Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.601

Squibs and Discussions

On the referential capacity of language models: An internalist rejoinder to Mandelkern & Linzen

Giosuè Baggio[1], Elliot Murphy[2]

[1] Language Acquisition and Language Processing Lab, Department of Language and Literature, Norwegian University of Science and Technology
giosue.baggio@ntnu.no

[2] Vivian L. Smith Department of Neurosurgery, University of Texas Health Science Center at Houston, TX, USA
Elliot.Murphy@uth.tmc.edu

Mandelkern and Linzen (2024) argue that words generated by language models (LMs) are linked to causal histories of use within human linguistic communities, and ultimately to their referents. Therefore, LMs' words refer. We qualify Mandelkern and Linzen's claim as applicable to a narrow class of expressions. Thus qualified, the claim is valid and motivated by the need to evaluate LMs' outputs for relevance and truth. Next, we discuss the actual scope of their claim, and we conclude that the bounds of sense and reference in LMs are more restricted than in humans. We close with some considerations on the status of LMs as members of human linguistic communities.

Mandelkern and Linzen (2024) (M&L) develop an externalist defense of the capacity of language models' (LMs) words to refer, i.e., to "achieve 'word-to-world' connections". In externalist theories, causally uninterrupted chains of usage, tracing every occurrence of a name back to its bearer, guarantee that, for example, 'Peano' refers to the individual Peano (Kripke 1980). This account is 'externalist', both because words pick out referents 'out there' in the world, and because reference is determined by actions in a community of speakers, not by individual mental states. For M&L, the "central question" is whether LMs, too, belong to human linguistic communities, such that words by LMs would also trace back to their referents. Their response is a cautious 'yes': occurrences of 'Peano' in LMs' outputs are as causally connected to Peano as any use of 'Peano' in human speech or text; therefore, occurrences of 'Peano' in LMs' outputs refer to Peano.

Below, we present a critique of M&L's externalist analysis and defend an internalist alternative. Our argument has four parts. In Section 1, we state why the question of LMs' reference matters. In Section 2, we argue that the causal-historical apparatus, developed for proper names and natural kind terms, encounters two main obstacles in applications to machines: first, the distinction between words and tokens, and the potential of tokens

Action editor: Michael White. Submission received: 9 April 2025; revised version received: 6 September 2025; accepted for publication: 10 October 2025.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.601

to disrupt chains of word usage; second, the fact that successful reference for indexicals and other expressions requires embodied and situated agents. In Section 3, we introduce a number of conditions on reference-capable agents from an internalist perspective, and we conclude that LMs lack the cognitive capacities necessary for reference. In Section 4, we consider the idea that LMs implement 'simulacra' that function as atypical members of linguistic communities, acknowledging the significant 'cognitive work' human users must perform to make this fiction viable.

1. Why Reference Matters: Machines Under Public Scrutiny

The question of whether LMs' words refer matters, not least because we ought to be able to judge whether LMs' outputs are relevant to a given topic, plausible, true, etc., and we should extend similar judgments to entailed, presupposed, and implicated meanings. If a chatbot composed a biography of Peano, containing a mix of true and false statements, we would not be able to say that what is true is true and what is false is false, if we took 'Peano' in the chatbot's outputs *not* to refer to anyone. Denying that machine speech and text may be taken to refer, or be about individuals or states of affairs, risks undermining the public enterprise of evaluating LMs' outputs for pertinence, truth, and other norms. Arguments that LMs are unable to recover meaning from input data and infuse meaning into output strings (Bender and Koller 2020) risk shifting the 'burden of truth' largely or entirely onto human interpreters. If 'Peano taught in Bologna' does not mean or refer to anything 'for' the LM that generates it, then it would be false (which it is) only relative to the interpretations it receives from users. Machine language would only be evaluable in an arena of competing interpretations, not also against the constraints of language. One may reply that we should interpret LMs as we would any human member of a linguistic community (DeVault, Oved, and Stone 2006; Cappelen and Dever 2021; Lederman and Mahowald 2024): 'Peano taught in Bologna' is false when produced by human speakers, and so it must be false also when generated by LMs. This position is attractive, but needs to be qualified: we will do that in Section 4. Let us first examine M&L's argument.

