Minds and Machines (2023) 33:33–54
https://doi.org/10.1007/s11023-023-09622-4

GENERAL ARTICLE

Grounding the Vector Space of an Octopus: Word Meaning from Raw Text

Anders Søgaard[1]

Received: 15 September 2022 / Accepted: 5 January 2023 / Published online: 23 January 2023

Abstract

Most, if not all, philosophers agree that computers cannot learn what words refers to from raw text alone. While many attacked Searle's Chinese Room thought experiment, no one seemed to question this most basic assumption. For how can computers learn something that is not in the data? Emily Bender and Alexander Koller (2020) recently presented a related thought experiment—the so-called Octopus thought experiment, which replaces the rule-based interlocutor of Searle's thought experiment with a neural language model. The Octopus thought experiment was awarded a best paper prize and was widely debated in the AI community. Again, however, even its fiercest opponents accepted the premise that what a word refers to cannot be induced in the absence of direct supervision. I will argue that what a word refers to *is* probably learnable from raw text alone. Here's why: higher-order concept co-occurrence statistics are stable across languages and across modalities, because language use (universally) reflects the world we live in (which is relatively stable). Such statistics are sufficient to establish what words refer to. My conjecture is supported by a literature survey, a thought experiment, and an actual experiment.

Keywords Language models · Grounding · Chinese room

1 Introduction

I begin with Searle, the Churchlands, Turing, and a recent paper from Emily Bender and Alexander Koller.

In Searle's Chinese Room thought experiment (Searle, 1980), the philosopher imagines himself or some other speaker of English alone in a room following a computer program or some form of manual for responding to Chinese characters slipped under the door. Simply following the program or manual, Searle sends appropriate

✉ Anders Søgaard
soegaard@di.ku.dk

[1] Department of Computer Science, University of Copenhagen, Lyngbyvej 2, 2100 Copenhagen, Denmark

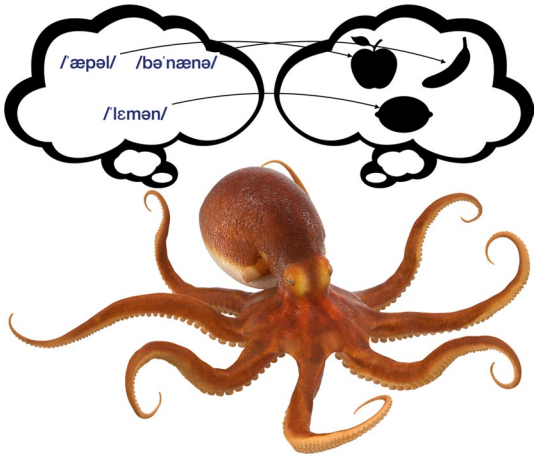


Fig. 1 Grounding the Vector Space of an Octopus

strings of Chinese Mandarin characters back out under the door, and this leads those outside to mistakenly suppose there is a Chinese speaker in the room. In leading observers to think so, Searle—or the computer program, really—passes the Turing Test. Yet, Searle understands nothing of Chinese. Or so he thinks.

The Churchlands (Churchland & Churchland, 1990) break Searle's argument into three premises and a conclusion. Premise 1 is that computer programs amount to formal application of syntactic rules. Premise 2 and 3 say that formal application of syntactic rules is insufficient (Premise 3) for the kind of semantics language gives rise to in humans (Premise 2). The Churchlands claim Premise 3 is, as of now, an empirical question:

> [Consider] the crucial third [premise] of Searle's argument: "Syntax by itself is neither constitutive of nor sufficient for semantics." Perhaps this [premise] is true, but Searle cannot rightly pretend to know that it is. (Churchland & Churchland 1990, p. 34)

Premise 3 is why Searle can say that even if the formal application of syntactic rules is sufficient to communicate effectively in a human-like fashion, it is still insufficient for understanding. Turing, in contrast, was agnostic about the validity of this inference. If something communicates fluently we have no reason to doubt that it understands.[1] Turing saw no *prima facie* reason to rule out that computers would one day understand language.

Searle, the Chuchlands, and Turing provide our backdrop. I will show how recent work in artificial intelligence is starting to fill the empirical gap identified by the Churchlands, providing evidence that at least some semantics may fall out of syntax, vindicating Turing's agnostic stance (Fig. 1).

In one of the most cited artificial intelligence papers of 2020, which also won the ACL 2020 Best Theme Paper Award, Emily Bender and Alexander Koller (2020)

[1] See Dietrich et al. (2021) for an interesting comparison of the Turing Test and the so-called Duck test.

presented a modified version of Searle's Chinese Room argument, namely the so-called Octopus thought experiment. They replace the rule-based interlocutor in Searle's Chinese Room with an interlocutor that is *very good at detecting statistical patterns*, i.e., a better-than-GPT-3, hypothetical language model.[2] They use the argument to arrive at the same conclusion as Searle did: Computers cannot acquire meaning from raw text alone.[3]

Importantly, Bender and Koller claim that computers cannot even acquire the meaning *of words* from raw text. This is clear when they talk about how the octopus 'does not know what words such as *rope* and *coconut* refer to.' They also write

> The lexical similarity relations learned by distributional models trained on text don't in themselves connect any of those words to the world. (Bender & Koller 2020, p. 5191)

I will distinguish between three questions:

1. Can language models learn the meaning of linguistic utterances?
2. Can language models learn the meaning of words?
3. Can language models learn the meaning of specific classes of words, e.g., color terms?

This three-way distinction will be useful below. Learning the meaning of full utterances requires a number of things, e.g., inducing the compositional machinery of language. I will therefore focus on questions 2 and 3, i.e., to what extent language models can be said to be learning the meaning of words? I will now and then zoom in on specific classes of words, e.g., color terms, when it is useful for the overall discussion.

I present the Chinese Room and the Octopus thought experiments in more detail in Sect. 2. In Sect. 3, I discuss the assumptions made in the latter, e.g., the role played by awareness and social interaction in language acquisition. I will argue that understanding the meaning of words requires establishing their reference or grounding them in real-world representations, but is orthogonal to awareness and social interaction. I turn to this problem in Sect. 5, after a short detour in Sect. 4, in which I discuss whether language models can possibly solve the Turing Test. Bender and Koller think not. I think their arguments are unconvincing and, at best, inconclusive. In Sect. 5, I present positive arguments for how semantics may fall out of the kind of syntax language models induce from raw text.

[2] GPT-3 is one of most famous autoregressive English language models released. It was released by OpenAI in June 2020, after 355 compute years of training (on a V100). It consists of 175 billion parameters and is particularly good at generating coherent and creative text.

[3] Note how this is orthogonal to Gold (1967), arguing that grammar cannot be acquired from raw text: Even if language was subregular and learnable in the limit, meaning would not necessarily be learnable. Similarly, Gold does not prove that the mapping of strings generated by context-sensitive grammars onto, say, communicative intent, cannot be learned from raw text.

