Synthese (2022) 200:443
https://doi.org/10.1007/s11229-022-03931-4

ORIGINAL RESEARCH

Understanding models understanding language

Anders Søgaard[1]

Received: 22 December 2021 / Accepted: 12 October 2022 / Published online: 27 October 2022

Abstract
Landgrebe and Smith (Synthese 198(March):2061–2081, 2021) present an unflattering diagnosis of recent advances in what they call *language-centric* artificial intelligence—perhaps more widely known as natural language processing: The models that are currently employed do not have sufficient expressivity, will not generalize, and are fundamentally unable to induce linguistic semantics, they say. The diagnosis is mainly derived from an analysis of the widely used Transformer architecture. Here I address a number of misunderstandings in their analysis, and present what I take to be a more adequate analysis of the ability of Transformer models to learn natural language semantics. To avoid confusion, I distinguish between inferential and referential semantics. Landgrebe and Smith (2021)'s analysis of the Transformer architecture's expressivity and generalization concerns inferential semantics. This part of their diagnosis is shown to rely on misunderstandings of technical properties of Transformers. Landgrebe and Smith (2021) also claim that referential semantics is unobtainable for Transformer models. In response, I present a non-technical discussion of techniques for grounding Transformer models, giving them referential semantics, even in the absence of supervision. I also present a simple thought experiment to highlight the mechanisms that would lead to referential semantics, and discuss in what sense models that are grounded in this way, can be said to understand language. Finally, I discuss the approach Landgrebe and Smith (2021) advocate for, namely manual specification of formal grammars that associate linguistic expressions with logical form.

Keywords Artificial intelligence · Language · Mind

1 Introduction

Cross-disciplinary investigations, such as when philosophers put artificial intelligence under scrutiny, are healthy, if not crucial. Any discipline has its blind spots, and

✉ Anders Søgaard
soegaard@di.ku.dk

[1] Department of Computer Science, Pioneer Centre for Artificial Intelligence, and Department of Philosophy, University of Copenhagen, Lyngbyvej 2, 2100 Copenhagen, Denmark

sometimes it takes a new set of eyes to push research horizons onward. Needless to say, cross-disciplinary investigations require considerable knowledge of at least two scientific fields, and it is both brave and praiseworthy when researchers embark on such endeavors.

Landgrebe and Smith (2021) present a very critical analysis of contemporary *language-centric* artificial intelligence (natural language processing)—in particular of models based on the Transformer architecture (Vaswani et al., 2017). Their article has two parts: In Sects. 1 and 2, they present their analysis of Transformer models; in Sect. 3, they present an alternative approach to modeling language. I will focus mostly on Sects. 1 and 2, but also briefly discuss the approach advocated for in Sect. 3. In these sections, Landgrebe and Smith (2021) argue that Transformer models are insufficiently expressive, exhibit poor generalization, and will never acquire linguistic semantics, never 'understand' language. There are *many* reasons to be critical of recent developments in artificial intelligence, but in this paper, I will argue that the diagnosis presented by Landgrebe and Smith (2021) is misleading in important respects, and I will show why, on the contrary, there are reasons to believe that Transformers suffer from *none* of the above weaknesses.

2 Understanding transformers

The most widely used models in natural language processing today rely on the Transformer architecture (Vaswani et al., 2017). This includes most popular pretrained language models such as BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2020), and GPT-3 (Brown et al., 2020), but Transformer models are also used across a wide range of downstream applications, including syntactic parsing (Mohammadshahi & Henderson, 2021), summarization (Gu et al., 2019), and semantic parsing (Shiv & Quirk, 2019). We first present a rough outline of how Transformer models work, and then review how they are presented in Landgrebe and Smith (2021).

Transformer models are deep neural networks, typically comprised of dozens of layers. Layers are commonly referred to as Transformer blocks. The neurons of each layer connect to the neurons of the next layer, leading to models in which learning amounts to carefully adjusting millions of numerical nobs. The input to a Transformer is typically a short text.[1] The first step is so-called tokenization of the text, a translation of the original input into a series of more meaningful entities, called tokens. There are many ways to do this, but the meaningful units roughly correspond to what we understand as words or morphemes. It is important to note that already at this point, the Transformer architecture injects an inductive bias by segmenting the input into words and morphemes. This inductive bias is a linguistically motivated bias: While learned tokenizations do not always align perfectly with how linguists would break a text into meaningful units, it correlates much better with a gold-standard segmentation than a random segmentation would.