2. From Words to Reference: Stumbling Blocks for Machines

Pursuing an externalist account of reference for LMs requires that the relevant concepts from the causal-historical framework (i) are equally applicable to machines and (ii) have sufficient scope to warrant the desired generalizations, including M&L's conclusion that LMs' words refer. The concepts of 'word' and 'causal history' are especially problematic in the context of machine language (Section 2.1), and even the assumption, independent of externalism, that *all* words, as generated by LMs, refer is hard to uphold (Section 2.2).

2.1 Causal Histories of What? Words *vs* Tokens

M&L observe that LMs' inputs include "strings of symbols with certain natural histories that connect them to their referents". The concept of 'natural histories' is not suitable for machine language: there is nothing natural nor historical about strings of symbols. LMs' tokens might not correspond to the (parts of) expressions that have causal histories, such as characters, morphemes, or words (Haspelmath 2023; Murphy 2024): e.g., the GPT-3.5, -4, and -4o tokenizers analyze 'incomprehensibility' as *in, com, preh, ens, ibility*. What are the causal histories of these tokens? Within current subword approaches to tokenization (Apidianaki 2023; Mielke et al. 2024), misalignments between tokens and linguistic units are expected (Riedl and Biemann 2018). Indeed, whereas morphological decomposition

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.601

preserves structure-meaning correspondences (e.g., the prefix 'in-' negates the meaning of the stem it applies to), subword tokenization creates arbitrary subword strings driven by statistical frequency, rather than lexical structure and meaning. For instance, Haslett (2025) shows that GPT-4, GPT-4o, and Llama 3 draw incorrect inferences about semantic similarity from token similarity. Tokenization disrupts chains of usage, both at the point of interaction with users, as Haslett implies ("subword tokens corrupt representations of word meanings [...] and systematically corrupted representations will impede people's ability to instruct and interpret LLMs"), and during pretraining. This raises the question of how exactly LMs' words may be connected to causal-historical chains of word usage. Our concern is with the adequacy of externalist explanations. If causal-historical chains are doing the explanatory work, then disruptions in those chains would undermine the explanation. An account in which learners must deploy internal resources to reconstruct meaningful linguistic units from arbitrary parts, and are only able to reconnect to usage chains in virtue of such internal reconstruction, is no longer an externalist story.

2.2 Do All Words Refer? The Bounds of Machine Sense and Reference

M&L's focus is on the externalist's favorite cases: proper names and natural kind terms. There is no mention of other classes of expressions in their article. This not an oversight: a causal-historical account of reference, particularly for LMs, cannot be easily extended to other expressions. This creates both a scope problem and a series of deeper theoretical problems (Section 3) for externalist theories of machine reference.

The scope problem is as follows. There are histories of usage for all words, but there is a variety of ultimate anchors for those histories: not all anchors are of the ontological types that externalism favors. What would 'love', 'democracy', 'English', etc. trace back to, if anything? Not to entities or events in the world, but rather to *concepts* that organize cognitive and bodily information, knowledge, etc. The word-to-word relationships that LMs come to internalize at best *approximate* human conceptual networks. This can make it easier to see how LMs' outputs could have conceptual or inferential semantics, if not quite reference (Piantadosi and Hill 2022; Pavlick 2023), but it does not help to formulate viable externalist explanations of the semantic capabilities of LMs (Søgaard 2025).

Consider expressions that require a situated speaker, such as indexicals, pronouns, and demonstratives. What would 'now' and 'here' refer to in LMs' outputs? Where is an LM, such that 'here' may be interpreted as denoting its current location? Though human speakers may occasionally fail to establish reference through indexicals, such failures in LMs are architectural and systematic: LMs cannot occupy deictic spaces or establish the spatial, temporal, and social coordinates necessary for indexical reference.