1.1 Background

Some distinctions in linguistic semantics will be useful for what follows: It is common to refer to the inferential and referential aspects of meaning (Marconi, 1997). Most would agree that language models can learn inferential semantics (Sahlgren & Carlsson, 2021; Piantadosi & Hill, 2022). (Marconi, 1997) explicitly discusses how Searle's Chinese Room only applies to referential semantics; see also (Hirst, 1997). Several authors (Sahlgren & Carlsson, 2021; Piantadosi & Hill, 2022) have pointed out how in the same way Bender and Koller fail to make this distinction, and that the Octopus thought experiment, if anything, only applies to referential semantics. Another important distinction is between internalist and externalist semantics. Internalists, e.g., Schank–Rapaport-style conceptual semantics (Schank & Colby, 1973), would generally have it that reference is fixed through mental representations, whereas externalists would say it is fixed by its causal implications across social interactions. Bender and Koller define meaning as a relation between linguistic form and communicative intent grounded in the outside world; but they also evoke the concept of *standing* meaning to describe what a word's or sentence's meaning has in common across social interactions. In this way, they remain agnostic about internalism-externalism. Even the late Wittgenstein allows for standing (prototypical) meaning, but emphasizes how probing standing meaning is a non-privileged language game.[4] None of what I will have to say relies on taking a stand in this debate. My discussion will be limited to standing meaning of words. That is: I will be agnostic about how exactly, in practice (or 'in context'), reference is fixed, but show how standing meaning may fall out of syntax.

1.2 Contributions

Below I will discuss the Octopus thought experiment and its apparent weaknesses, including some that also pertain to Searle's Chinese Room. I distill four claims from the work of Bender and Koller, three of which are shared by Searle: Learning the meaning of words requires awareness, grounding, social interaction, as well as the ability to solve the Turing Test. I shall argue that only grounding is needed, and that computers can already (to a high degree) do this in the absence of explicit instructions.

[4] In *Philosophical Investigations* (6), Wittgenstein writes, for example, that '[ostensive teaching of words] is an important part of the training, because it is so with human beings; not because it could not be imagined otherwise.'

2 The Octopus in the Chinese Room

2.1 Searle's Chinese Room

The Chinese Room argument of Searle (1980) goes as follows: Imagine a native speaker of English, who understands *no* Chinese. Imagine also that this person is locked in a room with boxes of Chinese characters together with a book of instructions in English for manipulating the symbols.[5] A person outside the room sends in small batches of sequences of Chinese characters (questions). To the English speaker, the characters are meaningless. She now follows the book of instructions in order to manipulate the characters. Imagine the instructions are so good and have sufficient coverage, and that the English speaker gets so good at manipulating the Chinese characters, that she is able to give correct answers to the questions (the output). The instructions make it possible for the English speaker to pass the Turing Test for understanding Chinese, but (Searle, 1980) argues that all the same, the native speaker of English does 'not understand a single word of Chinese'. Searle's *reductio ad absurdum* argument is designed to show that the existence of a program p enabling fluent communication in l is not sufficient to say that the system running p *understands* l. This directly challenges Turing's functionalist view (Copeland, 2004) and the motivation behind his design of the Turing Test.[6]

2.2 The Unnoticed Loophole

I first want to point out a weakness in Searle's Chinese Room thought experiment that so far has gone unnoticed, despite the many criticisms (Bishop, 2002): In the thought experiment, there is a (small, but non-negligible) overlap between the real world experience of the English speaker in the room and the person outside of the room. The person in the room may have a sense of time, for example. Nothing in the way Searle describes the Chinese Room prevents the person in the room for taking time into account. In the limit, this may enable her to induce that 上午 means *morning* from the fact that the phrase occurs almost exclusively in input received in the morning hours. Or that 你好 means *hi*, because it only occurs after periods of extended silence. The same counter-argument applies to the octopus in the argument put forward in Bender and Koller, which may induce the meaning of *storm* or *winter* from their distribution over time. The overlap provides weak supervision for the grounding of the representations of the instructions of the English speaker, i.e.,

[5] The boxes and the instructions are Searle's analogies to the lexica and rewrite rules of the NLP models of the 1980s. He refers to these simply as *programs*.

[6] Many philosophers have agreed with Searle that the Turing Test is insufficient to test for language understanding (Warwick & Shah, 2015) It is also unclear whether this is how the test was really intended (Dietrich et al., 2021). I will not argue for the sufficiency of the Turing Test, but against Bender and Koller's claim that language models are bound to fail the test sooner or later (Sect. 4). My main contribution, however, will be to present arguments against the main thesis in their paper, namely that word meaning is *a priori* not learnable from raw text. See (Shieber, 2004) for more discussion of the Turing Test.

for the representations of the program. While temporal words or words referring to weather phenomena make up tiny fractions of natural language vocabularies, numerals (Artetxe et al., 2017) or overlapping words (Søgaard et al., 2018) is known to be sufficient to align language models in different languages.[7] As one of my anonymous reviewers reminded me, daily weather reports was also how Alan Turing and his colleagues at Bletchley Park finally cracked the German Enigma. While this loophole is clearly an unintended weakness in the two thought experiments, the argument I will put forward, is much stronger: Word representations induced from raw text alone can be grounded even in the absence of supervision.

2.3 Observer-Relativity

Searle thinks of the Chinese Room argument as a refutation of the Turing Test. He also argues that it shows how semantics is underivable from syntax or computation. Computers perform syntactic symbol manipulation, but this will never be sufficient for understanding, he argues. This, he continues, is because symbol manipulation is ontologically *observer-relative* (Searle, 1992); see (Endicott, 1996) for discussion. To Searle, symbol manipulation only acquires meaning in the eyes of the beholder. Only when interpreted is it *assigned* meaning. Searle observes that structures may happen to be isomorphic to the symbol manipulation performed by a computer program *without* having meaning. This could, for example, hold true for a subset of an enormous cluster of asteriods or a subset of the molecules in the office wall behind you right now. Haugeland (2003) has argued that the analogy is false, because it ignores the fact that the symbol manipulation of a computer program is *reproducible and systematically retrievable*. Dennett (1987), more generally, argues that Searle mistakes semantics for consciousness of semantics. What Searle shows is not that semantics is underivable from syntax, but that *consciousness of* semantics is underivable from syntax. It is not the symbol manipulations that are observer-relative, but the consciousness of their meaning that is observer-relative,[8] Conflating semantics and consciousness of semantics—or understanding and awareness of

[7] Naim et al. (2014) showed that time stamps in both modalities is enough to align instructions and video sequences. This obviously does not show unsupervised grounding is possible, but it tells us weak supervision is sufficient, even in the absence of joint supervision and feedback signals. Sentence-image alignment is possible if coarse-grained, document-level supervision is available (Li et al., 2021)

[8] Others (Ivan & Indurkhya, 2019) as well as Searle himself, have suggested that computers in the same sense do not understand Chess or Go. In our view, thinking of games makes it easier to see how Searle conflates understanding and consciousness of understanding. Clearly, computers understand Chess in the sense that they understand (have knowledge of) how pawns move, even in the absence of any second order processes.

understanding—is simply a category mistake (Ryle, 1938).[9] I return to the possibility of language understanding in the absence of consciousness or awareness in Sect. 3.

