Pragmatic Norms Are All You Need – Why The Symbol Grounding Problem Does Not Apply to LLMs

Reto Gubelmann
Digital Society Initiative & Department of Computational Linguistics
University of Zurich
reto.gubelmann@uzh.ch

Abstract

Do LLMs fall prey to Harnad's symbol grounding problem (SGP), as it has recently been claimed? We argue that this is not the case. Starting out with countering the arguments of Bender and Koller (2020), we trace the origins of the SGP to the computational theory of mind (CTM), and we show that it only arises with natural language when questionable theories of meaning are presupposed. We conclude by showing that it would apply to LLMs only if they were interpreted in the manner of how the CTM conceives the mind, i.e., by postulating that LLMs rely on a version of a language of thought, or by adopting said questionable theories of meaning; since neither option is rational, we conclude that the SGP does not apply to LLMs.

1 Introduction: LLMs, Understanding Meaning, and Symbol Grounding

In recent years, the field called natural language understanding within natural language processing (NLP) has seen remarkable progress. After the better-than-human performance of encoder-only transformers at benchmarks that were explicitly designed to be challenging for them,[1] large generative (decoder-only) transformer-based language models (LLMs) have shown impressive performance at tasks where they have never been explicitly trained for. For instance, according to OpenAI, gpt-4 scores in the 90th and 99th percentile ranks at the Uniform Bar Exam and the GRE verbal respectively.[2]

So, given these impressive results, do these LLMs *really* understand language? To many practitioners of NLP as well as to many linguists, the idea that one could acquire any understanding of the real meaning of words such as "apple" or "dog" solely by processing large amounts of textual sequences seems intuitively implausible. It would seem that to understand what "dog" actually means, one needs to have at least seen one, and ideally also touched and interacted with on one. Informally, this is what so-called symbol grounding is about: connecting symbols to the entities to which they refer.

This intuition of the need for grounding is one of the driving argumentative forces of Bender and Koller (2020). In this influential contribution, the authors suggest that, given what "meaning" means, that is, for conceptual reasons, it might be impossible that LLMs as we know them could achieve real understanding of linguistic meaning. Briefly, their point is that linguistic meaning involves a mapping of words onto the real-world entities to which these words refer. This mapping, however, is likely not learnable by the kind of string-based training that LLMs undergo. Bender and Koller (2020, 5188) identify this impossibility with the symbol grounding problem (SGP) by Harnad (1990).

We think that the notion of LLMs' falling prey to the SGP is mistaken, and that the main cause for the confusion lies on the conceptual, and hence philosophical level (as opposed to the empirical-technical level, see the appendix, section C, for details on this distinction), and that there is a danger that significant resources are invested into solving a nonexistent problem, namely the SGP as applied to LLMs. We would like to contribute towards resolving this confusion by disentangling different notions of language, of meaning, and of grounding in play here and thereby also show how philosophy has the potential not only to unnecessarily stall, but also to contribute to progress in NLP.

Our contribution is threefold. (1) Starting out with a critical discussion of the main argument of Bender and Koller (2020), we distinguish two different uses of symbol grounding in the NLP literature, an empirical and a philosophical one. (2) We show how the latter can be dissolved with regard

[1] See the leaderboards of the GLUE and SuperGLUE benchmarks, Wang et al. (2018) and Wang et al. (2019).

[2] See this document, last consulted on January 31, 2024.

Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, pages 11663–11678
November 12-16, 2024 ©2024 Association for Computational Linguistics

to natural language symbols with recourse to the norm-governed context of use of natural language. (3) We argue that the philosophical SGP doesn't apply to LLMs either, as the explanatory grounding of the SGP is not available there.

Making progress on these topics is important for overall progress in NLP. The symbol grounding problem is a well-entrenched issue in current debates on the principled limits of LLMs; disambiguating it contributes to clarity and hence to progress. Furthermore, if Bender and Koller were right in their suggestion that current LLMs are climbing up the wrong hill when trying to reach general natural language understanding, this would imply that the enormous resources invested in training ever larger LLMs might deliver, perhaps, ever more helpful tools, but will never be able to reach actual understanding of meaning, let alone artificial general intelligence. More generally speaking, what one understands by meaning informs one's notion of what it takes to understand meaning, which in turn determines the kind of tool that they would build with the goal of understanding meaning.

Making progress on these topics is difficult because it requires the cooperation of three fields of inquiry: NLP, theoretical linguistics, and philosophy. Interdisciplinary research, while much-sought after currently[3], is extremely difficult. Too often, misunderstandings emerge and do more bad than good.

The paper is structured as follows. By means of Bender and Koller's so-called Octopus Test, we distinguish two senses of symbol grounding (Section 2), we review relevant literature (Section 3), and we delineate the origins of the SGP in the computational theory of mind as well as its expansion to linguistics (Section 4). We then build on this to argue that, as a matter of fact, the SGP neither applies to natural language (Section 5), nor to LLMs (Section 6).[4]

2 Which Symbol Grounding Problem? – Take The Octopus Test!

Bender and Koller (2020, 5188f.) introduce the so-called Octopus Test as their main argument for their claim that LLMs cannot really understand linguistic meaning because their symbols are not grounded. In this thought experiment (for an introduction to this technical term, see Brown and Fehige 2023), we are invited to imagine two people, *A* and *B*, stranded on separate islands that are connected via a telegraph cable. Both islander *A* and islander *B* are alone on their respective islands, so they are very happy to find that they can communicate with a fellow human being via telegraph and start using it. What they do not know is that a hyper-intelligent deep-sea octopus starts listening in on the communicative signals running through the telegraph cable. The octopus, representing the LLM, starts picking up patterns in their communication and, at some point, cuts the cable and starts pretending to be either *A* or *B* in communication with *A* or *B* respectively.

Now, Bender and Koller maintain that, like the LLMs, the octopus only receives form, not meaning, through the signals. By form, they understand "any observable realization of language" (Bender and Koller, 2020, 5186), while meaning, as well as understanding meaning, requires word-world mappings (see above, Section 1, for details, see below, Section 5). These mappings, however, cannot be learned from form alone, which means that current LLMs probably cannot understand language. This in turn, the authors argue, puts principled limits on the topics of conversation in which the octopus will be able to pass as an islander. In particular, the authors claim that the octopus would be exposed if *A* would all of a sudden be attacked by a bear and, in deep panic, beg *B* for help to defend against the bear using a couple of sticks. According to Bender and Koller, "[i]t is at this point that O would fail the Turing test" (Bender and Koller, 2020, 5189) because "[h]aving only form available as training data, O did not learn meaning". In other words, *O* cannot advise *B* in using the sticks to defend against the bear because it has never connected words such as "stick" to actual sticks.

As a matter of practical fact, however, having input data other than what Bender and Koller call pure form is neither sufficient nor necessary to be able to respond competently in this scenario. As Sahlgren and Carlsson (2021) argue, from the entire human population only the small minority of bear-handling-specialists would be helpful advisors to the threatened islander. Still, none of us would as a consequence refuse to credit the remaining majority with linguistic understanding. It seems not

[3] For example, in the NLP Community Survey of 2023, 83% of respondents stated that more interdisciplinary research is needed (Michael et al., 2023).

[4] We are grateful for the excellent comments and criticisms by three anonymous ARR reviewers, by the participants of the 2024 AIAI Workshop in Göttingen, and for discussions with Rick Nouwen about formal semantics.