Languages also include non-deictic expressions whose meaning requires a situated speaker with specific bodily and mental characteristics: psychological and physiological verbs (e.g., 'remember', 'wonder', 'believe', 'wake up', 'fall asleep', 'be sick'), movement and posture verbs (e.g., 'walk', 'sit'), appearance verbs ('seem'), perspective-dependent adjectives ('heavy', 'distant'), etc. The list is long. First-person uses of these expressions would be meaningless when generated by an LM, which raises the question whether the same expressions in different persons than the first *could* then have sense and reference. In general, *the bounds of sense and reference are not the same for humans and for (different kinds of) machines*. Endowing an LM with a robotic body could extend its 'semantic reach' for personal and spatial deixis, movement verbs, etc., but some of its outputs would remain meaningless, even in cases where externalist theories fare best, say, for a combination of a possessive determiner and a natural kind term: e.g., 'My stomach'. When prompted 'Do you have a stomach ache?', an LM cannot meaningfully respond either affirmatively

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.601

or negatively, as 'my stomach' lacks referential grounding in embodied experience. The fact that most LM-based chatbots would now reply 'I don't have a stomach or a physical body, so I can't experience stomach aches or any physical sensations. I'm an AI assistant without physical form' (Claude Opus 4.1), supports our point: currently, LMs are largely aligned with what we intuit to be the bounds of sense for disembodied machines. Still, the use of 'I' remains problematic: who does the pronoun refer to in this case? In general, how could we explain that, for LMs, *reference would fail* for a range of expressions (those listed above), *in spite of the fact* that, for all such expressions, usage histories in human communities of speakers exist and are reflected in the models' training data?

Our own view is not just that causal histories "cannot ground the reference relation" (Ostertag 2025) and cannot explain the referential limitations of LMs. Those limitations *may only be explained internalistically*, by appealing to constraints on the architectures and information-processing characteristics of LMs. For example, in most cases, pronouns are only interpretable relative to hypotheses about the speaker's referential intentions: who are 'I', 'us', or 'them' in machine outputs, when LMs cannot form referential intentions? We return to this important issue in Section 3.3.

One could grant LMs the capacity to produce 'simulacra' that would appear to have communicative intentions and other mental states, to the extent that LMs "convincingly play the role of a character that does" (Shanahan, McDonell, and Reynolds 2023). A 'role play' perspective cannot endow machine outputs with referential properties, as neither the LM nor the simulacra it supports occupy the sorts of deictic spaces that would make uses of 'us', 'here', etc. pick out contextually appropriate referents. The simulacra inherit restrictions on the bounds of sense and reference from the LMs that support them. They also inherit the LMs' information-processing constraints, including the inability to form communicative intentions. More and different data, including multimodal inputs, could enrich the LMs' internal representations of usage patterns and conceptual roles for some words, but cannot address problems that are *architectural* at heart. Reference minimally requires a grammar interfaced with perceptual, conceptual, and pragmatic systems, and speakers capable of forming communicative intentions, situated in social and physical space and time (Scott-Phillips 2014; Hinzen 2016; Murphy 2016; Murphy, Holmes, and Friston 2024). Only machines with bodies and brains with specific capacities will be able to break the current semantic barriers of LMs: LM words will then refer also in virtue of those capacities, and not only by participating in causal histories of word usage.

3. Enter Internalism: Conditions on Reference-Capable Agents

We have argued that attributing reference to LMs' words is required to assess the claims they make for relevance, truth, etc. (Section 1). However, not all expressions by LMs *can* refer (Section 2). For the remaining cases, we will need to clarify who does the 'cognitive work' behind reference (Section 4). In this Section, we show how internalism provides a framework for explaining why LMs largely lack the capacities necessary for reference.

3.1 No Speaker Communities Without Cognitive Capacities

Let us suppose that tokenization does disrupt *causal* chains of word usage, but that LMs achieve the relevant outcomes *statistically*: the sequences of tokens we call 'words', and that emerge from pretraining, are used in similar ways by humans and LMs. According to this position, *correlations* between human and machine language usage would suffice for LMs to count as members of linguistic communities, as per M&L's thesis.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.601