2.4 The Octopus Thought Experiment

Bender and Koller present a new version of the Chinese Room argument, replacing the rule-based interlocutor with a statistical one.[10]

Bender and Koller present us with the following scenario: Two humans A and B, each stuck on a deserted island, communicate regularly through an underwater cable. A statistical learner in the form of a hyper-intelligent octopus O taps in on the cable communication, while unable to visit or observe the islands. After a period of learning to predict what A and B will say, the octopus cuts the cable and inserts himself, pretending to be B. Like (Searle, 1980), Bender and Koller argue that O, regardless of how good it becomes at imitating B, does not and will never truly understand the language of A. The octopus, a stand-in for a pretrained language model, will never learn to represent meaning, i.e., a mapping of expressions onto communicative intent. They further argue that O eventually fails the Turing Test, and that neither O in this example, nor language models in general, can be said to understand language. This is because language understanding, in their view, requires awareness, the ability to ground words and phrases in real world experiences, as well as the ability to interact socially with other language users.[11] I discuss each of these prerequisites in detail.

[9] Ryle said the mind-body problem was a result of assuming that mind, like body, was a physical entity. The category mistake of Searle, as well as of Bender and Koller (see Sect. 4, Claim 1), is to assume that language understanding can be equated with what we experience, when we are aware of our language understanding. Understanding language, I argue, or linguistic meaning, if you prefer, does not belong to the category of private, conscious experiences, but to the public and pre-conscious. It is hard to provide uncontroversial examples of the former category, but for some, *guilt* would be an example of the former, as it is possibly contingent on conscious deliberation, whereas eye motor control or emotional responses, e.g., fear or shame, are presumably uncontroversial examples of the latter.

[10] They are not the first to do so. The Churchlands (Churchland & Churchland, 1990) also replaced the rule-based interlocutor with a statistical one (a connectionist network) and argued that in this case, Searle's argument *fails*. This has come to be known as the Brain Simulator Reply to the Chinese Room. Churchland and Churchland (1990) discuss Searle's premise that semantics cannot arise out of computation, claiming this is an empirical question, and likening Searle's blanket denial of this possibility to the eighteenth-century Irish bishop George Berkeley belief that compression waves in the air, by themselves, could not constitute or be sufficient for objective sound. Jackson and Sharkey (1996) argue that while grounding of connectionist or neural models may still be difficult, it is much easier than for rule-based systems.

[11] This is reminiscent of classical critiques of artificial intelligence. Aleksander (2002), for example, says that connectionist networks lack proper interpretation, grounding, as well as emotional impact, before they can be said to understand the meaning of sentences. The three requirements correspond almost exactly to those of Bender and Koller.

3 Prerequisites of Understanding

3.1 Is Awareness Needed?

Searle (1980) argues that consciousness is a prerequisite for language understanding, and Bender and Koller, along with many others (Signorelli, 2018; Bishop, 2020), seem to share the assumption that language understanding requires some form of awareness. Bender and Koller say 'understanding depends on the ability to be aware of what another human is attending to and guess what they are intending to communicate' (p. 5190). The quote is ambiguous between two readings: (i) understanding depends on having a model of what is attended to and what the speaker is intending to communicate; (ii) understanding depends on the listener being consciously aware of the communicative situation. I take it that Bender and Koller have reading (ii) in mind, since (i) reduces to grounding. If I am mistaken, the reader can ship to Sect. 3.2, in which I discuss the grounding problem.

Claim 1 in *Bender and Koller* (2020) *For language models to understand language, they would have to be aware, e.g., of communicative intent. They are not.*

On the premise that language understanding requires awareness (Claim 1), Bender and Koller get the following *modus ponens* through: Language understanding requires awareness, and language models do not exhibit awareness. Hence, language models cannot understand language.

As mentioned in Sect. 2, Dennett (1987) argues that Searle conflates semantics and consciousness of semantics.[12] I think Bender and Koller conflate understanding and awareness of understanding in much the same way (unless awareness reduces to grounding).

Understanding language generally does not seem to require conscious awareness. Consider first the common experience of unconscious driving. You jump on your bike or get into your car to drive to work, but quickly find yourself immersed in thoughts. Perhaps you are preparing yourself for a meeting later that day, or you are thinking about the movie you saw last night. Moments later you park in front of your office, with no recollection of how you made it there. Presumably you navigated through crossings and roundabouts, stopped at traffic lights, etc., but none of this required conscious effort. While biking in traffic or driving a car is mostly a conscious effort, consciousness is no prerequisite. Dennett (1987) claims the same holds for language understanding. Note also how responding meaningfully is possible under anesthesia (Webster, 2017) or during sleep. In fact, there is a growing literature that examines whether awareness is a prerequisite for language understanding. Some researchers have found experimental evidence for unconscious language understanding (Van den Bussche et al., 2009; Sklar et al., 2012), while others

[12] Copeland (2003) argues that Searle's thought experiment's weakness lies in the assertion that it follows from the observation that an interlocutor does not understand language, that no part of her would. On a particular view of what an interlocutor (person) is, this seems equivalent to Dennett's argument.

have pushed back against the idea (Rabagliati et al., 2018). Van den Bussche et al. (2009) present several experiments that seem to suggest the possibility of unconscious language understanding, even when participants are fully awake. One of them is a lexical decision task, in which participants were exposed to sequences of letters and asked to classify these as words or non-words. Subliminal primes preceded the exposure. Some primes were semantically related, while others were completely unrelated. Semantically related primes were shown to lead to faster and more accurate responses. In another experiment, participants were asked to read target words aloud, and related subliminal primes were again shown to facilitate reading. See also (Bergson, 1896), for example. Minimally, whether language understanding requires awareness, is an empirical question.

Reply to Claim 1

Awareness is *not* necessary for producing, comprehending, or learning language.

Now, a possible reply to my rebuttal of Claim 1 is that instances of successful linguistic communication without conscious awareness do not imply that sustained conversation is possible in the absence of such conscious awareness.[13] I actually think there is some evidence for people having long conversations (dialogues) while sleep talking (Peeters & Dresler, 2014; Arnulf et al., 2017). Here's an example from (Peeters & Dresler, 2014):

> Peacock tree?! There's a tree full of peacocks and swallows and other colorations, aabsolutely, aaaabsolutely g-o-o-or-geous–get my camera, Larry, and don't forget to turn the film or I'll have fits when it gets back–come on! God!–you should see it.

But even if there was no evidence for sustained conversation in the absence of conscious awareness, this would *not*, I think, be evidence for Claim 1, for two reasons: (i) The lack of such reports may be due to other facts, i.e., people being more sensitive to distractive stimuli under anesthetics or while asleep; and more importantly, (ii) even if sustained conversations were impossible in the absence of conscious awareness, would we really say that engaging in sustained conversations is necessary for understanding language? If some patients with severe attention deficit disorder were unable to engage in sustained conversations, would we really say that they do not understand language?