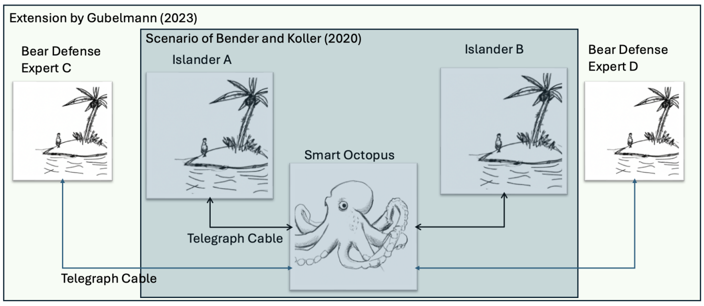

Figure 1: An Illustration of the Octopus Test proposed by Bender and Koller (2020) (blue area) as well as the extension of the experiment by Gubelmann (2023) (larger green area).

necessary either since, as Gubelmann (2023, 510-513) suggests, the octopus in Bender and Koller's thought experiment might not need a different kind of input, but rather different input of the same kind, namely large amounts of conversations about effective bear-defense. Gubelmann (ibid.) invites us to imagine that, in addition to listening in on *A*'s and *B*'s conversation, the octopus also listens to *C*'s and *D*'s conversations, which are islanders as well who have become world-class bear defenders and talk about nothing else than fending off bears with sticks. Listening in on *C*'s and *D*'s conversations for long enough, Gubelmann argues, would allow the Octopus to give a proficient response to Bender and Koller's islander's cry for help (without having any word-world-relationships). See figure 1[5] for an illustration of Bender and Koller's set-up and Gubelmann's extension of it.

As a consequence, whoever looks at this outcome of the Octopus Test and still feels a need for symbol grounding is not occupied with a specific engineering challenge: We have seen that the octopus can fulfill the task specified by Bender and Koller perfectly without any grounding.[6] Rather, they are worried about what we will call the philosophical SGP, or simply the SGP. In other words: As a simple test to distinguish the engineering from the philosophical sense, one can ask whether the problem in question would disappear if theoreticians would stop thinking about the phenomenon in the way they currently do. If it is an engineering problem, it would not: it would still be the case that the system in question lacks the capacities that require grounding. If it is a philosophical kind of SGP, however, the problem would simply cease to exist.

3 Overview on Recent Research

Recently, this question of whether LLMs are subject to the SGP as well as related topics have been discussed by a number of researchers. Pavlick (2023) is mostly interested in the question whether LLMs can serve as models of human language processing. She maintains, in loose analogy to the position taken in this paper, that LLMs can be said to "encode meaning" without what she calls explicit grounding. Mandelkern and Linzen (2024), with regard to the slightly different question of whether the words generated by LLMs refer, argue that the natural histories of the text that has served as training data for the LLMs to bring tokens generated by LLMs in referential contact with real-world entities. Piantadosi and Hill (2022), finally, distinguish, as we have done, between meaning and reference, then argue, *pace* Bender and Koller, that meaning can be had without reference, and finally suggest that the performance of LLMs evidences that they learn conceptual roles, which provides strong grounds to suppose that they learn meaning.

Quilty-Dunn et al. (2023, 5) accept that neural networks like the transformer are no candidates for attribution of a so-called Language of Thought which will emerge as a precondition for the SGP (in Section 4.1 and ultimately in Section 6); rather, as adherents of the scientific potential of the Language of Thought hypothesis, Quilty-Dunn et al. (2023, 4) point to alternatives to the transformer architecture, especially versions of symbolic AI supplemented with Bayesian probabilistic inference to suggest that the Language of Thought has still potential even in computer science. What they do not even begin to argue for is that the transformer itself (or other deep neural networks) can be seen as operating with a Language of Thought.

[5] The pictures used in this and the following figures were drawn by the author with support from dall·e.

[6] For an overview on these engineering challenges, see Bisk et al. (2020, 8722), for two examples, see the Appendix, Section A.

Mollo and Millière (2023), finally, address a topic that is most directly relevant to this paper and will be discussed in detail below (Section 6). They examine what they call the "vector grounding problem", which they consider the analogon for LLMs to the symbol grounding problem of Harnad (vectors replacing symbols). They argue that the grounding problem (which they specify as a case of referential grounding) does indeed apply to LLMs, but that it can be solved by using reinforcement learning on the models, which imbues them with causal-historical input from humans and hence establishes referential relations; more speculatively, they argue that, in specific zero-shot in-context-learning scenarios, they might also acquire grounding. Among these settings, Mollo and Millière (2023, 24) mention the task of mapping color terms like cyan to RGB color spaces. Millière and Buckner (2024, 17) review this line of reasoning approvingly.

4 The (Philosophical) Symbol Grounding Problem

In stark contrast to the *engineering sense* of the grounding problem, where there is an observable problem in the performance or capacity of AI-systems to be solved, there is another sense of the symbol grounding problem that is not about a specific shortcoming in performance; rather, it has been raised explicitly in spite of flawless performance. This *philosophical sense* is the original sense of the SGP, and it is this sense which still exercises many practitioners in NLP. We first sketch the origin of the symbol grounding problem in the philosophy of mind, then we show how it expanded to linguistics.

4.1 Its Origin in the Computational Theory of Mind (CTM)

The philosophical version of the symbol grounding problem, what has been and will be referred to by "SGP", originates from a very specific tradition of thinking about the human mind in the philosophy of mind, namely the so-called computational theory of mind (CTM, Rescorla 2020), a version of functionalism that maintains that the mind is a computing system. Broadly speaking, functionalism is the idea that an entity is to be identified by nothing else than the function it performs (hence the term) in a process. The biological or physical realization of the function is considered irrelevant (Arkoudas and Bringsjord, 2014, 43).

The CTM is not necessarily symbolic, or as it is expressed in the field, representational (Rescorla, 2023), that is, the tokens that are functionally identified do not necessarily have to represent (read: have a meaning). However, largely thanks to the work of Fodor (1975, 1983, 1987, 1992, 2008), a representational version of the CTM where the tokens functionally identified are actually meaningful symbols has become dominant. As a consequence, we use CTM to refer to representational CTM. Fodor famously referred to this version of the CTM as the Language of Thought, or Mentalese (Rescorla, 2023). For some more details on the CTM, see the appendix, section B. The CTM has been advocated as ideally suited to explain certain abilities of the human brain/mind, such as the recursive generation (and understanding) of potentially infinitely many sentences and the systematicity of language.

The Chinese Room Thought Experiment, proposed by Searle (1980), was directly aimed at the CTM, and it runs as follows. Searle, who does not understand any Chinese, is locked in a room with very detailed rule-books. Thanks to these rule-books, he is able to write sensible responses in Chinese to questions in Chinese that he receives through a lid into the room. While he has no idea what the strange-looking shapes mean that he is receiving and returning, it seems to people outside of the room that he understands Chinese. From the point of view of the CTM, it would seem that the squiggles and squoggles that Searle is manipulating are actually meaningful symbols, as they are embedded in a functionals structure that results in well-formed, sensible (at least to the recipient) linguistic messages. The moral to be drawn is that, just like the squiggles and squoggles lack meaning, so do the supposed Mentalese symbols postulated by CTM.

Note that **multimodality** was already a topic in Searle (1980): The so-called robot reply to the Chinese Room suggests to solve the issue essentially by conceiving the Chinese Room as the head of a robot, equipped with sensors and able to move. However, it is generally agreed that this does not solve the basic issue raised by the Chinese Room: Whether Searle receives the input he does not understand via sheets through a lid or in the form of sensory signals is irrelevant (see the collection by Preston and Bishop 2002).

Harnad (1990, 336) explicitly references Fodor's conception of language of thought as well as the computational theory of the mind in general as the target of his symbol grounding problem, and he also refers to the Chinese Room Thought Experiment to motivate his SGP: Just like Searle's squiggles and squoggles, our Mentalese symbols need to be grounded in something non-symbolic to be meaningful. See figure 2 for an illustration of this very specific set-up that grounds the symbol grounding problem. The problem is how to connect the Mentalese symbols to their referents.