[1] In Vaswani et al. (2017), the input is a pair of sentences, but more commonly, the input is a 512 or 1024 token window of running text or a sequence of sentences. Some Transformer models accept longer input (Zaheer et al., 2020), and others accept structured input (Brown et al., 2020).

The input tokens are then translated into vectors that represent their meaning out of context. The token vectors are combined with vectors that represent where in the short text the different tokens were located. These vectors are called position encodings. It is the combined vectors that thus represent the different tokens and where they are in the text, that are sent through the stacked Transformer blocks.

What is a Transformer block then? A Transformer block is first and foremost a way of combining information about different tokens that takes into account that tokens may be more or less important in a particular context and with a specific purpose in mind. The word *good* has a particular meaning in the sentence *Huawei's new phone is good*, and is particularly important if we were to decide whether the review provides a rationale for buying the phone or not. Consider now the sentence *Many say that Huawei's new phone is good, but I think it is average*. The word *good* obviously has the same meaning, but is less important in deciding the sentiment of the sentence. The importance of the word *good* depends on the other words in the sentence, and the Transformer architecture presents a particular way of combining the encodings of the different words with a specific purpose in mind.

This is where the calculations in Transformer blocks get a little complicated. In brief, the vectors that represent situated tokens, are multiplied into three different number matrices. This gives us three new vectors u_i, v_i, w_i for each token t_i. These are now combined with the vectors of other tokens. For each token t_i, the first vector u_i is multiplied by the second vector v_j for the other tokens, giving us a scalar value that is used to weight the third vector w_j of the second token. Each token t_i is now represented by the sum of these vectors (w_j). This summed vector now contain not only information about the original word, but also about the context in which it appeared. In each stacked Transformer block, the same operation is repeated. Each layer contains more and more abstract vector representations of the original text, and the various vector representations have been found to contain useful information for a wide range of applications in natural language processing.

Misunderstanding Transformers Landgrebe and Smith (2021), in their criticism, focus on what they see as the limited capacity of Transformer models, but their description of the capacity of Transformer models is factually wrong. They describe, for example, how a Transformer model 'encodes each single-sentence input into an encoding vector of 1024 real numbers.' This is simply not true. In a Transformer, the input is a series of, say, 512 or 1024 tokens, each represented as, say, 512- or 1024-sized vectors, but then combined through stacked Transformer blocks of multi-headed self-attention into complex sentence representations.[2] Landgrebe and Smith (2021) also claim that 'these sentence embeddings lose relations between words within the sentence.' As my presentation of the inner workings of Transformer blocks just showed, this is also false: Transformer models were specifically designed to model the relations between words. Note also that the dimensionality of neural networks are hyper-parameters that can be modified, and that the capacity of networks is only practically bounded by computational resources.

[2] It would, if you like, be more accurate to say that even a small Transformer model has at least $3 * 512 * n^2$ dimensions at its disposal at *each of its layers* for describing an n token sequence.

Landgrebe and Smith (2021) also point to a more serious limitation of (some) Transformer models, namely 'the discarding of all information pertaining to the contexts of the input sentences.' As already mentioned, some applications of the Transformer architecture model sentences independently of each other. This clearly is an important limitation, preventing resolution of inter-sentential anaphora and bridging, or disambiguation based on preceding discourse. However, most applications of the Transformer architecture *do* model context, when relevant, typically in one of three ways: either by (a) simply processing larger chunks of texts sequentially (Liu et al., 2019); (b) conditioning on context presentations (Wang et al., 2020); or (c) applying Transformer models hierarchically (Ging et al., 2020). In sum, Landgrebe and Smith (2021) thus misrepresent Transformers in three ways: by claiming Transformers have limited expressivity, fail to capture relations between words, and fail to model inter-sentential context. None of these are true. Note how all three claims pertain to inferential semantics.