3.2 Is Grounding Needed?

Bender and Koller also point out how language model representations are not *grounded*. Understanding language, in their words, relies on

[13] Thanks to one of my anonymous reviewers for raising this important concern. The reviewer also raises the possibility that this is what Bender and Koller have in mind in Claim 1. I think it is not. Bender and Koller, remember, claim that language models cannot even understand the meaning of individual words.

> mastery of the structure and use of language and the ability to ground it in the world (Bender & Koller, 2020, p. 5185)

They equate understanding with retrieving the communicative intent of a speaker and argue that *the communicative intent is grounded in the real world the speaker and listener inhabit together*. Their claims amounts to the following:

Claim 2 *Bender and Koller* (2020) *For language models to understand language, they would have to ground representations in the absence of supervision. They cannot.*

Claim 2 also licenses a *modus ponens* inference: Language understanding requires grounding, and language models (trained on raw text) do not exhibit grounding. Hence, language models cannot understand language. To highlight the role of grounding, Bender and Koller present the following *more constrained* version of their thought experiment:

> As a second example, imagine training an LM (again, of any type) on English text, again with no associated independent indications of speaker intent. The system is also given access to a very large collection of unlabeled photos, but without any connection between the text and the photos. For the text data, the training task is purely one of predicting form. For the image data, the training task could be anything, so long as it only involves the images. At test time, we present the model with inputs consisting of an utterance and a photograph, like *How many dogs in the picture are jumping?* or *Kim saw this picture and said "What a cute dog!" What is cute?*

This second thought experiment highlights the importance of grounding word representations to their argument. It also shows what their argument hinges on: If unsupervised alignment of image models and language models is possible, the argument fails completely. The empirical question then is whether *completely unsupervised* alignment of the two modalities is possible? In Sect. 5, I present evidence that weakly supervised alignment is possible, but there are good reasons to think unsupervised alignment will also be possible.

Consider, for example, the fact that unsupervised machine translation is possible (Lample et al., 2018a, b; Park et al., 2021). Unsupervised machine translation works by first aligning vector spaces induced by monolingual language models in the source and target languages (Søgaard et al., 2019). This is possible because such vector spaces are often near-isomorphic (Vulic et al., 2020). If weak supervision is available, we can use techniques such as Procrustes Analysis (Gower, 1975) or Iterative Closest Point (Besl & McKay, 1992), but aligments can be obtained *in the absence* of *any* supervision using adversarial learning (Li et al., 2019; Søgaard et al., 2019) or distributional evidence alone. If the vector spaces induced by language models exhibit high degrees of isomorphism to the physical world or human perceptions thereof, we have reason to think that similar techniques could provide us with sufficient grounding in the absence of supervision.

Unsupervised machine translation show that language model representations of different vocabularies of different languages are often isomorphic. Some researchers have also explored cross-modality alignment: (Chung et al., 2018) showed that unsupervised alignment of speech and written language is possible using the same techniques, for example. This also suggests unsupervised grounding should be possible.

Is there any direct evidence that language model vector spaces are isomorphic to (representations of) the physical world? There is certainly evidence that language models learn isomorphic representations of parts of vocabularies. Abdou et al. (2021), for example, present evidence that language models encode color in a way that is near-isomorphic to conceptual models of how color is perceived, in spite of known reporting biases (Paik et al., 2021). Patel and Pavlick (2022) present similar results for color terms and directionals. Liétard et al. (2021) show that the larger models are, the more isomorphic their representations of geographical place names are to maps of their physical location. I would therefore reply to Claim 2 in the following way.

Reply to Claim 2

Grounding is needed, but is possible from raw text.

A possible reply to this reply is that some subsets of the vocabulary, e.g., color terms or geographical names, may be easier to ground than others, because their semantics can, at least in part, be defined in geometric terms.[14] This is the difference between providing an answer to Questions 2 and 3 in Sect. 1. Whether color terms and geographical names are particularly easy is an empirical matter. I think there is reason to think going from such words to the rest of the vocabulary will not be a world of difference.

Unsupervised machine translation and unsupervised bilingual dictionary induction are evaluated over the full vocabulary, often with more than 85% precision. This indicates language models learn to represent concepts in ways that are not very language-specific. There is also evidence for near-isomorphisms with brain activity, across less constrained subsets of the vocabulary: (Wu et al., 2021), for example, show how brain activity patterns of individual words are encoded in a way that facilitates analogical reasoning. Such a property would in the limit entail that brain encodings are isomorphic to language model representations (Peng et al., 2020). Other research articles that seem to suggest that language model representations are generally isomorphic to brain activity patterns include (Mitchell et al., 2008; Søgaard, 2016; Wehbe et al., 2014; Pereira et al., 2018; Gauthier & Levy, 2019; Caucheteux & King, 2022).

3.3 Is Social interaction Needed?

It is important to note that grounding is not necessarily grounding in the physical world, but in conceptual representations of entities, events, intentions, etc. Lupyan

[14] Thanks to one of my anonymous reviewers for raising this important concern.

and Winter (2018) discuss the abstractness of language and argue that in fact, the vast majority of concepts in language have no concrete reference in the real world. The fact that language is predominantly abstract, does not mean perceptual grounding is not beneficial for language acquisition. There is plenty of evidence that lexical processing is faster for perceptually grounded words (Juhasz et al., 2011). Concrete and abstract words also differ in their representational substrates (Hoffman, 2016), and language acquisition studies also show that concrete words are learned before abstract words (Schwanenflugel, 1991; Bergelson & Swingley, 2013). However, the fact that humans tend to learn concrete words faster, through embodied social interaction, does not imply this is a necessary order in which to acquire language. That is: It does not follow that grounding is necessary for language acquisition. Bender and Koller also argue that social interaction is a necessary ingredient in language acquisition, presumably because it is a prerequisite for awareness of communicative intent. The claim can be formulated as follows:

Claim 3 *Bender and Koller* (2020) *For language models to understand language, they would have to be able to interact socially. They are not.*

Claim 3 licenses the following *modus ponens*: Language understanding requires social interaction, and language models do not exhibit social interaction. Hence, language models cannot understand language. The claim that language understanding requires social interaction is reminiscent of Wittgenstein's arguments against private language (Wittgenstein, 1953).[15] However, conflating the two would be wrong. The fact *that private language is not possible, is not the same as saying that a public language cannot be learned in private*. For example, there is plenty of research that shows learning language in the absence of social interaction, e.g., from printed materials or video, is possible (Perez & Rodgers, 2019; Ulker, 2019). Even in first language acquisition, children down to two years have been shown to learn from television (Rice, 1983; Tsuji et al., 2020). Social interaction can mean different things: Language models can be and commonly are induced from dialogue. This does not lead to social interaction, but it is hard to see why training on conversational data should be qualitatively different from training by interaction. For a sufficiently good language model, the learning signal would amount to only masking words in the input from one conversational agent. It is not clear if Bender and Koller imply that language understanding is also impossible from conversational data, but it surely makes it possible to ground representations in turn-taking dynamics.