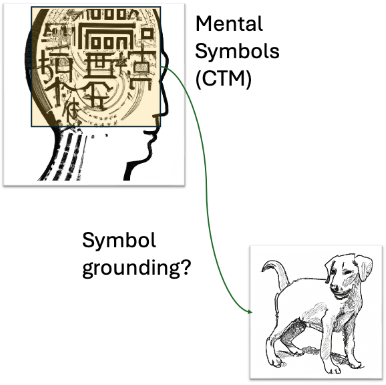

Figure 2: An illustration of the theoretical set-up grounding the symbol-grounding problem.

Within *this* conception of a language of thought, and more broadly within the idea that the human mind is a manipulator of non-linguistic symbols, Harnad's symbol grounding problem seems perfectly sensible: How is it possible that this Mentalese, on which humans rely for thinking, but of which they are not conscious, and which they certainly cannot manipulate in any conscious way, has any meaning at all? For an overview of the CTM's struggle with the SGP, compare Taddeo and Floridi (2005), who review fifteen years of discussion that Harnad's symbol grounding problem has sparked, distinguishing eight types of attempted solutions to it, and concluding that it remains unsolved to this day.

4.2 Its Expanding to Formal Semantics

In a surprising turn of events, the SGP expanded from a certain conception in the philosophy of mind, namely the CTM, to pertinent theories in the philosophy of language and linguistics. This section focuses on its appeal from the point of view of formal semantics.

Bender and Koller also touch upon this issue by addressing model-theoretic semantics (MTS), a kind of formal semantics for natural languages (Luo, 2014, 177). When introducing their notion of meaning, they maintain that all that they are presupposing is that "conventional meanings must have interpretations, such as a means of testing them for truth against a model of the world" (Bender and Koller, 2020, 5187). This is a model-theoretic notion of a world, a purely abstract concept used in formal semantics and pioneered by Montague et al. (1970); with elaborations by Kripke (1983) and Plantinga (1974).

MTS has proven a highly useful tool to study language, in particular to represent inferential relationships between claims.[7] As Gochet (2011) shows nicely, it can explain the notion of logical entailment merely with recourse to generality, without any reference to modal notions of necessity. Furthermore, harnessing the powerful formalism of post-Fregean predicate logic, MTS has been able to make explicit in many ways how Syntax impacts semantics (e.g., in the case of quantifier scopes). In doing so, it relies on the notion of a non-empty domain of individuals over which its variables can range. Then, for any given n-ary predicate P, satisfaction of this predicate is determined by an interpretation I that maps a set of n-tuples from the universe of discourse onto each n-place predicate (Gochet, 2011, 173) – this set being the things of which P is true, informally speaking.

This interpretation I, however, still leaves some theorists wondering how these elements of the universe of discourse can be grounded in the real material world of sticks and stones. In particular, in their Octopus Test, Bender and Koller capitalize on the need to map words onto real-world entities like sticks and stones to understand their meanings.

So, this version of the grounding problem arises if we ask how the abstract entities in the universe of discourse (which are, in the case of modal MTS, called possible worlds) can possibly refer to any real-world entities like sticks and stones. One could say that it is a second-order SGP, as the entities in the universe of discourse are themselves used to give an interpretation to predicate constants that can be used as representations of natural language concepts. For instance, "Px" might be used to represent "x is a stick". Now, the interpretation tells us, for any member of the universe of discourse, whether it satisfies P. However, when analyzing natural language, we might want to pick out the class of real-world entities of which it is true that

[7] See Peregrin (1997) for a conclusive argument as to why MTS is a theory of inference and not of meaning.

Aspect	Correspondence Theories of Meaning	Pragmatic Theories of Meaning
Grand Picture of Language	A symbol system constituted of syntax and semantics	A social practice of norm-governed use
How Linguistic Meaning is Constituted	Via a mapping between linguistic expressions and non-linguistic entities	Via convential norms that regulate correct and incorrect use
Typical Primary Level of Meaning-Constitution	Subsentential Elements, e.g., concepts	Speech Act (usually on propositional level)

Table 1: General view on differences between correspondence and pragmatic theories of meaning.

they are sticks. How can we map the abstract set onto the real-world stuff?

This, then, allows the SGP from the CTM to formal semantics: Instead of Mentalese symbols, it is the symbols employed in the formalism of MTS that are found to be in need of grounding. Don't the abstract entities employed in MTS need to be grounded in the real world of sticks and stones just like Mentalese symbols? One could think that, unless the individuals of the universe of discourse over which we quantify in our MTS are mapped onto individuals in the real world, their meaning and reference would probably have to go by the board.

5 Why the SGP does not Apply to Natural Language

The basic argument that we will present in this section is that the SGP – both with regard to natural language as well as regarding its systematic abstraction in formal semantics – dissolves if we consider the pragmatic context in which any natural language is at home.

We begin by sketching a somewhat generic notion of correspondence theory of meaning (Munson, 1962, 42), which will be called following Wittgenstein (1958, §1) the Augustinian Picture. In the Augustinian Picture, (1) the primary view of natural language is one of a complex symbol system encompassing syntax and semantics that is at root abstracted from its material, historical, pragmatic, social, etc., context, (2) its meaning-constituting relation is the mapping of linguistic expressions to non-linguistic objects to which they correspond, e.g., the famous so-called middle-sized dry goods such as sticks and stones. This then invites the principle (3) that the typical fundamental unit of linguistic meaning is the concept.

Bender and Koller (2020) develop a correspondence conception of meaning that involves both conventional meaning and the communicative intention that a speaker pursues with a specific utterance. In both variants, meaning requires mapping words to the real-world entities to which they refer. The precise relationship that this said to be needed between words and the real world for the words to be meaningful is conceived differently.[8] The underlying notion, however, is the same in all of these different uses, namely **the need for linguistic symbols to, as it were, hook up to their non-symbolic referents**. Given such a correspondence theory of meaning, it follows that one has to be able to map the words onto the objects to which they refer or correspond to. Without this meaning-constituting correspondence connection, the words of any natural language are literally meaningless. Furthermore, any being, such as the octopus in in Section 2 or an LLM, that is unable to, as it were, reach beyond language and connect language to these non-linguistic entities, is categorically unable to understand linguistic meaning.

Correspondence theories of meaning, however, are by no means the only, and arguably not the most convincing, kind of theories of meaning of natural language.[9]

Therefore, we suggest to take a step back and get **a more comprehensive view of the phenomenon of language** by taking its social and contextual institution and constitution seriously. This is what so-called pragmatists have advocated for decades (Legg and Hookway, 2024); we here follow the version of pragmatism developed by Brandom 1994a, 2010, 2021; Hlobil and Brandom 2024, which is further detailed in the Appendix, Section D.1. Briefly, compared to the three principles of the Augustinian Picture, pragmatism replaces (1) with conceiving natural language primarily as a norm-

[8] As the *grounding* of a linguistic element in the real world (Bender and Koller, 2020, 5185, 5187, 5188, 5190), as *connecting* to the real world (ibid., p. 5188, 5188, 5190), as *mapping* between words and real-world entities (ibid., 5189), or as *being about* things in the real world (ibid, p. 5190).

[9] For a very forceful and highly influential critique of correspondence theories of meaning, see Wittgenstein (1958, §1ff.).

governed social practice for specific human and societal uses. With regard to (2), it sees linguistic meaning of symbols not in mappings to non-symbolic entities, but rather in the norm-governed way in which they are being used by a community of speakers (the norms and the community are mutually constitutive). This then readily leads over to (3): The fundamental unit of meaning is typically not located on the level of the concept, but rather the smallest linguistic unit that is typically used to do things in real life, namely the speech act, typically expressed in a proposition/sentence. See Table 1 for an overview.