3 Transformers' understanding

The main argument presented by Landgrebe and Smith (2021) against using models such as Transformers in natural language processing, has to do with referential semantics, and is reminiscent of Searle (1980) and similar thought experiments in philosophy of mind.[3] Landgrebe and Smith (2021) claim that a Transformer model is necessarily shallow because 'the vector space it uses is merely morpho-syntactical and lacks semantic dimensions.' This is clearly a false statement under most definitions of morpho-syntax and semantics, since Transformer models obtain very good results on tasks such as topic classification and machine translation. If Transformer models only encoded morpho-syntactic information, they would not be able to distinguish between *I just ate an apple* and *I never painted a lion*,[4] making topic classification and machine translation blind guessing.

Here is what I think they mean, though: Transformer models at best learn inferential semantics, not referential semantics (Marconi, 1997).[5] Landgrebe and Smith (2021) define understanding language as seeings its relevance to actions and thoughts, and argue this is what Transformer models cannot do. Seeing the relevance of words and phrases to actions and thoughts seems to decompose into the following two properties:

(a) Lexical representations are grounded in representations of the (physical, social, and mental) world; and

[3] The first of these goes back to Leibniz and is often referred to as Leibniz' Mill. See also Bender and Koller (2020) for a more recent such thought experiment, albeit with more limited scope.

[4] To see this, note that both sentences are transitive sentences in past tense with a first person pronominal subject and a singular common noun object, modified by an adverb. From a strictly syntactical perspective, the two sentences are thus indistinguishable.

[5] Inferential semantics concerns relationships between expressions; referential semantics concerns relationships between expressions and referents. That is: If you are asked to name animals that are vertebrates, you are asked to do inferential semantics. If you are asked to point to animals (in images or real life), you are asked to do referential semantics.

(b) the agent is aware or conscious of such grounding.[6]

Inferential semantics refers to the part of semantics that is concerned with valid inferences. In lexical semantics, this involves establishing relations of synonymy, antonymy, hyponymy, etc. The output of such lexicographic exercises is often a database, which is best thought of as a graph with lexemes as nodes and with edges corresponding to lexical relations. The situation gets more complex at the sentence or discourse level, revolving around discourse relations such as entailment, contrast, consequence and explanation. Referential semantics is the part of semantics that concerns denotation, whether in terms of truth-conditions, mental representations or situations in which using a word is deemed appropriate.

Both (a) and (b) clearly concern *referential* semantics. Marconi (1997) remarks that Searle's Chinese Room only applies to referential semantics, not inferential semantics, since category effects, word associations, etc., are unconscious processes. Landgrebe and Smith (2021) thus seem to agree with Searle (1980) on the inherent limitations of Transformer models: lack of proper grounding and lack of consciousness.[7] In Sect. 4, I will discuss handwritten grammars, which Landgrebe and Smith (2021) claim do *not* suffer from such limitations. (Searle, however, would.)

Grounding The grounding problem (Harnad, 1990; Jackson & Sharkey, 1996) is the problem of learning a mapping from words and phrases to the objects and events they refer to, or to cognitive representations thereof. How can deep neural network models such as Transformer models learn to ground their representations in this way? Inferential semantics, i.e., relations between words and sentences, are induced implicitly by most learning objectives used to train these models. The most commonly used learning objective today is perhaps masked language modeling (Devlin et al., 2019), but the same holds for the translation objective in Vaswani et al. (2017), for example. Transformer models can therefore be straight-forwardly evaluated as models of inferential semantics. How can Transformer models encode referential semantics, though?

Well, the details of this will depend on what we think it is that linguistic expressions refer to. Let us, for example, assume linguistic expressions refer to vectors in an embedding space of neural activations or (fMRI/EEG) images thereof. If mental imagery is defined broadly enough, this should be compatible with some forms of internalist semantics (Rapaport, 1994; Schank & Colby, 1973), but note the vector space could also be a perceptual or physical space. Referential semantics or grounding now amounts to learning a mapping between the Transformer model vector space and this target space.[8]

[6] Strictly speaking, defining understanding in terms of relevance attribution (or grounding) does not imply awareness of relevance attribution (or grounding). Most mental verbs are ambiguous in this way. If relevance attribution is not intended by Landgrebe and Smith (2021) to carry this connotation, some of my arguments below can be ignored.