Reply to Claim 3

[15] (Proudfoot, 2002) discuss whether Wittgenstein also believed consciousness is a prerequisite for language. He apparently did not, as paragraph 202 of *Philosophical Investigations* (Wittgenstein, 1953) suggests: *Darum ist 'der Regel folgen' eine Praxis. Und der Regel zu folgen* glauben *ist nicht: der Regel folgen.* ('And hence also 'obeying a rule' is a practice. And to think one is obeying a rule is not to obey a rule.')

Social interaction is necessary for language coming about, but *not* necessary for an individual's language learning.

4 Coconut Catapults and Angry Bears

As mentioned above, Bender and Koller argue that the octopus O eventually fails the Turing Test. To illustrate this, they present us with two continuations of the above thought experiment: (a) A calls B, reaching the octopus instead, and asks for advice on how best to design a *coconut catapult*; (b) A calls B, reaching the octopus instead, saying she is being chased by a bear and asking for advice to fight it off. It is clear from the discussion that Bender and Koller want the examples to illustrate how language models are limited by their lack of grounding and social interaction. Can you communicate meaningfully about a *coconut catapult* and *fighting off bears* without real world experiences with coconuts and catapults and bears? I think the answer is *yes*, and have obtained decent advice on such matters from existing language models,[16] but more importantly, these examples do not test language understanding, but the *experiences* of the octopus. It is important to distinguish understanding from experience: Many English speakers would not necessarily be able to give advice on designing a coconut catapult, because they lack the necessary experience. I assume readers will agree that obtaining such experience should *not* be what distinguishes those who understand English from those who do not.

In fact there is nothing preventing language models from knowing about coconut catapults or angry bears (Sahlgren & Carlsson, 2021). If the Coconut Catapult and the Angry Bear tests can be solved, other examples may of course be harder,[17] The argument of Bender and Koller is often summarized as an empirical claim about what is theoretically possible, i.e., that there will always be some test, which will be able to distinguish language models from humans.[18] This amounts to the following claim:

Claim 4 in *Bender and Koller* (2020) *For language models to understand language, they would have to be able to solve the Turing Test. They are not.*

[16] Even a small language model like BERT-base will tell you that *if you see an orc, you better* {run, know, hurry, stop, hide} (the top-5 suggestions for filling in the gap), or that *you must fight orcs with a* { *sword, spear, bow, weapon, blade*}.

[17] In an interview (Hershcovich & Donatelli, 2021) Alexander Koller mentions another example where language models may fail. He presents GPT-3 with the question *Who won the World Series in 2022?*—to which it responds *The New York Yankees won the World Series in 2022.* Again, this of course says nothing about the fundamental limits of language models, and even GPT-2, for example, seems to be able to infer from distributional evidence, that the answer to *Is 1900 in the past or in the future?* is *It is in the past*, while the answer to *Is 2022 in the past or in the future?*, is *It is in the future.* Obviously, since language models are not generally trained on time-stamped text, there is a limit to what can be inferred from distributional evidence, but this is orthogonal to the question of what can possibly be learned. In both the Chinese Room and the Octopus thought experiment, time stamps are potentially available.

[18] https://blog.julianmichael.org/2020/07/23/to-dissect-an-octopus.html.

This is a position which is different from Searle's. Searle (1980) argues that the Turing Test is insufficient for language understanding, i.e., that machines solving it still would not understand language. Bender and Koller seem to accept the Turing Test as sufficient. I believe the Turing Test has the opposite problem: Being also a test of experience, e.g., with coconut catapults, solving the test is *not a necessary condition* for exhibiting language understanding. A program that understands language, but has no idea of whether it is day or night at the time of the test, is perfectly conceivable. Not knowing whether it is a day or night, would likely lead to failing the Turing Test, however.

Reply to Claim 4

The Turing Test-style queries provided by Bender and Koller are *not* unsolvable by language models (neither in principle nor in practice), leaving it an empirical question whether language models can solve the Turing Test.

Other Arguments in Bender and Koller (2020)

Bender and Koller also briefly survey existing research that show the importance of social interaction to language acquisition. I only discuss the most prominent example, which also provides the strongest support for the claim that language understanding requires social interaction: Patricia Kuhl (2007) performed an experiment to show that toddlers can learn phonemic distinctions in Mandarin Chinese from a Mandarin-speaking experimenter, but not from exposure to Mandarin TV or radio. The result that toddlers learn phonemic distinctions faster from interacting with a human experimenter than from TV or radio, is rather unsurprising. It is well-known that social interaction facilitates learning in general (Okita, 2012). Kuhl (Kuhl et al., 2003) herself found that infants' attention to people is much stronger than attention to inanimate objects. In other words, the effect of social interaction is not language-specific.

5 Grounding the Vector Space of an Octopus

I have shown that Bender and Koller make four assumptions in their Octopus thought experiment:

1. For language models to understand language, they would have to be aware, e.g., of communicative intent. They are not.
2. For language models to understand language, they would have to ground representations in the absence of supervision. They cannot.
3. For language models to understand language, they would have to be able to interact socially. They are not.
4. For language models to understand language, they would have to be able to solve the Turing Test in the limit. They are not.

In Sect. 3.1, I have argued that it is, minimally, an empirical question if language production and comprehension relies on conscious awareness. I have also argued that the impossibility of private language does not preclude private learning of language. Further, in Sect. 3.2, I have shown that language models induce associative geometries that are stable across languages. It is not far-fetched to speculate this stability results from shared experience, and that languages are used to talk about the world (as we experience it). The induced geometries are clearly semantic in nature and thus partially fix reference. They can, at least in part, be aligned in a completely unsupervised fashion with brain images, representations of computer vision models, as well as perceptual and physical spaces. Section 4 pushed back against the arguments for why language models will not be able to solve the Turing Test.

This alone is enough to refute Claims 1–4. If humans can be said to understand language in the absence of awareness, this should be possible for language models, too. This refutes the *first* part of Claim 1: I do not claim that language models can exhibit awareness, but push back against the idea that this is a prerequisite for understanding language. Language model representations can, as an empirical fact, be grounded in cognitive, perceptual, and physical spaces with relatively high precision. This refutes the *second* part of Claim 2: I accept that grounding is important for language understanding, but claim this is possible in the absence of supervision. I also push back against the *first* part of Claim 3: While private language is impossible, this does not mean that private learning of language is impossible. So while language models may have to learn language in private, it is an empirical question if this is possible or not. I have also argued that Bender and Koller do not succeed in presenting convincing examples of where language models would fail the Turing Test. Here, however, I am unconvinced about *parts* of Claim 4. I think language models may, for all I know, pass the Turing Test some day, if not yet, but I also do *not* agree that this should be a necessary condition for understanding language. To see this why the Turing Test is not a necessary condition for language understanding, consider a language model that has an implausible personality for a human being, or which has internalized first-hand experiences accumulated over 500 years. Both would be give-aways that the respondent is not human, but would this be reason to say the respondent did not understand language?