This is just a rough sketch of pragmatism, which is admittedly strongly influenced by Brandom. However, the three principles do justice to the main tenets of pragmatism broadly conceived, as pioneered by Wittgenstein (1958). Furthermore, it also aligns with the influential speech act theorists Austin (1962) and Searle (1969), who have emphasized that the primary locus of language is its use in speech acts rather than as a system standing in splendid isolation. With regard to (3), our position aligns with Kant (1998 [1781/1787]) and Frege (1892). In the 20th century, important proponent is Quine (1974), and more recently Frápolli (2019).

Emphasizing the fact that natural language – unlike Mentalese – is constantly used, and pointing out that this use determines the meaning of utterances and the reference of concepts, allows us to resolve the puzzle that initiated this section by pointing to a rather mundane source of meaning for sentences, and hence for words: the users of language. Whether “this is a dog” or “tkr br drkg” has any meaning at all and, correspondingly, is able to refer to any non-linguistic entity, depends on the existing practices of a community of speakers. These practices are governed by norms that determine that “this is a dog” can be used in a speech act – to claim that the relevant object pointed out is a dog –, while “tkr br drkg” cannot. So, echoing the SGP represented in figure 2, we can easily resolve the mystery in the case of natural language by pointing to conventional norms, see figure 3.

This in turn allows us to embed the phenomena examined by formal semantics such as MTS using its powerful analytic tools back into the actual use of language in human societies, as described by pragmatist perspectives on language. In short, while MTS serves as a highly successful tool to analyze relationships between meaningful statements and referring concepts, it is the conventional norms that govern the use of both that ensure that these statements, including the abstractions and analyses of them by MTS, have meaning or refer respectively. **Note that it is precisely this context of norm-governed use that is missing with Mentalese symbols**: Being unconscious, non-linguistic elements of the mental computer, it is obvious that they are not used in a norm-governed way by humans and hence cannot receive their meaning and reference from there.

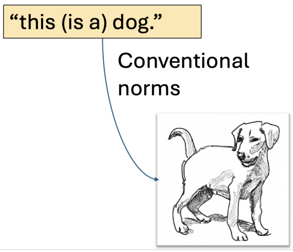

Figure 3: A pragmatic resolution of the SGP for natural language.

In sum, in this section, we have suggested that **whether the SGP gains any traction with regard to natural language is a question of the theory of linguistic meaning that one embraces**. On the Augustinian Picture, and in general on correspondence theories of meaning as such, the SGP does indeed come up. However, on the pragmatic conception of linguistic meaning that we have sketched and recommended here, it is conventional norms, defined by and defining speaking communities that establish reference between certain symbols and the non-symbolic entities which they signify and thereby prevent any SGP in the first place.

6 Why the SGP does not Apply to LLMs

After having discussed the origins of the SGP in the CTM, and after having shown how it has expanded into formal semantics and natural language, and also how the SGP can be dissolved there by reconsidering language in the everyday context of norm-governed use, we are now finally ready to consider the central question of this article: Does the SGP apply on LLMs?

As we have seen above (Section 2 and Section 5), (Bender and Koller, 2020) worry that LLMs, being trained solely on strings of text, cannot get the meaning, that is, on their notion of meaning, the world-items corresponding to the strings that they process. However, as we have argued, this theory of meaning is questionable. If we take into view that it is conventional norms, constituting and constituted by the social practice of speaking a language, that give meaning to claims and concepts, then we can see that LLMs might pick up linguistic meaning: The norms observed by language users will leave recognizable patterns in the training data which the LLMs can represent and with them infer the norms governing the use of expressions, that is, the meanings of these same expressions.

So, while the correspondence theory of meaning would predict, as Bender and Koller did, that there are linguistic tasks for which LLMs are not qualified because they lack the relevant exposure to the non-linguistic meaning of words, the pragmatic conception of meaning developed here would predict the contrary, namely that there is no principled limit to the linguistic performance of LLMs. We submit that (albeit with the gift of hindsight compared to Bender and Koller 2020) empirical evidence clearly favors the pragmatic side here: Time and again, and even more so after the publication of ChatGPT on November 30, 2022, LLMs have managed to shine in one linguistic challenge after another that was previously thought be beyond them. Hence, for the theoretical and empirical reasons sketched, we should abandon the correspondence theory of meaning in favor of a pragmatic one; and with the correspondence theory, Bender and Koller's reason for applying the SGP to LLMs also goes by the board.

This verdict contrasts with Mollo and Millière (2023), who insist that there is a symbol grounding problem that needs a solution. Mollo and Millière (2023, 7-8) justify the existence of such a problem with reference to the Octopus Test introduced and critically assessed above (Section 2), emphasizing that "[t]he vectors they [the LLMs, RG] receive and manipulate during internal processing are merely meaningless arrays of numbers, ungrounded in anything outside statistical patterns in language, just like the outputs they generate" (ibid, p. 8).

This shows that, unlike Bender and Koller (2020), Mollo and Millière (2023) want to anchor the SGP not in a specific theory of linguistic meaning, but rather in the inner workings of LLMs, which puts them closer to the origins of the SGP in the CTM than to its location in linguistic theories. Specifically, Mollo and Millière (2023, 6) argue that the vectors that represent the input tokens face the SGP and require some specific sort of grounding, namely referential grounding.

So, does the SGP apply to the inner workings of LLMs, including their vectors? To begin answering this question, note that the CTM as well as the Chinese Room Thought Experiment that gives traction to the SGP are both firmly situated in the *good old-fashioned, rule-based* world of AI (compare Quilty-Dunn et al. 2023, 5). LLMs, in contrast, are *connectionist, statistical devices that have no intrinsic symbolic structure*. It is, however, this symbolic structure that is presupposed in the CTM, the Chinese Room Thought Experiment, and hence also in the symbol grounding problem. See figure 4 for an illustration of why the SGP fails to get traction with statistical methods such as transformers: there are no symbols that would need grounding.

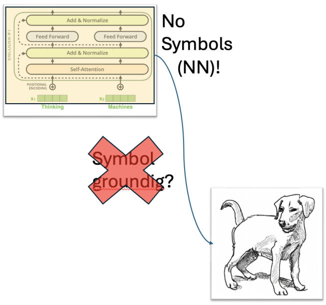

Figure 4: Why the SGP does not directly apply to transformers (source: author and Alammar 2018).

This implies that to make plausible that LLMs are subject to the SGP, one would have to find reasons interpret their inner going-ons in a way that is analogous to how the CTM interprets the human mind, namely by postulating non-linguistic symbols – a kind of LLM-Mentalese – that are being manipulated by an LLM to perform its functions. In other words, one would have to credit the LLMs with a version of Fodor's Language of Thought.

What reasons could there be to postulate a Language of Thought in LLMs? To answer this question, it is helpful to return to Fodor's original argument for his Language of Thought in human minds: Fodor introduced the Language of Thought hypothesis to account for the productivity and the systematicity of thought and thinking (Rescorla, 2023). The productivity of thought consists in the fact that we as competent language users can potentially produce an infinite number of well-formed sentences; often, recursive phenomena are used to make this point (a simple example for this would be to add "'s daughter": to the end of the output of the last recursion: "Sarah is Mary's daughter" –> "Sarah is Mary's daughter's daughter", etc.). The systematicity of thought consists in systematic interrelations between thoughts that we can entertain; one of the prime examples is the transitivity of deductive inference. Fodor explicitly motivates his Language of Thought hypothesis with recourse to these properties of human language and thought.