However, agents are not endowed with the capacity to use words according to their 'true' sense and reference only or principally by virtue of the fact that they participate in human linguistic communities, whether their participation is anchored causally or only correlationally to usage patterns by others within the same communities, and ultimately to the words' referents. What delimitations can be placed on the notion of 'community', that would help with explaining the semantic properties of a language, including what words refer to? Coherent, coordinated linguistic behaviors by members of a community are constrained by how their cognitive system 'carves up' and conceptualizes the world. Some words might acquire meaning through linguistic conventions, without necessarily being grounded in concrete objects and properties. Conventions may get usage histories off the ground and stabilize form-meaning pairings. However, the meanings themselves and their articulations as modality-specific signs are internal constructions: conventions correlate perceptual and cognitive representations of form and meaning (Baggio 2018). Even if we accept first-order externalism (reference is determined by factors external to individuals' mental states), we still have to consider cognitive architecture to account for how reference is possible, how words refer, and what makes theories of meaning true or false: this is what Cohnitz and Haukioja (2013, 2025) call *semantic meta-internalism*. They argue that its negation is "an implausible view about semantics" (Cohnitz and Haukioja 2013), largely because explanations of human communication, of referential success and failure, rest on cognitive factors, such as what speakers know about lexical meaning and reference (or what semantic theories they have internalized), and how they intend to use expressions in specific circumstances. Facts about the mind-brain have to be invoked to explain *how* words refer: not what they refer to, but what determines what they refer to (Block 1987; Chomsky 2000; Jackendoff 2002; Carey 2009; Pietroski 2017; Baggio 2018; Murphy 2023; Cohnitz and Haukioja 2025).

Reference and referring require at least: (a) the capacity to parse linguistic structure and access lexical information; (b) compositional mechanisms for building structure and meaning from constituent parts; (c) bodily and mental characteristics enabling situated, perspective-specific speech; (d) pragmatic capabilities, including forming and inferring referential intentions; (e) conceptual systems interfaced with systems for perception and action. In what follows, we elaborate on these points in more detail.

3.2 No Reference Without Cognition, 1: Internal Models of Language and World

Let us return to what we agree on with M&L. They write that if a friend, Luke, sends us the message 'Peano proved that arithmetic is incomplete', then "his text says something false, namely something about Peano: that he proved incompleteness. So Luke's use of the word 'Peano' refers to Peano." This is correct, but we want to frame the observation differently. Under the assumption that 'Peano' refers to Peano, in a language or idiolect where that is the case, that sentence is false. Luke's text says something false according to our own internal models of the world and how our language capacity interfaces with them. We get the feeling that Luke is — whether he intends to or not, and whether what he writes is true or false — referring to Peano: our own generative inferences are about Peano. This is neither solipsism nor coherentism: the external world constrains the truth of our claims. Our point is that determining what Luke's sentence refers to, and whether it is true, requires accessing internal representations of language and the world.

We have seen that LMs' words cannot refer by virtue of LMs' participation in causal chains of language usage: even if those chains were not disrupted by tokenization, they would not suffice to explain how all words refer. Cognitive structure, on the part of both producers and receivers of linguistic signals, is necessary to account for *how* words refer

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.601

(semantic meta-internalism) and even for the fact *that* words can refer and sentences can be true. By and large, LMs lack the mental structures that, in humans, explain referential capacities, but human cognition suffices to explain how LMs' words can be (and, in fact, are routinely) interpreted as having meaning and reference: this is why 'Peano' in LMs' outputs will refer to Peano, according to *our own* models of language and the world.

When we utter or interpret (1), we are not necessarily presupposing any connections to states of the external world, but to possible states within our models of reality:

(1) The previous emperor of Kansas is about to announce a new mixtape.

Nothing in the usage histories of the expressions in (1) explains its referential properties. The reference of 'The previous emperor of Kansas' results from composing the meanings of 'emperor', 'Kansas', 'previous', and the definite determiner, according to the phrase's syntactic structure: these operations depend more on internal cognitive algorithms than on causal histories of use (van Lambalgen and Hamm 2004; Pietroski 2018; Pustejovsky and Batiukova 2019). Reference is established relative to a *model* (in some fictional world, a non-actual possible world, etc.) in the course of this computation (van Lambalgen and Hamm 2005; Baggio 2018). Truth requires that the computation is extended to linguistic objects of a certain syntactic complexity, e.g., a sentence. Future machines might manage this referential calculation owing to their architecture and computational capacities (van Deemter 2016), not primarily or exclusively to their anchoring in language use.

To appreciate this point from a different angle, consider an event in which John and James are hunting each other, and suppose that these descriptions are both true:

(2) John chased James athletically but not skillfully.

(3) James chased John unathletically but skillfully.