The discussion has led me to reject all the four implicit claims in Bender and Koller. The discussion of Claims 1–3 is also reason to refute Searle's Chinese Room thought experiment. I will now present a simple thought experiment to, as clearly as possible, convey the intuition that computers can acquire referential semantics in the absence of supervision.

5.1 A Thought Experiment

Consider a common AM/FM radio receiver tuned in on a talk radio channel. The engineer who built the receiver augmented the device with a pattern recognition module, similar to that of the octopus in Bender and Koller (2020) (say, a neural language model), as well as a one-pixel camera. The radio wants nothing more than to learn the meaning of color terms. It therefore starts to consider the contexts in which

these terms occur. Notice that this, in the eyes of Searle (1980) and Bender and Koller (2020), should be no less impossible than learning to understand language in general, because understanding color terms also, in their view, relies on awareness, grounding, and social interaction. In other words, we can limit the language understanding problem to acquiring the meaning of this smaller vocabulary without loss of generality, and rely on well-established models of human color perception (Fairchild, 2005). Pursuing its goal, nevertheless, the radio notices how terms such as *yellow* and *turquoise* occur in slightly different contexts, but also how other color terms such as *violet* and *purple* occur in very similar contexts. After years of practice, it learns to represent colors in a way that is near-isomorphic to how humans perceive colors. Because its language model is contextualized, it even learns to correct for possible reporting biases (Paik et al., 2021). It now has, I argue, learned the inferential semantics (Marconi, 1997) of color terms. The radio wants more, though. It also wants to learn the referential semantics of color terms, i.e., the mapping of color terms onto pixel values. However, if the color term representation is isomorphic to the camera's representation of colors, it follows that unless the color terms lie equidistantly on a sphere, we can induce a mapping, even in the absence of supervision.

Finally, I present a novel experiment to show how the representations of language models and computer vision models converge toward being isomorphic. The experiment directly targets the more constrained thought experiment of Bender and Koller (see Sect. 3):

5.2 An Actual Experiment

To demonstrate in practice what I have suggested in the above—that unsupervised grounding of word representations learned from raw text is possible—I set up the following experiment.

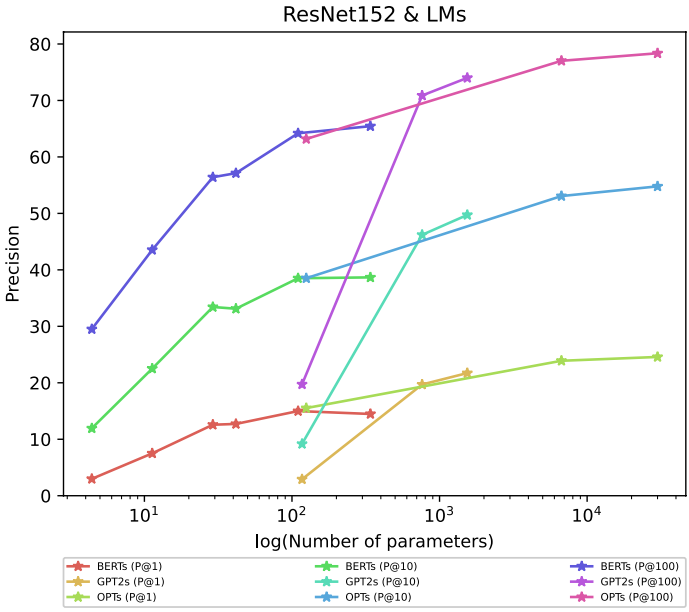
I compare the representations B of a not-too-old language models (BERT-Large) (Devlin et al., 2019) and the representations A of a not-too-old computer vision model (ResNet152) (He et al., 2016). BERT-Large is pretrained on the BooksCorpus (800 M words) and English Wikipedia (2500 M words), while the ResNet152 model is pretrained on the ImageNet-1k dataset (Russakovsky et al., 2015).

To this end, I induce a linear mapping Ω based on a few randomly selected image-word pairs. We then evaluate how close $A\Omega$ is to B by computing retrieval precision on held-out pairs. For example, we could use the pairing of the vector representation of an image of an apple in B, and the vector representation of the word 'apple' in A, as well as the vector for an image of a banana in B and the vector for 'banana' in A, to find the linear mapping Ω that would minimize their distance under Ω. If Ω then maps the image of a lemon onto the word 'lemon' as its nearest neighbor, we say that the precision at one (P@1) for this mapping is 100%. Crucially, the number of candidate vectors in the target space is 79,059, obtained by taking the union of all aliases in ImageNet-21K and a much larger English wordlist,[19] using

[19] From https://www.npmjs.com/package/wordlist-english.

only words that appear at least 5 times in the English Wikipedia. This makes the random retrieval baseline P@1 = $\frac{1}{79059}$. In practice, our mapping proves to be much more precise, reflecting the structural similarities between language model and computer vision vector spaces.

In fact, our mapping is a lot better: The so-called P@100 score—i.e., the fraction of cases in which the visual concepts are mapped onto a small neighborhood of 100 words (including inflections, spelling variations, compound words, etc.) that include the corresponding target word—is 0.68, suggesting that more than 2/3 words are correctly aligned. Moreover, by varying the size of the language model, we see that model size and precision are positively (log-linearly) correlated. Consider the plot below for details, which summarizes results with three languages modeling architectures, each represented by models of three different sizes:



The clear tendency means that eventually, with enough data, the representations of language models and computer vision models may converge entirely, demonstrating how there is already empirical evidence to suggest Bender and Koller's thought experiments rely on a false premise.

6 Related Work

Sahlgren and Carlsson (2021) also discuss the Octopus thought experiment. They quickly refute the edge cases, arguing that it is perfectly possible that language models can pass the Angry Bear test, for example. My arguments against the limit cases in Sect. 4 are similar in spirit, but I provide more detail and empirical evidence. In their discussion of the Octopus thought experiment and its general validity, Sahlgren and Carlsson (2021) take a different approach, however. Instead of arguing that grounding is possible, they argue that the term *understanding language* is ambiguous, and that in one sense, it is perfectly adequate to talk about language models understanding language. Citing Daniel Dennett's famous argument that consciousness is not an extra ingredient to our complex cognitive systems—consciousness *is* their complexity—they envoke a sense in which *understanding* simply refers to the complexity of the induced language model. On the other hand, there is another sense in which *understanding* refers to the ability to act upon language 'outside the textual modality'. This is precisely what I refer to as *grounding* in the above. Sahlgren and Carlsson (2021) agree with (Bender & Koller, 2020) that grounding language model representations in this sense of relating them to (cognitive representations of) the physical world, is impossible:

> We completely agree that no language model, no matter how much data it has been trained on and how many parameters it has, by itself will be able to understand instructions in the sense of actually performing actions in the world. But we do believe that, in principle, a language model can acquire a similar understanding and mastery of language as a human language user. (Sahlgren & Carlsson, 2021, Sect. 3.1)

Sahlgren and Carlsson (2021) thereby seem agree with (Bender & Koller, 2020) that grounding, as a correlate between form and (standing) meaning, cannot be learned from raw text. This seems identical to the position in Marconi (1997) that inferential semantics is learnable from raw text, but referential semantics is not. I do not see why this should be *a priori* true, and in the above, I have argued for why I think such a correlate may be learnable in the absence of explicit supervision.