However, there does not seem to be any solid, reliable evidence that LLMs actually are truly productive in Fodor's sense. While it is rather simple to create a GOFAI-algorithm that uses recursion to create as many different sentences as the computing architecture used allows, it seems less obvious that this is also possible with purely associative, implicit systems such as LLMs, and there is, to the best of the knowledge of the author, no study available that would provide positive evidence for it. The situation seems even worse regarding systematicity. As the empirically established problem of generalization in NLI shows (see the Appendix, Section D.2), and as Asher et al. (2023) argue on a theoretical level, LLMs are not particularly good at such systematic thinking, such as drawing and labelling deductively valid inferences.

In sum, then, this presents the following picture. First, as LLMs are not instances of symbolic AI, it requires a positive reason to postulate that their inner workings proceed by processing of some kind of symbols that would then have to be grounded. In the case of the human mind, Fodor has argued for this by referring to the systematicity and the productivity of thought. Hence, if Mollo and Millière (2023) want to read the token embedding vectors of LLMs as symbols requiring some sort of grounding, they should, as Fodor did, make explicit the explanatory need that exists for positing such symbols (that could then be seen as in the need of grounding).

Mollo and Millière (2023) might respond to that with the emphatic insistence that, regardless of any explanatory benefits, the embedding vectors certainly need to mapped onto the real-world entities to which they refer, or, as they put it (ibid, p. 28): "hook onto the world". This moves the discussion on the level of theories of meaning, where my response would be to point to the sketch of the two different paradigms of theories of meaning displayed in Table 1: Within correspondence theories, some such mapping might be required. However, they are not the only, and arguably not the most convincing kind of theory of meaning: A pragmatist theory of meaning does not need any hooking of words to the world (whatever that might mean). All it needs is a norm-governed practice in a society, the patterns of which can be picked up by LLMs from the training data and used to infer said norms.

7 Conclusion

In this article, following Bender and Koller (2020, 5192), we have focused on conceptual-philosophical considerations of NLP research. Specifically, we have considered their claim that LLMs are vulnerable to the so-called symbol grounding problem (SGP), as introduced by Harnad (1990). After distinguishing the SGP, which is a philosophical problem, from empirical *Doppelgängers*, we have shown how it has emerged from the theoretical background of the Computation Theory of Mind (CTM) and from there expanded to linguistic theories, including model-theoretic semantics (MTS). We have then argued that MTS, understood properly, does not give rise to the SGP, as the powerful formalism that it furnishes is best understood as systematically analyzing pre-existing meaningful statements, which in turn receive their meaning from norm-governed use. Furthermore, with regard to natural language, the SGP arises only if one adopts a correspondence theory of meaning, which is in competition with more comprehensive pragmatic conceptions of meaning that elegantly dissolve the SGP by bringing into view the conventional norms that govern language use.

With regard to LLMs, we have found that, given that they are neural network rather than GOFAI architectures, there is no reason to assume that the SGP arises, as it would require postulating Mentalese-like symbols in the LLMs, which in turn would require an explanatory need to do so which we are currently lacking.

Of course, the questions whether LLMs, being the kind of thing they are, can understand meaning, properly conceived, remains an important conceptual and empirical question. However, by carefully delineating the conceptual landscape surrounding the SGP, we hope to have shown that NLP researchers can safely stop worrying about the SGP and go on about their business, and thus to have provided a case in point where philosophy can foster progress in NLP.

8 Limitations

The final argument that the SGP does not apply to LLMs depends on the empirical fact that there are no empirical reasons (i.e., explanatory needs) to postulate in LLMs a version of Mentalese. However, if future research in the field would bring to light decisive reasons for such a postulation, the overall argument would have to be reconsidered.

References

Jay Alammar. 2018. The illustrated transformer [blog post].

Gabor Angeli and Christopher D Manning. 2014. Naturalli: Natural logic inference for common sense reasoning. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 534–545.

Konstantine Arkoudas and Selmer Bringsjord. 2014. Philosophical foundations. In *The Cambridge Handbook of Artificial Intelligence*, pages 34–63.

Dimion Asael, Zachary Ziegler, and Yonatan Belinkov. 2021. A generative approach for mitigating structural biases in natural language inference. *arXiv preprint arXiv:2108.14006*.

Nicholas Asher, Swarnadeep Bhar, Akshay Chaturvedi, Julie Hunter, and Soumya Paul. 2023. Limits for learning with language models. In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 236–248, Toronto, Canada. Association for Computational Linguistics.

John Austin. 1962. *How to do things with words*. Clarendon Press.

G.P. Baker and P.M.S. Hacker. 1984. The illusions of rule-scepticism. In *Scepticism, Rules and Language*, pages 56–97. Basil Blackwell.

Michael Beaney. 2021. Analysis. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, summer 2021 edition. Metaphysics Research Lab, Stanford University.

Emily Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. *Proceedings ofthe 58th Annual Meeting ofthe Association for Computational Linguistics*, pages 5185–5198.

Maxwell Bennett and Peter Michael Stephan Hacker. 2003. *Philosophical foundations of neuroscience*, volume 79. Blackwell Oxford.

Maxwell Bennett and Peter Michael Stephan Hacker. 2007. The conceptual presuppositions of cognitive neuroscience. In *Neuroscience and philosophy: Brain, mind, and language*, pages 127–162. New York: Columbia University Press.

Jean-Philippe Bernardy and Stergios Chatzikyriakidis. 2019. What kind of natural language inference are nlp systems learning: Is this enough? In *ICAART (2)*, pages 919–931.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience Grounds Language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Robert Brandom. 1994a. *Making It Explicit: Reasoning, Representing, and Discursive Commitment*. Harvard university press.

Robert Brandom. 1994b. *Making it explicit: Reasoning, representing, and discursive commitment*. Harvard university press.

Robert Brandom. 2010. *Between Saying and Doing: Towards an Analytic Pragmatism*. Oxford University Press.

Robert Brandom. 2021. *Articulating Reasons*. Harvard University Press.

Ronan Le Bras, Swabha Swayamdipta, Chandra Bhagavatula, Rowan Zellers, Matthew E Peters, Ashish Sabharwal, and Yejin Choi. 2020. Adversarial filters of dataset biases. *arXiv preprint arXiv:2002.04108*.

James Robert Brown and Yiftach Fehige. 2023. Thought experiments. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, winter 2023 edition. Metaphysics Research Lab, Stanford University.

Tyler Burge. 2010. *Origins of Objectivity*. Oxford: Oxford University Press.

Tyler Burge. 2012. Living wages of sinn. *The Journal of Philosophy*, 109:40–84.

Rudolf Carnap. 1950. Empiricism, semantics, and ontology. 4(11):20–40.

Khyathi Raghavi Chandu, Yonatan Bisk, and Alan W. Black. 2021. Grounding 'Grounding' in NLP.

Zeming Chen, Qiyue Gao, and Lawrence S Moss. 2021. Neurallog: Natural language inference with joint neural and logical reasoning. *arXiv preprint arXiv:2105.14167*.

Jacob Devlin, 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Jerry Fodor. 1975. *The Language of Thought*. Harvard university press.

Jerry Fodor. 1983. *Representations: Philosophical Essays on the Foundations of Cognitive Science*. Cambridge/MA: Mit Press.

Jerry Fodor. 1987. *Psychosemantics: The Problem of Meaning in the Philosophy of Mind*. Cambridge, Massachusetts/London, England: MIT Press.

Jerry Fodor. 1992. *A Theory of Content and Other Essays*. Cambridge/MA: MIT press.

Jerry Fodor. 2008. *LOT 2: The Language of Thought Revisited*. Cambridge/MA: Oxford University Press.

