related information than human participants, even when they have access to more (training) data.

Another important parameter that needs to be accounted for in studies on language learning and processing in humans and LLMs is the *ecological validity* of the considered cognitive mechanisms. Many studies on language development focus on cognitively plausible mechanisms, but their implementation typically does not scale to natural language or relies on drastic simplifications to do so. This makes it hard to evaluate the ecological validity of these mechanisms and their practical advantages, especially compared to LLMs. In the case of some cognitively inspired mechanisms, a promising strategy is to integrate them as inductive biases within LLMs, thereby merging the insights of cognitive science with the scalability of LLMs. This is still a largely unexplored area of research, although promising results were reported using transfer learning and meta-learning techniques (Papadimitriou and Jurafsky 2023; McCoy and Griffiths 2023; Lake and Baroni 2023).

Notably, the scientific approach adopted in cognitive science and LLM research has, until now, been different. In psycholinguistics, as in the cognitive sciences more generally, scientific progress involves the *interplay of theory building and empirical investigation* (Haig 2014), as reflected in the two main types of articles published in this field. In contrast, the study of LLMs has (at least currently) inherited the focus on empirical work from NLP, with a corresponding lack of theory building. This is not surprising, as a "cognitive theory of LLMs" has yet to be clearly articulated.

In particular, not all research on LLMs that uses cognitive measures or data is necessarily interested in developing models that are cognitively plausible. Many researchers simply believe in the usefulness of cognitive assessments as parts of a general-purpose evaluation benchmark. In addition, it is unclear what cognitive status one should assign to the growing body of work on prompt optimization, if any. Fundamentally, there is a tension between building models that are *as human-like as possible*, including by incorporating human constraints (e.g., cognitive architectures [Newell 1990]); and models that are optimized to perform *as well as possible in practical tasks*, which often requires removing human constraints such as limitations on working memory. We believe that studies in the LLM area should position themselves with respect to this distinction, which is currently often not the case.

A crucial challenge is the need for *experiments that manage to compare human and machine language learning under more similar conditions*, for example, in terms of data size and access to different modalities. Prompts must also be carefully designed in order to not disadvantage one side of the comparison over the other. In terms of *evaluation*, there is a need for benchmarks that address various aspects of language learning. Existing resources used for evaluation (e.g., McRae et al. 2005; Devereux et al. 2014) contain features of much different nature than the implicit perceptual features they are supposed to stand for (Bruni, Tran, and Baroni 2014). There is also a need for benchmarks addressing the models' grounding capabilities, and how well they capture empirical phenomena associated with situational word learning (Ebert and Pavlick 2020; Vong and Lake 2022; Jiang et al. 2023).

Regarding the communication between the areas of cognitive science and LLM research, findings and insights from one are already being exploited in the other. For example, standard tasks in psycholinguistics allow us to examine how LLMs fare in comparison to humans. However, if the model's behavior shows a discrepancy with humans, we often lack *insights on how the models can be improved* or better aligned with human knowledge. Commonly used probing techniques, which typically point to correlations between model representations and linguistic properties, do not always help in this respect (Feder et al. 2021; Lyu, Apidianaki, and Callison-Burch 2024). The challenge can potentially be mitigated using methods that can identify causal relationships between model structures and outputs, such as ablation studies and counterfactual methods.

Conversely, causal intervention in LLMs could provide insights into human language learning. High-performing LLMs are typically hard to (re-)train with academic computing resources. This limits the type of investigation that can be pursued. For example, the papers in this special issue typically test knowledge in pre-trained models, the same way we test knowledge in humans. While this approach can provide a wealth of insights, as it has done with humans in the field of experimental psychology, the ability to re-train LLMs in different diagnostic conditions (at the architecture or input level) would be a game-changer by providing the opportunity for *causal interventions*. Such interventions could provide insights into the mechanisms of knowledge emergence in LLMs in a way that could not be obtained in humans. In fact, this approach—if made logistically feasible (see, for example, the effort in Warstadt et al. 2023)—can also provide insights into the emergence of knowledge in humans.

## References


Allaway, Emily, Chandra Bhagavatula, Jena D. Hwang, Kathleen McKeown, and Sarah-Jane Leslie. 2024. Exceptions, instantiations, and overgeneralization: Insights into how language models process generics. *Computational Linguistics*, 50(4):1211–1275. https://doi.org/10.1162/coli_a_00530

Bengio, Yoshua, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*, volume 13.