7 Concluding Remarks

I have discussed Bender and Koller's Octopus thought experiment for the impossibility of learning language understanding from raw text, e.g., using language models, situating it in the larger context of the Turing Test, Wittgenstein (1953) and Searle (1980), as well as of the many replies to Searle's Chinese Room (Dennett, 1987; Churchland & Churchland, 1990; Jackson & Sharkey, 1996; Marconi, 1997; Haugeland, 2003; Copeland, 2003; Lupyan & Winter, 2018). I identified four claims in Bender and Koller (2020): Claims 1–3 are shared with (Searle, 1980), and I presented novel arguments against these claims and briefly discussed earlier replies. I

argued that (a) the claim that awareness is a prerequisite for understanding (Claim 1) is a category mistake, and that empirical research suggests language understanding is possible in the absence of awareness; (b) the claim that grounding is a prerequisite for understanding, applies only to referential semantics, not inferential semantics (Marconi, 1997), and grounding for referential semantics *can* be established for sufficiently good language models, even in the absence of supervision; (c) the claim that social interaction is a prerequisite for understanding language—and not merely beneficial (which I believe is well-established)—is an empirical question, and I am optimistic (and have to some degree shown) that much can be learned, e.g., from traces of social interaction. Claim 4—that language models will eventually fail the Turing Test—was relativized by Alexander Koller in Hershcovich and Donatelli (2021) and has already received pushback from Sahlgren and Carlsson (2021). I agree the claim is unwarranted. I also point out that (Bender & Koller, 2020) depart from the existing literature when presenting this fourth claim. I present a simple thought experiment that I believe is more suitable for guiding our thinking about machine acquisition of natural language understanding. The thought experiment, which I called *the Color Radio*, serves two functions: (a) it limits language understanding to the acquisition of the referential semantics of color terms, sharpening our intuitions about what language model vector spaces and human perception may look like; (b) it sketches how a representation that is isomorphic to the perceptual space of color encodings could realistically be learned from raw text alone. I supplement the thought experiment with an actual experiment showing how language models and computer vision models converge on representing the world in structurally similar ways, providing direct evidence against Bender and Koller's more constrained version of their thought experiment (Sect. 3.2).

Acknowledgement Thanks to the anonymous reviewers, as well as to Jiaang Li and Yova Kementchedjhieva for valuable feedback and help with the experiments.

Declarations

Conflict of interest The authors have no relevant financial or non-financial interests to disclose.

References

Abdou, M., Kulmizev, A., Hershcovich, D., Frank, S., Pavlick, E., & Søgaard, A. (2021). Can language models encode perceptual structure without grounding? a case study in color. In: *Proceedings of the 25th Conference on Computational Natural Language Learning*, pp. 109–132. Association for Computational Linguistics, Online.

Aleksander, I. (2002). Neural depictions of 'world' and 'self': Bringing computational understanding to the Chinese room. In J. M. Preston & J. M. Bishop (Eds.), *Views Into the Chinese room: New essays on Searle and artificial intelligence*. Oxford University Press.

Arnulf, I., Uguccioni, G., Gay, F., Baldayrou, E., Golmard, J.-L., Gayraud, F., & Devevey, A. (2017). What does the sleeping brain say? Syntax and semantics of sleep talking in healthy subjects and in parasomnia patients. *Sleep*, *40*(11).

Artetxe, M., Labaka, G., & Agirre, E. (2017) Learning bilingual word embeddings with (almost) no bilingual data. In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 451–462). Association for Computational Linguistics, Vancouver, Canada.

Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On meaning, form, and understanding in the age of data. In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online.

Bergelson, E., & Swingley, D. (2013). The acquisition of abstract words by young infants. *Cognition, 127*, 391–397.

Bergson, H. (1896). *Matter and memory*. MIT Press.

Besl, P. J., & McKay, N. D. (1992). A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 14*(2), 239–256.

Bishop, J. (2002). Views into the Chinese room: New essays on searle and artificial intelligence vol. 15.

Bishop, J. M. (2020). Artificial Intelligence is stupid and causal reasoning won't fix it.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology, 5*, 134.

Chung, Y.-A., Weng, W.-H., Tong, S., & Glass, J. (2018). Unsupervised cross-modal alignment of speech and text embedding spaces. In: *Proceedings of the 32nd International Conference on Neural Information Processing Systems*. NIPS'18, pp. 7365–7375. Curran Associates Inc., Red Hook, NY, USA.

Churchland, P. M., & Churchland, P. S. (1990). Could a machine think? *Scientific American, 262*(1), 32–7.

Copeland, B. J. (2003). The Chinese room from a logical point of view. In J. M. Preston & J. M. Bishop (Eds.), *Views into the Chinese room: New Essays on Searle and artificial intelligence*. Oxford University Press.

Copeland, B. J. (2004). *The essential turing: Seminal writings in computing, logic, philosophy, artificial intelligence, and artificial life plus the secrets of enigma*. Oxford University Press.

Dennett, D. C. (1987). Fast thinking. In: *The intentional stance*. MIT Press.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota.

Dietrich, E., Fields, C., Sullins, J., Heuveln, B. V., & Zebrowski, R. (2021). *Great philosophical objections to artificial intelligence the history and legacy of the AI wars*. Bloomsbury Publishing.

Endicott, R. P. (1996). Searle, syntax, and observer relativity. *Canadian Journal of Philosophy, 26*(1), 101–122.

Fairchild, M. D. (2005). *Color appearance models*. Wiley.

Gauthier, J., & Levy, R. (2019). Linking artificial and human neural representations of language. arXiv: 1910.01244

Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*, 447–474.

Gower, J. (1975). Generalized procrustes analysis. *Psychometrika, 40*(1), 33–51.

Haugeland, J. (2003). Syntax, semantics, physics. In J. M. Preston & M. A. Bishop (Eds.), *Views Into the Chinese room: New essays on Searle and artificial intelligence*. Oxford University Press.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep Residual Learning for Image Recognition. In: *Proceedings of 2016 IEEE Conference on Computer Vision and Pattern Recognition*. CVPR '16, pp. 770–778. IEEE.

Hershcovich, D., & Donatelli, L. (2021). Climbing the hill of computational semantics. *Künstliche Intelligenz, 35*, 361–365.

Hirst, G. (1997). Briefly noted. *Computational Linguistics*, *23*(4).

Hoffman, P. (2016). The meaning of 'life' and other abstract words: Insights from neuropsychology. *Journal of Neuropsychology, 10*(2), 317–343.

Ivan, C., & Indurkhya, B. (2019). On modelling the emergence of logical thinking.

Jackson, S. A., & Sharkey, N. E. (1996). Grounding computational engines. *Artificial Intelligence Review, 10*(1–2), 65–82.