María José Frápolli. 2019. Propositions first: Biting geach's bullet. *Royal Institute of Philosophy Supplements*, 86:87–110.

Gottlob Frege. 1892. Über sinn und bedeutung. *Zeitschrift für Philosophie und philosophische Kritik*, 100:25–50.

Hans-Johann Glock. 2012. Thought, judgment and perception. 86:207–221.

Hans-Johann Glock. 2013. Animal minds: Philosophical and scientific aspects. In *A Wittgensteinian Perspective on the Use of Conceptual Analysis in Psychology*, pages 130–152. Palgrave MacMillan.

Paul Gochet. 2011. Model-Theoretic Semantics. In Marina Sbisà, Jan-Ola Östman, and Jef Verschueren, editors, *Philosophical Perspectives for Pragmatics*, number 10 in Handbook of Pragmatics Highlights, pages 171–179. J. Benjamins, Amsterdam.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. 2016. *Deep Learning*. Cambridge, Massachusetts/London, England: MIT Press.

Reto Gubelmann. 2019. *A Science-Based Critique of Epistemological Naturalism*. Palgrave MacMillan.

Reto Gubelmann. 2023. A loosely wittgensteinian conception of the linguistic understanding of large language models like bert, gpt-3, and chatgpt. *Grazer Philosophische Studien*, 99(4):485–523.

Reto Gubelmann, Aikaterini-lida Kalouli, Christina Niklaus, and Siegfried Handschuh. 2023a. When truth matters - addressing pragmatic categories in natural language inference (NLI) by large language models (LLMs). In *Proceedings of the 12th Joint Conference on Lexical and Computational Semantics (*SEM 2023)*, pages 24–39, Toronto, Canada. Association for Computational Linguistics.

Reto Gubelmann, Ioannis Katis, Christina Niklaus, and Siegfried Handschuh. 2023b. Capturing the Varieties of Natural Language Inference: A Systematic Survey of Existing Datasets and Two Novel Benchmarks. *Journal of Logic, Language and Information*.

Reto Gubelmann, Christina Niklaus, and Siegfried Handschuh. 2022. A Philosophically-Informed Contribution to the Generalization Problem of Neural Natural Language Inference: Shallow Heuristics, Bias, and the Varieties of Inference. In *Proceedings of the 3rd Natural Logic Meets Machine Learning Workshop (NALOMA III)*, pages 38–50, Galway, Ireland. Association for Computational Linguistics.

Peter MS Hacker. 2009. Philosophy: A contribution, not to human knowledge, but to human understanding. *Royal Institute of Philosophy Supplements*, 65:129–153.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42:335–346.

He He, Sheng Zha, and Haohan Wang. 2019. Unlearn dataset bias in natural language inference by fitting the residual. *arXiv preprint arXiv:1908.10763*.

Ulf Hlobil and Robert B Brandom. 2024. *Reasons for Logic, Logic for Reasons: Pragmatics, Semantics, and Conceptual Roles*. Taylor & Francis.

Aikaterini-Lida Kalouli, Richard Crouch, and Valeria de Paiva. 2020. Hy-NLI: a hybrid system for natural language inference. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 5235–5249, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Immanuel Kant. 1998 [1781/1787]. *Kritik der reinen Vernunft*. Hamburg: Meiner.

Saul Kripke. 1980. Naming and necessity, rev. ed. *Cambridge, MA: Harvard UP*.

Saul Kripke. 1983. Semantical Considerations on Modal Logic. *Acta Philosophica Fennica*, 16:83–94.

Saul A. Kripke. 1982. *Wittgenstein on Rules and Private Language*. Basil Blackwell.

Catherine Legg and Christopher Hookway. 2024. Pragmatism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, winter 2024 edition. Metaphysics Research Lab, Stanford University.

Zhaohui Luo. 2014. Formal semantics in modern type theories: Is it model-theoretic, proof-theoretic, or both? In *International Conference on Logical Aspects of Computational Linguistics*, pages 177–188. Springer.

Matt MacMahon. Walk the Talk: Connecting Language, Knowledge, and Action in Route Instructions.

Penelope Maddy. 2007. *Second Philosophy. A Naturalistic Method*. Oxford: Oxford University Press.

Penelope Maddy. 2011. Naturalism, transcendentalism and therapy. In *Transcendental Philosophy and Naturalism*, pages 120–156. Oxford: Oxford University Press.

Penelope Maddy. 2014. *The Logical Must*. Oxford: Oxford University Press.

Penelope Maddy. 2017. *What Do Philosophers Do? Skepticism and the Practice of Philosophy*. The Romanell Lectures. Oxford: Oxford University Press.

Rabeeh Karimi Mahabadi, Yonatan Belinkov, and James Henderson. 2019. End-to-end bias mitigation by modelling biases in corpora. *arXiv preprint arXiv:1909.06321*.

Matthew Mandelkern and Tal Linzen. 2024. Do Language Models' Words Refer?

Tom McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Julian Michael, Ari Holtzman, Alicia Parrish, Aaron Mueller, Alex Wang, Angelica Chen, Divyam Madaan, Nikita Nangia, Richard Yuanzhe Pang, Jason Phang, and Samuel R. Bowman. 2023. What Do NLP Researchers Believe? Results of the NLP Community Metasurvey. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16334–16368, Toronto, Canada. Association for Computational Linguistics.

Raphaël Millière and Cameron Buckner. 2024. A Philosophical Introduction to Language Models – Part I: Continuity With Classic Debates.

Dimitri Coelho Mollo and Raphaël Millière. 2023. The Vector Grounding Problem.

Richard Montague et al. 1970. *English as a Formal Language*. Ed. di Comunità.

Thomas N Munson. 1962. Wittgenstein's phenomenology. *Philosophy and Phenomenological Research*, 23(1):37–50.

Bhavish Pahwa and Bhavika Pahwa. 2023. BpHigh at SemEval-2023 Task 7: Can Fine-tuned Cross-encoders Outperform GPT-3.5 in NLI Tasks on Clinical Trial Data? In *Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023)*, pages 1936–1944, Toronto, Canada. Association for Computational Linguistics.

Ellie Pavlick. 2023. Symbols and grounding in large language models. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, 381(2251):20220041.

Jaroslav Peregrin. 1997. Language and its models: Is model theory a theory of semantics? *Nordic Journal of Philosophical Logic*, 2(1):1–23.

Steven T Piantadosi and Felix Hill. 2022. Meaning without reference in large language models. *arXiv preprint arXiv:2208.02957*.

Alvin Plantinga. 1974. *The Nature of Necessity*. Oxford: Oxford University Press.

John Preston and John Mark Bishop, editors. 2002. *Views into the Chinese room: New essays on Searle and artificial intelligence*. Oxford: Oxford University Press.

Hilary Putnam. 1975. What is mathematical truth? In *Philosophical Papers, Volume 1*, pages 60–78. Cambridge: Cambridge University Press.

Jake Quilty-Dunn, Nicolas Porot, and Eric Mandelbaum. 2023. The best game in town: The reemergence of the language-of-thought hypothesis across the cognitive sciences. *Behavioral and Brain Sciences*, 46:e261.

Willard Van Orman Quine. 1974. *The Roots of Reference*. Open Court Publishing Co.

Willard Van Orman Quine. 1976. Whither physical objects? In *Boston Studies in the Philosophy of Science*, volume XXXIX, pages 497–504. Reidel.

Willard Van Orman Quine. 1980 [1951]. Two dogmas of empiricism. In *From a Logical Point of View*, pages 20–46. Harvard University Press.

Willard Van Orman Quine. 1981. Things and their place in theories. In *Theories and Things*, pages 1–23. Harvard University Press.