What may be the real-world anchors, the endpoints of causal histories, of the predicates 'athletically' and 'skillfully'? An action cannot be skillful and unskillful, a person cannot be athletic and unathletic at the same time — this has less to do with the 'external world' than with peculiarities of human cognition, e.g., how aspects of the world are attended to, selected, categorized, conceptualized, or lexicalized (Jackendoff 2002; Pietroski 2005; Carey 2009; Murphy 2023). These and other expressions, along with the examples from Section 2 and others, resist an externalist treatment. Internalism offers better prospects for explaining both why (words as used by) humans *can* refer and why machines *cannot* refer: LMs lack some or all of the capacities and properties listed in (a)-(e), Section 3.1.

3.3 No Reference Without Cognition, 2: Communicative Intentions

M&L discuss cases where machines are exclusively trained on the kind of data that Luke may have received about Peano and other historical figures. They write: "The inputs to LMs are not just forms, but forms with particular histories of referential use. And those histories ground the referents of those forms, whether or not those histories are known or accessible." They ask what the philosophical consequences of this are, but they do not raise a more pertinent question: would an LM trained only on photographs and 'tokens' of ant trails accidentally spelling out English sentences yield any philosophical import, even if it achieved comparable accuracy and fluency to a model trained on Wikipedia or the entire internet, and regardless of whether it generated true or false sentences about historical figures? We consider the answer fairly obvious.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.601

M&L discuss the case of ants unintentionally carving out a long path that spells out 'Peano proved that arithmetic is incomplete'. In contrast, as above, we receive a message from Luke saying the same thing. M&L then argue that "Luke's words mean something on their own (regardless of whether you or anyone else interprets them): namely, that Peano proved that arithmetic is incomplete. What Luke said is false: It was Gödel who proved incompleteness. But Luke said something, whereas the ants didn't say anything at all." The claim that Luke's words "mean something on their own" confuses the 'sense data' of orthographic patterns on a screen with "meaningful words", or the internalized algorithms we trigger from such data (van Lambalgen and Hamm 2004, 2005; Pietroski 2018; Pustejovsky and Batiukova 2019; Baggio 2018; Murphy 2024). The words alone do not do any referring, for the same reason that the orthographic information 'generated' by the ants does not do any referring. Luke's text message has a definite meaning not as an inscription, but in virtue of the communicative intentions we attribute to him. Luke said something that could be true or false only if there was a communicative act, backed by a particular referential intention.

M&L's premise here appears to us to be set on a problematic foundation. They write: "We are interested in the question of whether the outputs of LMs are more like the ants' patterns or like Luke's text: Do they merely resemble meaningful sentences, or are they in fact meaningful sentences?" But what would be the difference between resembling a sentence and being one? Luke's message also 'resembles' a sentence: How can we know that our phone is not infected by a virus or bug, and is genuinely showing the results of a human's communicative action? Both the ant's trail and Luke's text provide the exact same type of data to our language capacity. In the case of Luke, we assume there was a communicative intent, and so pragmatics kicks into action. But the difference is entirely in our internal states, that are triggered by or accompany the same or similar data. There are currently no compelling arguments that LMs can either form or infer communicative intentions (Bender and Koller 2020; Piantadosi and Hill 2022). M&L leave it as an "open question" whether communicative intentions require cognitive capacities that LMs lack. Our view is that LMs indeed lack those capacities. Consequently, LMs' outputs can only be evaluated relative to our own internal models of language and the world.

4. Crawl Out Through the Fallout: Machines in Linguistic Communities

In the Fallout's universe, the robot assistant Codsworth's voice is taken from a recording of a human being long since deceased. It animates the wasteland with accurate echoes of this voice, providing the illusion of it being part of a linguistic community, although in fact it is only a stochastic reassembly of sampled speech. The inclusion of LMs as *bona fide* members of linguistic communities is not unlike Codsworth's capacity to bring up encyclopedia entries in its memory and recite historical facts. They are both based upon the artificial agents' ability to simulate characters that appear to have the cognitive states that we would impute to humans showing comparable behaviors. Thus, one question is what counts as a 'member of a linguistic community': the LM, the simulacra it supports, both, or neither? Another question is: in what sense would any of these be members of linguistic communities? To try to answer these questions we must return to the original motivation for attributing reference to LMs' outputs.