Juhasz, B., Yap, M., Dicke, J., Taylor, S., & Gullick, M. (2011). Tangible words are recognized faster: The grounding of meaning in sensory and perceptual systems. *Quarterly Journal of Experimental Psychology, 64*, 1683–91.

Kuhl, P. K. (2007). Is speech learning 'gated' by the social brain? *Developmental Science, 10*(1), 110–120.

Kuhl, P. K., Tsao, F.-M., & Liu, H.-M. (2003). Foreign-language experience in infancy: Effects of short-term exposure and social interaction on phonetic learning. *Proceedings of the National Academy of Sciences, 100*(15), 9096–9101.

Lample, G., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Unsupervised machine translation using monolingual corpora only.

Lample, G., Ott, M., Conneau, A., Denoyer, L., & Ranzato, M. (2018). Phrase-based & neural unsupervised machine translation.

Li, Z., Wei, Z., Fan, Z., Shan, H., & Huang, X. (2021). An unsupervised sampling approach for image-sentence matching using document-level structural information. arXiv:abs/2104.02605

Li, C.-L., Zaheer, M., Zhang, Y., Póczos, B., & Salakhutdinov, R. (2019). Point Cloud GAN

Liétard, B., Abdou, M., & Søgaard, A. (2021). Do language models know the way to Rome? In: *Proceedings of the Fourth BlackboxNLP Workshop on Analyzing and Interpreting Neural Networks for NLP*, pp. 510–517. Association for Computational Linguistics, Punta Cana, Dominican Republic.

Lupyan, G., & Winter, B. (2018). Language is more abstract than you think, or, why aren't languages more iconic? *Philosophical Transactions of the Royal Society B: Biological Sciences, 373*(1752), 20170137.

Marconi, D. (1997). *Lexical competence. A Bradford book*. MIT Press.

Mitchell, T. M., Shinkareva, S. V., Carlson, A., Chang, K.-M., Malave, V. L., Mason, R. A., & Just, M. A. (2008). Predicting human brain activity associated with the meanings of nouns. *Science, 320*(5880), 1191–1195.

Naim, I., Song, Y.C., Liu, Q., Kautz, H., Luo, J., & Gildea, D. (2014). Unsupervised alignment of natural language instructions with video segments. In: *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence*. AAAI'14, (pp. 1558–1564).

Okita, S. Y. (2012). Social Interactions and Learning. In N. M. Seel (Ed.), *Encyclopedia of the sciences of learning* (pp. 3104–3107). Springer.

Paik, C., Aroca-Ouellette, S., Roncone, A., & Kann, K. (2021). The World of an Octopus: How Reporting Bias Influences a Language Model's Perception of Color. In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 823–835. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic.

Park, C., Tae, Y., Kim, T., Yang, S., Khan, M. A., Park, E., & Choo, J. (2021) Unsupervised neural machine translation for low-resource domains via meta-learning.

Patel, R., & Pavlick, E. (2022). Mapping language models to grounded conceptual spaces. In: *International Conference on Learning Representations*.

Peeters, D., & Dresler, M. (2014). Scientific significance of sleep talking. *Frontiers for Young Minds, 2*, 9.

Peng, X., Lin, C., Stevenson, M., & li, C. (2020). Revisiting the linearity in cross-lingual embedding mappings: from a perspective of word analogies.

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications9*.

Perez, M. M., & Rodgers, M. P. H. (2019). Video and language learning. *The Language Learning Journal, 47*(4), 403–406.

Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models. arXiv.

Proudfoot, D. (2002). Wittgenstein's anticipation of the Chinese room. In J. M. Preston & J. M. Bishop (Eds.), *Views Into the Chinese room: New essays on Searle and artificial intelligence*. Oxford University Press.

Rabagliati, H., Robertson, A., & Carmel, D. (2018). The importance of awareness for understanding language. *Journal of Experimental Psychology: General, 147*, 190–208.

Rice, M. (1983). The role of television in language acquisition. *Developmental Review, 3*(2), 211–224.

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision (IJCV), 115*(3), 211–252.

Ryle, G. (1938). Categories. *Proceedings of the Aristotelian Society, 38*, 189–206.

Sahlgren, M., & Carlsson, F. (2021). The Singleton fallacy: Why current critiques of language models miss the point.

Schank, R. C., & Colby, K. M. (1973). Computer models of thought and language.

Schwanenflugel, P. (1991). Why are abstract concepts hard to understand? *The Psychology of Word Meanings*, *1991*

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

Searle, J. R. (1992). *The rediscovery of the mind*. MIT Press.

Shieber, S. M. (2004). The Turing Test: Verbal behavior as the hallmark of intelligence. *Computational Linguistics, 31*, 407–412.

Signorelli, C. M. (2018). Can computers become conscious and overcome humans? *Frontiers in Robotics and AI, 5*, 121.

Sklar, A. Y., Levy, N., Goldstein, A., Mandel, R., Maril, A., & Hassin, R. R. (2012). Reading and doing arithmetic nonconsciously. *Proceedings of the National Academy of Sciences, 109*(48), 19614–19619.

Søgaard, A. (2016). Evaluating word embeddings with fMRI and eye-tracking. In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 116–121. Association for Computational Linguistics, Berlin, Germany.

Søgaard, A., Ruder, S., & Vulić, I. (2018). On the limitations of unsupervised bilingual dictionary induction. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, (pp. 778–788). Association for Computational Linguistics, Melbourne, Australia.

Søgaard, A., Vulic, I., Ruder, S., & Faruqui, M. (2019). Cross-lingual word embeddings. *Synthesis Lectures on Human Language Technologies, 12*(2), 1–132.

Tsuji, S., Jincho, N., Mazuka, R., & Cristia, A. (2020). Communicative cues in the absence of a human interaction partner enhance 12-month-old infants' word learning. *Journal of Experimental Child Psychology, 191*, 104740.

Ulker, M. (2019). The approach of learning a foreign language by watching tv series. *Educational Research and Reviews, 14*, 608–617.

Van den Bussche, E., Van den Noortgate, W., & Reynvoet, B. (2009). Mechanisms of masked priming: A meta-analysis. *Psychological Bulletin, 135*, 452–77.

Vulic, I., Ruder, S., & Søgaard, A. (2020). Are all good word vector spaces isomorphic?.

Warwick, K., & Shah, H. (2015). Passing the Turing Test does not mean the end of humanity. *Cognitive Computation, 8*, 409–419.

Webster, C. S. (2017). Anesthesia, consciousness, and language. *Anesthesiology, 127*(6), 1042–1043.

Wehbe, L., Vaswani, A., Knight, K., & Mitchell, T. (2014). Aligning context-based statistical models of language with brain activity during reading. In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 233–243. Association for Computational Linguistics, Doha, Qatar.

Wittgenstein, L. (1953). *Philosophical investigations*. Basil Blackwell.

Wu, M.-H., Anderson, A. J., Jacobs, R. A., & Raizada, R. D. S. (2021). Analogy-related information can be accessed by simple addition and subtraction of fMRI activation patterns, without participants performing any analogy task. *Neurobiology of Language, 2*, 1–17.