Beuls, Katrien and Paul Van Eecke. 2024. Humans learn language from situated communicative interactions. What about machines? *Computational Linguistics*, 50(4):1277–1311. https://doi.org/10.1162/coli_a_00534

Bingel, Joachim, Maria Barrett, and Anders Søgaard. 2016. Extracting token-level signals of syntactic processing from fMRI—with an application to PoS induction, Erk, Katrin and Noah A. Smith, editors. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–755. https://doi.org/10.18653/v1/P16-1071

Binz, Marcel and Eric Schulz. 2023. Using cognitive psychology to understand GPT-3. *Proceedings of the National Academy of Sciences*, 120(6):e2218523120. https://doi.org/10.1073/pnas.2218523120, PubMed: 36730192

Bruni, Elia, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *Journal of Artificial Intelligence Research (JAIR)*, 49(1):1–47. https://doi.org/10.1613/jair.4135

Caucheteux, Charlotte and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5(1):Article 134, 10 pages. https://doi.org/10.1038/s42003-022-03036-1

Chang, Tyler A. and Benjamin K. Bergen. 2024. Language model behavior: A comprehensive survey. *Computational Linguistics*, 50(1):293–350. https://doi.org/10.1162/coli_a_00492

Cichy, Radoslaw M. and Daniel Kaiser. 2019. Deep neural networks as scientific models. *Trends in Cognitive Sciences*, 23(4):305–317. https://doi.org/10.1016/j.tics.2019.01.009, PubMed: 30795896

Clark, Eve V. 2003. *First Language Acquisition*. Cambridge University Press, Cambridge, U.K.

Devereux, Barry J., Lorraine K. Tyler, Jeroen Geertzen, and Billi Randall. 2014. The Centre for Speech, Language and the Brain (CSLB) concept property norms. *Behavior Research Methods*, 46(4):1119–1127. https://doi.org/10.3758/s13428-013-0420-4, PubMed: 24356992

Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 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)*, pages 4171–4186. https://doi.org/10.18653/v1/N19-1423

Dunbar, Ewan, Xuan Nga Cao, Juan Benjumea, Julien Karadayi, Mathieu Bernard, Laurent Besacier, Xavier Anguera, and Emmanuel Dupoux. 2017. The zero resource speech challenge 2017. In *Proceedings of the IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, pages 323–330. https://doi.org/10.1109/ASRU.2017.8268953

Dupoux, Emmanuel. 2018. Cognitive science in the era of artificial intelligence: A roadmap for reverse-engineering the infant language-learner. *Cognition*, 173:43–59. https://doi.org/10.1016/j.cognition.2017.11.008, PubMed: 29324240

Ebert, Dylan and Ellie Pavlick. 2020. A visuospatial dataset for naturalistic verb learning. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 143–153.

Ettinger, Allyson. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48. https://doi.org/10.1162/tacl_a_00298

Feder, Amir, Nadav Oved, Uri Shalit, and Roi Reichart. 2021. CausaLM: Causal model explanation through counterfactual language models. *Computational Linguistics*, 47(2):333–386. https://doi.org/10.1162/coli_a_00404

Frank, Michael C. 2023. Bridging the data gap between children and large language models. *Trends in Cognitive Sciences*, 27(11):990–992. https://doi.org/10.1016/j.tics.2023.08.007, PubMed: 37659919

Georges, Marc-Antoine, Marvin Lavechin, Jean-Luc Schwartz, and Thomas Hueber. 2024. Decode, move and speak! Self-supervised learning of speech units, gestures, and sound relationships using vocal imitation. *Computational Linguistics*, 50(4):1345–1373. https://doi.org/10.1162/coli_a_00532

Goldberg, Yoav. 2017. *Neural Network Methods for Natural Language Processing*. Morgan & Claypool Publishers, San Rafael, CA. https://doi.org/10.1007/978-3-031-02165-7

Goldstein, Ariel, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emmanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2022. Shared computational principles for language processing in humans and deep language models. *Nature Neuroscience*, 25(3):369–380. https://doi.org/10.1038/s41593-022-01026-4, PubMed: 35260860

Haig, Brian D. 2014. *Investigating the Psychological World: Scientific Method in the Behavioral Sciences*. MIT Press, Cambridge, MA. https://doi.org/10.7551/mitpress/9780262027366.001.0001

Hewitt, John and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. 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)*, pages 4129–4138. https://doi.org/10.18653/v1/N19-1419

Hollenstein, Nora, Emmanuele Chersoni, Cassandra L. Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus. 2021. CMCL 2021 shared task on eye-tracking prediction. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 72–78. https://doi.org/10.18653/v1/2021.cmcl-1.7

Hu, Jennifer and Michael C. Frank. 2024. Auxiliary task demands mask the capabilities of smaller language models. In *Proceedings of the First Conference on Language Models*, pages 1–17.