The pressing question is whether and how expressions, including proper names and natural kind terms, 'refer' in LMs outputs, such that we may assess their pertinence and plausibility or truth. This is urgent also in connection with the growing use of language technologies in research (e.g., in literature summarization, hypothesis generation, etc.), where assessing just how accurate and trustworthy LMs' outputs are is crucial (Messeri

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.601

and Crockett 2024). We believe one reasonable path forward, which should however be trodden carefully, is to say it is simulacra (Shanahan, McDonell, and Reynolds 2023) that would count as (atypical) members of human linguistic communities. It is these virtual characters, and not LMs, that we typically interact with through language, and it is the quality of their outputs that we intend to evaluate. If an AI summarizes scientific articles about electrons, we want to consider 'electron' in its summaries as referring to electrons, a postulate in an explanatory scientific theory. The same would apply to outputs of other tasks, such as hypothesis generation, and to some other expressions, e.g., proper names. The virtual agent thus makes statements about the real world: this would make it a more viable member of a linguistic community than, say, a character in a science fiction story who talks about electrons, but whose claims we could only assess in a fictional world in which electrons might have properties they lack in the actual world, or vice versa.

What kind of linguistic community members would simulacra be, and in what ways would they be atypical? First, they would be limited by their own bounds of sense and reference. As noted above, the simulacra would inherit the restrictions of the LMs they are implemented by. For example, one might prompt a model with meta-rules for deictic resolution ('Imagine we are facing each other'), but this would create a fictional scenario rather than genuine indexical reference. The LM may well correctly generate outputs *as if* the simulacrum's bounds of sense and reference are not fixed by its physical structure. Still, this is not the same as achieving deictic grounding in the same physical and social spaces that the human interlocutor occupies. We suspect that this problem would apply to other versions of fictionalism, too (Mallory 2023) — to the extent that the object of the fiction includes the LM's behavior —, and that only endowing LMs with robotic bodies, integrated with their linguistic and cognitive abilities, will begin to alleviate it.

Second, LMs would be atypical members of linguistic communities also in that they would rely on humans to do the 'cognitive work' behind reference: e.g., we would have to project onto virtual agents the referential intentions that they cannot form, but that we would attribute to people who produce the same utterances. Third, they would be prone to errors that humans might not make, including drawing incorrect semantic inferences from subword tokens. Fourth, referential attributions to LMs would depend entirely on human willingness to sustain the fiction of community membership and the continuous cognitive supplementation it requires from human interpreters.

The extent of 'role play' is therefore significant: we pretend simulacra are temporary, atypical members of linguistic communities to evaluate their claims for truth and other norms, and we fill in cognitively for them. Historically, we do not have experience with other human members of linguistic communities that we expect are going to be limited in their semantic abilities, and that will depend on others for their possibility of making sense through language. No child or adult language learner can be subjected to anything less than the assumption that they could learn to use any language to its fullest capacity. LMs invite us to explore, both philosophically and empirically, alternative scenarios. We hope to have shown here that traditional answers are already off the table, and that even our currently more sophisticated answers are not entirely satisfactory. A combination of semantic internalism and role-play fictionalism is a promising framework for explaining the divergent semantic capabilities and limitations of humans and LMs. That machines are capable of meaning is a necessary fiction, sustained by human semantic cognition at all stages, from the generation of training data to the interpretation of machine outputs.

Acknowledgments
We wish to thank Harvey Lederman, Tal Linzen, and Matthew Mandelkern for their comments on an earlier version of this paper.