[7] Such a claim about Transformers would also find support among *some* computer scientists (Sahlgren & Carlsson, 2021; Piantadosi & Hill, 2022).

[8] Mapping means projecting these spaces onto one another. In object recognition, this process is called *point set registration*, and is something that humans seem to be endowed with from nature's hand (Serre et al., 2007), enabling us recognize objects under different conditions, from different angles, or at different distances. Mapping of language model vector spaces have previously been used in unsupervised machine translation (Artetxe et al., 2017; Søgaard et al., 2019), and often does not require supervision (Aldarmaki et al., 2018; Conneau et al., 2017; Zhang et al., 2017).

But why, you may ask, would language model vector spaces be isomorphic to representations of our physical, mental and social world? After all, language model vector spaces are induced merely from higher-order co-occurrence statistics. I think the answer is straight-forward: Words that are used together, tend to refer to things that, in our experience, occur together. When you tell someone about your recent hiking trip, you are likely to use words like *mountain*, *trail*, or *camping*. Such words, as a consequence, end up close in the vector space of a language model, while being also intimately connected in our mental representations of the world. If we accept the idea that our mental organization maps (is approximately isomorphic to) the structure of the world,[9] the world-model isomorphism follows straight-forwardly (by closure of isomorphisms) from the distributional hypothesis.

There is plenty of evidence that Transformer-based language models encode words in ways that are near-isomorphic to where neural activation occurs when listening to or reading these words (Pereira et al., 2018; Caucheteux & King, 2021),[10] to how our perceptual spaces are organized (Abdou et al., 2021; Patel & Pavlick, 2022), as well as to how physical spaces are organized (Liétard et al., 2021).[11] Such global similarities can also be induced from local ones: Wu et al. (2021) show how brain activity patterns of individual words are encoded in a way that facilitates analogical reasoning—the same analogical reasoning that language models facilitate. Such a property would in the limit entail that brain encodings are isomorphic to language model representations (Peng et al., 2020). To see this, consider an example of analogical reasoning: 'Berlin is to Germany, what Copenhagen is to ____'. In a language model, this is computed by subtracting the vector for Germany from the sum of the vectors for Berlin and Copenhagen, returning the nearest neighbor for the resulting vector. For today's language models, the result would most likely be the vector for Denmark. If you can compute all possible analogies using vector off-set this way, you have induced a structure that is isomorphic to (the current geopolitical) reality. If you can compute the same analogies by offset of brain imaging vectors, these two spaces must be near-isomorphic. And language models can thus be grounded in brain imaging spaces.

To flesh out my argument that Transformers (and similar neural network architectures) can learn grounding, I present the Color Radio thought experiment, about

[9] This view has a long history in the philosophy of mind, from Spinoza or Leibniz over Wittgenstein's *Tractactus* to cognitive scientists such as Ron Sheppard. Predictive coding (Clark, 2013; Schrimpf et al., 2021) is, among other things, an attempt to explain this second-order isomorphism between world and mind. Priming experiments in psycholinguistics also provide evidence for lexical organization in the brain structured around distributional similarity.

[10] This line of research began when researchers at Carnegie Mellon University, in a seminal study from 2014, had eight participants read Chapter 9 of Harry Potter and the Sorcerer's Stone in a brain scanner (Wehbe et al., 2014). The output was fMRI images with high spatial resolution, but low temporal resolution, and initially, the data was used for classical cognitive science studies. One or two years later, however, researchers interested in lexical semantics revisited the data to see what the Harry Potter data could tell us about words. Researchers also developed various smoothing techniques enabling more fine-grained, token-by-token analysis. Søgaard (2016) used the data to see whether there exists a linear transformation from language model representations of tokens into the vector space induced by these fMRI images. This fact there (almost) does, is exploited by researchers who work on brain-computer interaction.