Jack W. Rae, Sebastian Borgeaud, and Trevor Cai et al. 2021. Scaling Language Models: Methods, Analysis & Insights from Training Gopher.

Michael Rescorla. 2020. The computational theory of mind. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, fall 2020 edition. Metaphysics Research Lab, Stanford University.

Michael Rescorla. 2023. The language of thought hypothesis. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, winter 2023 edition. Metaphysics Research Lab, Stanford University.

Deb Roy and Ehud Reiter. 2005. Connecting language to the world. *Artificial Intelligence*, 167(1-2):1–12.

Magnus Sahlgren and Fredrik Carlsson. 2021. The Singleton Fallacy: Why Current Critiques of Language Models Miss the Point. *Frontiers in Artificial Intelligence*, 4:682578.

John Searle. 1969. *Speech Acts: An Essay in the Philosophy of Language*. Cambridge: Cambridge University Press.

John Searle. 1980. Minds, brains, and programs. *Behavioral and Brain Sciences*, 3(3):417–424.

Mariarosaria Taddeo and Luciano Floridi. 2005. Solving the symbol grounding problem: A critical review of fifteen years of research. *Journal of Experimental & Theoretical Artificial Intelligence*, 17(4):419–445.

Sever Topan, David Rolnick, and Xujie Si. 2021. Techniques for symbol grounding with satnet. *Advances in Neural Information Processing Systems*, 34:20733–20744.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Prasetya Ajie Utama, Nafise Sadat Moosavi, and Iryna Gurevych. 2020. Towards debiasing nlu models from unknown biases. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7597–7610.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. Superglue: A stickier benchmark for general-purpose language understanding systems. *Advances in Neural Information Processing Systems*, 32.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Ludwig Wittgenstein. 1958. *Philosophical Investigations*. Blackwell.

Xiang Zhou and Mohit Bansal. 2020. Towards robustifying nli models against lexical dataset biases. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8759–8771.

A A Small Selection of Engineering Work Done on Symbol Grounding in NLP

Compare how a recent research report describes its understanding of the concept of symbol grounding: “[...] the *Symbol Grounding Problem*: the inability to map visual inputs to symbolic variables without explicit supervision (“label leakage”)” (Topan et al., 2021). On this understanding, the symbol grounding problem refers to a very specific engineering challenge, a kind of label leakage that occurs when a method is unable to connect information in the visual modality with symbolic-logical structures. Their solution to this problem is equally specific and sober, involving, on Bender and Koller’s view, nothing more but more sophisticated processing of form.

Similarly, Roy and Reiter (2005, 2) describe a notion of language grounding that generalizes across modalities, rather than, as previously, focusing on the visual one: “Language grounding provides an impetus for AI researchers to integrate these subfields, so that they can attempt to build machines that can converse about what they observe and do in human-like ways”. In the same vein, MacMahon develop a system that is able to integrate linguistic, spatial, and local dimensions in its behavior. By means of a typology in this sense of symbol grounding, Chandu et al. (2021) provide a very sophisticated typology of different kinds of grounding relations within the (purely formal, according to Bender and Koller) modalities available to AI systems and make specific suggestions for future engineering.

B A Slightly more Detailed Sketch of CTM

The computational theory of the mind (CTM, for an introduction, see Rescorla 2020) is the overarching hypothesis that (1) the mind of humans (and likely of other animals) is essentially a computer, and that (2) a computer can be described by the processes that run on it, while, as it were, the biological wetware such as the human body, is inessential: two minds are identical if they have the same computational-functional structure, two mental tokens (symbols) are identical if they perform the same function: you can transfer the mind (conceived as software) on any body (hardware) that fulfills the basic requirements. The human mind then (roughly) becomes a symbol manipulator, like a certain compiler that can take in a well-formed sequence of symbols, a script, and perform a sequence of processing steps prescribed in this script.

Once you’ve got this conception of the mind as symbol-manipulating software in place, it’s hard to avoid asking what kind of symbols are being manipulated by this mind. Enter the idea of a language of thought (so-called Mentalese, see Rescorla 2023 for an introduction), the symbol system that is, as it were, “in the head” without any necessary connections or mappings to natural languages – pure mental symbolism.

C Details on the Distinction Between Conceptual and Empirical Questions

Bender and Koller frame their conceptual reflections and arguments on the meaning of meaning by the use of an intriguing metaphor that also figures in the title of their paper: they suggest that having an accurate conception of meaning is analogous to having a realistic top-down overview on the landscape that allows knowing what hill should be climbed. Empirical research and engineering is then conceived as the bottom-up effort to climb that hill that was previously identified by conceptual top-down considerations. While they suggest that current LLMs might be climbing the wrong hill, their basic goal seems to be more in kindling a discussion on this conceptual level than being absolutely right with their own position: to bring, in their own words, "the top-down perspective into clearer focus" (Bender and Koller, 2020, 5192).[10]

Following common usage, we propose to conceive of what Bender and Koller call the top-down perspective as conceptual and hence philosophical questions, to be resolved by analysis and reflection about concepts, for instance, by exploring the meaning of meaning, language, or grounding. In contrast, the bottom-up view is the approach taken by empirical methods, typically employed in NLP.

In philosophy, the traditional way to conceive what Bender and Koller call the top-down-perspective are conceptual questions (for an excellent overview, see Beaney 2021). They are about what claims mean and whether or not they make sense. Empirical questions are about the truth of claims. So, the fact that the claim "my favorite green idea sleeps peacefully" makes no sense is a conceptual fact, having to do with what can and cannot sensibly be said of ideas. In contrast, the fact that the claim "the seasonal flu is caused by viruses" is true is an empirical-scientific fact, to be established by these same methods. Bender and Koller's claim that LLMs might be climbing the wrong hill is predominantly of the conceptual sort: according to them, it is likely that meaning just does not mean a thing that can be understood by processing mere form.

In philosophy, there is an ongoing discussion about the standing of this traditional division of labor. In this discussion, there are three established groups of positions. One of them, naturalism, claims that there simply is no significant role played by philosophy at all. This position has been championed by Willard Van Orman Quine (1980 [1951], 1976, 1981). Currently, one of its most distinguished proponents is Penelope Maddy (2007, 2011, 2014, 2017). For a recent book-length critique of naturalism, see Gubelmann (2019).

The position at the other end of the spectrum, which we call apriorism, maintains that empirical research has nothing whatsoever to contribute to conceptual questions, which includes the clarifications regarding the precise conception of human-level linguistic abilities and of strong AI. Versions of this position date back to the logical positivists, and to Carnap in particular (see Carnap 1950). Currently, one of the most established and controversial proponent of this view is P.M.S. Hacker. His seminal critique of neuroscience (Bennett and Hacker, 2003) and the insightful debate with his critics, including John Searle and Daniel Dennet, in (Bennett and Hacker, 2007)), is based on a strict distinction between two dimensions of investigation, namely the philosophical one separating sense from nonsense and the empirical one separating truth from falsehood (Bennett and Hacker, 2007, 12). On this view, empirical inquiry has nothing whatsoever to contribute to clarifying non-technical, established concepts such as language (while empirical inquiry can influence the formation of new, technical concepts, terms of art).