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.601

References

Apidianaki, Marianna. 2023. From word types to tokens and back: A survey of approaches to word meaning representation and interpretation. *Computational Linguistics*, 49(2):465–523.
Baggio, Giosuè. 2018. *Meaning in the Brain*. MIT Press, Cambridge, MA.
Bender, Emily M. and Alexander Koller. 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*, pages 5185–5198.
Block, Ned. 1987. Advertisement for a semantics for psychology. *Midwest Studies in Philosophy*, 10:615–678.
Cappelen, Herman and Josh Dever. 2021. *Making AI Intelligible: Philosophical Foundations*. Oxford University Press, Oxford.
Carey, Susan. 2009. *The Origin of Concepts*. Oxford University Press, Oxford.
Chomsky, Noam. 2000. *New Horizons in the Study of Language and Mind*. Cambridge University Press, Cambridge.
Cohnitz, Daniel and Jussi Haukioja. 2013. Meta-externalism *vs* meta-internalism in the study of reference. *Australasian Journal of Philosophy*, 91(3):475–500.
Cohnitz, Daniel and Jussi Haukioja. 2025. *Foundations for Metasemantics*. Oxford University Press, Oxford.
van Deemter, Kees. 2016. *Computational Models of Referring: A Study in Cognitive Science*. MIT Press, Cambridge, MA.
DeVault, David, Iris Oved, and Matthew Stone. 2006. Societal grounding is essential to meaningful language use. In *Proceedings of the National Conference on Artificial Intelligence*, pages 747–754, Menlo Park, CA.
Haslett, David A. 2025. Tokenization changes meaning in large language models: Evidence from Chinese. *Computational Linguistics*, pages 1–30.
Haspelmath, Martin. 2023. Defining the word. *Word*, 69(3):283–297.
Hinzen, Wolfram. 2016. On the grammar of referential dependence. *Studies in Logic, Grammar and Rhetoric*, 46(59):11–33.
Jackendoff, Ray. 2002. *Foundations of Language*. Oxford University Press, Oxford.
Kripke, Saul. 1980. *Naming and Necessity*. Harvard University Press, Cambridge, MA.
van Lambalgen, Michiel and Fritz Hamm. 2004. Moschovakis' notion of meaning as applied to linguistics. In *Logic Colloquium '01, ASL Lecture Notes in Logic*, pages 255–280, A. K. Peters, Wellesley, MA.
van Lambalgen, Michiel and Fritz Hamm. 2005. *The Proper Treatment of Events*. Blackwell-Wiley, Oxford.
Lederman, Harvey and Kyle Mahowald. 2024. Are language models more like libraries or like librarians? Bibliotechnism, the novel reference problem, and the attitudes of LLMs. *Transactions of the Association for Computational Linguistics*, 12:1087–1103.
Mallory, Fintan. 2023. Fictionalism about chatbots. *Ergo an Open Access Journal of Philosophy*, 10.
Mandelkern, Matthew and Tal Linzen. 2024. Do language models' words refer? *Computational Linguistics*, 50(3):1191–1200.
Messeri, Lisa and M.J. Crockett. 2024. Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627:49–58.
Mielke, Sabrina J., Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Y. Lee, Benoît Sagot, and Samson Tan. 2024. Between words and characters: A brief history of open-vocabulary modeling and tokenization in NLP. *arXiv*, (2112.10508).
Murphy, Elliot. 2016. Phasal eliminativism, anti-lexicalism, and the status of the unarticulated. *Biolinguistics*, 10:21–50.
Murphy, Elliot. 2023. The citadel itself: Defending semantic internalism. *Global Philosophy*, 33(7).
Murphy, Elliot. 2024. What is a word? *arXiv*, (2402.12605).
Murphy, Elliot, Emma Holmes, and Karl Friston. 2024. Natural language syntax complies with the free-energy principle. *Synthese*, 203(154).
Ostertag, Gary. 2025. Language models and externalism: A reply to Mandelkern and Linzen. *Computational Linguistics*, 51(2):651–659.
Pavlick, Ellie. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A*, 381(20220041).
Piantadosi, Steven T. and Felix Hill. 2022. Meaning without reference in large language models. *arXiv*, (2208.02957).

Computational Linguistics Just Accepted MS.
https://doi.org/10.1162/COLI.a.601

Pietroski, Paul. 2005. *Events and Semantic Architecture*. Oxford University Press, Oxford.
Pietroski, Paul. 2017. Semantic internalism. In James McGilvray, editor, *The Cambridge Companion to Chomsky*. Cambridge University Press, Cambridge, pages 196–216.
Pietroski, Paul. 2018. *Conjoining Meanings: Semantics Without Truth Values*. Oxford University Press, Oxford.
Pustejovsky, James and Olga Batiukova. 2019. *The Lexicon*. Cambridge University Press, Cambridge.
Riedl, Martin and Chris Biemann. 2018. Using semantics for granularities of tokenization. *Computational Linguistics*, 44(3):483—524.
Scott-Phillips, Thom. 2014. *Speaking Our Minds*. Palgrave MacMillan, London.
Shanahan, Murray, Kyle McDonell, and Laria Reynolds. 2023. Role play with large language models. *Nature*, 623(7987):493–498.
Søgaard, Anders. 2025. Do language models have semantics? on the five standard positions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 25910–25922, Association for Computational Linguistics, Vienna, Austria.