[11] This body of evidence is supplemented by evidence that language model vector spaces for different languages are near-isomorphic (Søgaard et al., 2019). If the spaces are isomorphic to mental, perceptual, or physical spaces, this immediately explains the cross-lingual similarities.

grounding of color terms:[12] 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 or a modern 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 linguistic contexts in which these terms occur. Since the talk radio channel signal is not aligned with the input of its camera, it cannot use co-occurrence statistics to ground these terms in its color perception. Notice also that the problem of grounding color terms, in the eyes of Searle (1980), should be as impossible as learning to understand language in general. The representations of the receiver's language model were induced 'in a vat', so to speak. 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. Technically, it computes the co-occurrence statistics of these color terms and embeds these in a low-dimensional vector space. 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. It now has learned the inferential semantics 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, by straight-forward methods that are humans also seem to be endowed with.[13] The Color Radio thought experiment is designed to suggest the plausibility of unsupervised grounding, and is as such intended as *both* a rebuttal of Searle (1980) and Landgrebe and Smith (2021).

In sum, my argument for why (unsupervised) grounding of Transformers is possible, goes as follows:

Premise (P1)	'Transformer language model vector spaces are near-isomorphic across languages and often also with brain imaging, perceptual and physical spaces.'
Premise (P2)	'Two near-isomorphic vector spaces can be aligned with minimal supervision, and often without supervision.'
Conclusion	'Transformer language model vector spaces can be aligned with minimal supervision, and often without supervision.'

Both premises have empirical support, and the conclusion is derived by a simple application of *modus ponens*.

Awareness Landgrebe and Smith (2021)'s definition of understanding as seeing the relevance of words and phrases to actions and thoughts, was shown to decompose

[12] Color terms are useful in thought experiments, because we have intuitions about color, and there exists well-established models of human color perception.

[13] Unsupervised grounding in this case requires nothing except basic adversarial learning and linear projection. Both are commonly evoked in cognitive science.

into grounding and awareness. I will argue with Dennett (1987) that seeing awareness as a prerequisite for understanding, rests on a category mistake (Ryle, 1938).[14] The category mistake of Searle, as well as of Landgrebe and Smith (2021), is to assume that language understanding can be equated with what we experience, when we are aware of our language understanding. Understanding language, we argue, or linguistic meaning, if you prefer, does not belong to the category of private, conscious experiences, but to categories of processes that are orthogonal to consciousness. It is generally easy to conflate these, because our introspection suffers from a severe sampling bias: When we think of instances of our own language production, we naturally tend to think about instances in which we were conscious of our language production.

Now ask yourself this: Does linguistic meaning *really* imply awareness of linguistic meaning? Does understanding really imply awareness of understanding? Dennett (1987) argued that Searle (1980) conflates understanding and *awareness of* understanding. Leibniz already emphasized the importance of processes of understanding and reflection that we are unaware of. It certainly seems possible to produce semantically fluent sentences in the absence of conscious thought, e.g., during sleep or under anesthesia (Webster, 2017). Patients that are unconscious—as defined by the Glasgow Coma Scale—reportedly react to and can remember verbal communication, even if they are not able to respond. Comatose patients also seem to comprehend language. Van den Bussche et al. (2009) present several experiments that 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.[15] This all suggests that meaning does not require conscious reflection on relevance and attribution. And if so, machines simply do not need consciousness to acquire linguistic meaning.

That we are prone to this category mistake is unsurprising: When we recollect memories of communicating with others, memories of understanding what others were saying, we almost by definition recall events in which we were in fact aware of this process of understanding. It is much easier to recall events you were conscious of than events you were not. Our introspection thus suffers from severe sampling bias, so to speak. This holds true for things we do. Consider 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. Nevertheless, if you were asked

[14] Gilbert Ryle said the mind-body problem was a result of assuming that mind, like body, was a physical entity.

[15] Such effects have been explored before, including in Bergson (1896).

what it feels like to bike to work, you would likely recall events in which you were conscious of biking to work.