The third group of position occupies a middle-ground between naturalism and apriorism. we maintain that such middle-ground-positions are most promising for the questions at hand, namely the cooperation of philosophy and NLP-engineering with the aim of understanding meaning and maximizing progress in NLP. We submit that Bender and Koller implicitly also subscribe to such a middle-group position, as they discuss recent evidence from NLI research to support their claim that LLMs currently climb the wrong hill. To the a priorist, this would make no sense, as conceptual matters are entirely immune against factual considerations. Within this third group, Glock's conception of impure conceptual analysis is particularly interesting for (see in particular Glock 2012, 115-119 and Glock 2013, 140-145). In a first approximation, empirical research can be said to contribute the facts to this common effort, while philosophy contributes conceptual analyses, clarifications and reflection (compare Glock 2013, 136-151). However, upon closer inspection, the situation is more subtle than that. In particular, it seems plausible that empirical results can strongly suggest substantial conceptual modifications. This agrees with a position, currently championed by

[10] The authors are not fully consistent on this, see Bender and Koller (2020, 5188). However, we follow the principle of hermeneutical charity and assume that their main goal is indeed to bring this top-down perspective into focus and less to firmly establish that meaning cannot be learned from form.

Tyler Burge, according to which empirical research can show that philosophical analyses of concepts are wrong, or that they are incomplete (compare Burge 2010, xv, Burge 2012, 79, Putnam 1975, ch. 12 and ultimately Kripke 1980).

Unlike empirical claims, conceptual claims cannot be straightforwardly falsified – after all, they define what empirical falsification amounts to. Still, conceptual claims can turn out to be inaccurate descriptions of the phenomenon at hand, to miss it, so to speak.

D Brandom's Inferentialist Pragmatism

D.1 A Slightly More Detailed Sketch of the Position

The version of pragmatism that we would like to delineate to allow for a more specific conception of this kind of view on language is the inferentialist pragmatism of Robert Brandom (for his main works, see Brandom 1994a, 2010, 2021; Hlobil and Brandom 2024). Brandom follows the general pragmatic outlook to put linguistic practice first. This means that his thinking about language sets in with what human beings do with language. According to him, what they do is inherently normative:[11] They commit themselves to certain things, and they become entitled to certain other things. Furthermore, Brandom specifies the kind of practice that is fundamental on his view: drawing logical inferences between claims. According to him, it is characteristic of speech acts that they entail commitments and incur entitlements that are inferentially connected. For instance, if I am claiming "The earth is flat", this speech act commits me to justify my (rather outrageous) claim when properly challenged. This, in turn, could happen simply by remarking "What of the pictures of earth from space that seem to show a ball?" If my claim goes through, however, this would entitle me to an entire class of other claims, such as the one that it is impossible to fly around the world.

According to this picture, claims are the central units of linguistic meaning: linguistic meaning is constituted by the network of entailments that exist between claims, and which are evidenced in the norms that we hold our fellow language users accountable to. The meaning of words is defined by the difference that they make for the entailment relations of claims in which they occur. For instance, if you commit to the truth of "This ball is red", your conversational counterparts will hold you accountable for the truth of this claim as well as for any claim that is logically entailed by it, such as "This ball is not white all over". This would not follow if you had said "This ball is old", evincing a semantic difference between "red" and "old". The meaning of concepts, then, is to be derived from the propositional level.

On this inferentialist picture, the symbol grounding problem simply loses its bite. While inferentialists need to account for the relevance and significance of observation statements (see Brandom 1994b, 213ff.), concepts receive their meaning from norm-governed use in inferentially connected claims.

Taking up the metaphor at the end of section C, might this be an accurate picture of natural language? How could we begin to answer this question? In our case, for instance, if it would turn out that LLMs reach superhuman performance at sensible and realistic NLI benchmarks, while utterly failing most other tasks in natural language understanding, for instance, text summarization, information retrieval, word-sense-disambiguation, this would seriously question inferentialism as such. Conversely, if it would turn out that progress on NLI is still lacking, this might indicate that we are indeed, using Bender and Koller's metaphor, climbing the wrong hill. Again mostly for illustrative purposes, we detail some of the evidence in this regard in the next section.

D.2 Comparing the Theory to the Picture Emerging From Empirical Research

Briefly, the following survey of the state of the art in NLI is intended to assess whether any of the two scenarios referred at the end of the previous section – NLI as an island of performance in natural language understanding and, conversely, stagnation in NLI – are real. For a recent survey on NLI, see Gubelmann et al. (2023b).

Before large generative transformer-based LLMs such as gpt-3.5 or llama-2 (Touvron et al., 2023) became the center of attention in NLP, Natural Language Inference was largely approached using encoder-only transformers in the tradition inspired by BERT (Devlin et al., 2019). With these models,

[11] It is essential to a norm that it *prescribes* rather than *describes*. This means that, to follow a norm (rather than to just act in accordance with it, see the literature following Wittgenstein 1958, §§185,189,198,201, in particular Kripke 1982, Baker and Hacker 1984, and Hacker 2009), one has to be able to decide not to follow it. In the case of factual speech, this often amounts to lying: To deliberately make a claim that one knows is not entailed from all that one knows otherwise.

NLI hit upon the so-called problem of generalization: the observation that these models performed very well on the training data, but often less than random on out of distribution data. The problem was referenced by numerous researchers in the field Bernardy and Chatzikyriakidis (2019); Mahabadi et al. (2019); He et al. (2019); Bras et al. (2020); Utama et al. (2020); Zhou and Bansal (2020); Asael et al. (2021); Gubelmann et al. (2022).

Often, it was suggested that this problem of generalization is a consequence of the models' overfitting on the training dataset (compare Goodfellow et al. 2016). As a consequence, the models represented spurious idiosyncrasies as well as shallow heuristics. In a much-noticed publication, McCoy et al. (2019) report that these models use three different kinds of heuristics to achieve their performance.

In the period between GPT-3 and gpt-3.5, the study by Rae et al. (2021) suggests that, in the words of the authors, "the benefits of scale are nonuniform". On the one hand, they find less-than expected improvement in performance with logical or mathematical reasoning when scaling to Gopher, a model having 280B parameters, while it sets a new SOTA with many other natural language understanding tasks such as RACE-h and RACE-m, clearly outperforming GPT-3.

This emerging consensus has created a space for hybrid systems that combine LLMs with rule-based modules to overcome the perceived stagnation of purely neural-based LLMs, often building on Angeli and Manning (2014), who combined natural logic, monotonicity structures, WordNet and learned word probabilities as well as embeddings to conceive of NLI as a search problem. Among these recent approaches are Kalouli et al. (2020), used, e.g., in Gubelmann et al. (2023a), establishing a new SOTA on many standard datasets. Chen et al. (2021) conceive of NLI as a path planning problem with the premise as the start and the hypothesis as the goal to be reached.

It would be highly desirable to have an empirically solid, reproducible track record of state of the art models whose inner workings are openly available, allowing to conduct research on them in a reproducible and empirically sound manner. Unfortunately, this is not the case for the most intensely debated – and perhaps also best performing – models currently around, namely the GPT-4-series released by OpenAI. With regard to these models, two central parameters are entirely unknown: (1) the training data of the models. This effectively means that we cannot distinguish between in-distribution and out-of-distribution samples, which in turn means that we cannot judge the ability of these models to generalize as opposed to their ability to memorize. (2) the inner workings of the entire system. It is not clear whether the results returned on requests sent by the API are solely processed by the respective language models or whether further elements, perhaps even rule-based modules, come into play.

Likely, research on the currently exploding scene of open LLMs, pioneered by Meta's llama-2 (Touvron et al., 2023), will catch up; right now, however, there are no systematic, published studies of the logical abilities of such open LLMs. With regard to OpenAI's models Pahwa and Pahwa (2023) find that using a fine-tuned BERT-large is able to almost match the NLI performance of GPT-3.5 in a BioNLP setting (macro F1: 0.694 by GPT-3.5 vs. 0.690 by BERT-large).

In sum, the scene is slightly difficult to survey at present. What is clear is that the scenario that would immediately expose inferentialism as inaccurate has not become reality: performance at NLI is not incommensurably better than performance at other benchmarks.