Jiang, Guangyuan, Manjie Xu, Shiji Xin, Wei Liang, Yujia Peng, Chi Zhang, and Yixin Zhu. 2023. MEWL: Few-shot multimodal word learning with referential uncertainty. In *Proceedings of the 40th International Conference on Machine Learning (ICML'23)*, pages 15144–15169.

Jon-And, Anna and Jérôme Michaud. 2024. Usage-based grammar induction from minimal cognitive principles. *Computational Linguistics*, 50(4):1375–1414. https://doi.org/10.1162/coli_a_00528

Jones, Cameron R., Benjamin Bergen, and Sean Trott. 2024. Do multimodal large language models and humans ground language Similarly? *Computational Linguistics*, 50(4):1415–1440. https://doi.org/10.1162/coli_a_00531

Lake, Brenden M. and Marco Baroni. 2023. Human-like systematic generalization through a meta-learning neural network. *Nature*, 623:115–121. https://doi.org/10.1038/s41586-023-06668-3, PubMed: 37880371

Lampinen, Andrew. 2024. Can language models handle recursively nested grammatical structures? A case study on comparing models and humans. *Computational Linguistics*, 50(4):1441–1476. https://doi.org/10.1162/coli_a_00525

Ling, Sijie, Alex Murphy, and Alona Fyshe. 2024. Exploring temporal sensitivity in the brain using multi-timescale language models: An EEG decoding study. *Computational Linguistics*, 50(4):1477–1506. https://doi.org/10.1162/coli_a_00533

Lyu, Qing, Marianna Apidianaki, and Chris Callison-Burch. 2024. Towards faithful model explanation in NLP: A survey. *Computational Linguistics*, 50(2):657–723. https://doi.org/10.1162/coli_a_00511

Mahowald, Kyle, Anna A. Ivanova, Idan A. Blank, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2024. Dissociating language and thought in large language models. *Trends in Cognitive Sciences*, 28(6):517–540. https://doi.org/10.1016/j.tics.2024.01.011, PubMed: 38508911

McCoy, R. Thomas and Thomas L. Griffiths. 2023. Modeling rapid language learning by distilling Bayesian priors into artificial neural networks. *arXiv preprint:2305.14701*.

McRae, Ken, George Cree, Mark Seidenberg, and Chris Mcnorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behavior Research Methods*, 37(4):547–559.

https://doi.org/10.3758/bf03192726, PubMed: 16629288
Newell, Allen. 1990. *Unified Theories of Cognition*. Harvard University Press, Cambridge, MA.
Ohmer, Xenia, Elia Bruni, and Dieuwke Hupkes. 2024. From form(s) to meaning: Probing the semantic depths of language models using multisense consistency. *Computational Linguistics*, 50(4):1507–1556. https://doi.org/10.1162/coli_a_00529
Papadimitriou, Isabel and Dan Jurafsky. 2023. Injecting structural hints: Using language models to study inductive biases in language learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8402–8413. https://doi.org/10.18653/v1/2023.findings-emnlp.563
Petroni, Fabio, Tim Rocktäschel, Sebastian Riedel, Patrick Lewis, Anton Bakhtin, Yuxiang Wu, and Alexander Miller. 2019. Language models as knowledge bases? In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2463–2473. https://doi.org/10.18653/v1/D19-1250
Pouw, Charlotte, Marianne de Heer Kloots, Afra Alishahi, and Willem Zuidema. 2024. Perception of phonological assimilation by neural speech recognition models. *Computational Linguistics*, 50(4):1557–1585. https://doi.org/10.1162/coli_a_00526
Tenney, Ian, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R. Thomas McCoy, Najoung Kim, Benjamin Van Durme, Sam Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? Probing for sentence structure in contextualized word representations. In *Proceedings of International Conference on Learning Representations (ICLR 2019)*, 17 pages.
Tomasello, Michael. 2009. *Constructing a Language*. Harvard University Press.
de Varda, Andrea Gregor, Daniele Gatti, Marco Marelli, and Fritz Günther. 2024. Meaning beyond lexicality: Capturing pseudoword definitions with language models. *Computational Linguistics*, 50(4):1313–1343. https://doi.org/10.1162/coli_a_00527
Vigliocco, Gabriella, Pamela Perniss, and David Vinson. 2014. Language as a multimodal phenomenon: Implications for language learning, processing and evolution. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 369(1651):20130292. https://doi.org/10.1098/rstb.2013.0292, PubMed: 25092660
Vong, Wai Keen and Brenden M. Lake. 2022. Cross-situational word learning with multimodal neural networks. *Cognitive Science*, 46(4):e13122. https://doi.org/10.1111/cogs.13122, PubMed: 35377475
Warstadt, Alex, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. 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*, pages 1–34. https://doi.org/10.18653/v1/2023.conll-babylm.1