My argument for why awareness is irrelevant for the ability of Transformers to learn referential semantics, is simply that awareness is irrelevant for this pursuit. This follows directly from the empirical observation that language understanding can be unconscious.

Approximation Transformer models are induced from finite amounts of data and hence *approximative*. If trained on more (representative) data, they will likely learn to better approximate the inferential and referential aspects of semantics. Landgrebe and Smith (2021) find this disturbing and write: 'Even at their very best, they remain approximative, and so any success they achieve is still, in the end, based on luck.' This, though, is a fallacy. Mastery of archery or talent for counseling is also approximative, but not a matter of luck. While no one—neither a chess computer nor a grand champion—is able to compute the optimal next chess move in real time, because of the doubly exponential search space, a skilled chess player will nevertheless win over me a hundred games in a row. While his craft is in the same sense approximative, any attempt to reduce the difference between us to luck would be ridiculous. Human language acquisition, by the way, is also approximative. This was the most prominent counter-argument against another classical argument for the impossibility of machine understanding of language, namely Gold's Theorem (Gold, 1967). In fact, Gold's Theorem seems to provide some motivation for saying approximation is *necessary* for language understanding. This follows from the fact that language is a moving target, and that members of a linguistic community exhibit a great deal of variation, speaking slightly different dialects, sociolects, and idiolects. A learning algorithm that would iterate through all possible grammars and only discriminate between (exactly) correct and incorrect ones, would never terminate in the face of such variation.

That is: I argue that the approximative nature of Transformer models, like their possible lack of awareness, is orthogonal to their ability to learn referential semantics. This follows from two relatively uncontroversial assumptions, namely that language exhibits drift and inter-speaker variation, and that this makes it possible to identify a language exactly:

Premise (P1)	'Language is a moving target, over time and between speakers.'
Premise (P2)	'Moving targets can be approximated, not modeled exactly.'
Conclusion	'Language models can only approximate language.'

The Robustness of Transformers In a final point of criticism, Landgrebe and Smith (2021) also suggest that Transformer models

> will quickly become invalid if the input-output relationship changes on either side even in some minor way. This is because the model does not generalize. Once fed with data as input that do not correspond to the distribution it was trained with, the model will fail.

Table 1 Landgrebe and Smith (2021)'s claims about Transformer architectures. Transformers *are* approximative, but this is not prohibitive of language understanding

Claim	True	False
Limited expressivity		×
No word-word interactions		×
No context-sensitivity		×
No referential semantics		×
No generalization		×
Approximative	×	

Landgrebe and Smith (2021) are here concerned with the robustness of deep neural networks, e.g., Transformer models, under distributional shift. This is an important subfield of artificial intelligence, and many researchers have devoted their careers to learning good models under distributional shift. Sometimes this literature is referred to as *domain adaptation* or *transfer learning* (Søgaard, 2013). While domain adaptation remains a challenge, language models based on Transformers are among the most robust models in artificial intelligence, and it is certainly false to say that they *become invalid* if the input-output relationship changes moderately.[16] In other words: The claim that Transformers generally exhibit poor generalization and low performance is inconsistent with empirical observations.

In Sect. 2, I showed how Landgrebe and Smith (2021) misrepresented how inference works in Transformers in three ways. In this section, I have discussed three other claims by Landgrebe and Smith (2021), pertaining to their learning capacity: Landgrebe and Smith (2021) claim Transformers cannot acquire referential semantics and cannot learn to generalize outside of their training data. I presented a mixture of arguments and empirical evidence in an atttempt to refute both claims. Moreover, on my way, I also discussed how awareness is generally not a prerequisite for understanding, and how the fact that machine learning models, including Transformer models, are approximate by nature by no means disqualify them as models of language. I summarize my discussion of Landgrebe and Smith (2021)'s critique of Transformers in the table below.

4 Handwritten grammars

Earlier critiques of Transformers and related architectures in natural language processing focused on showing language understanding is unlearnable from raw text (Bender & Koller, 2020), i.e., in the absence of supervision, that language models based on

[16] Hsieh et al. (2019), for example, show how Transformer models tend to be much more robust than earlier models, including so-called *recurrent* neural networks, across tasks such as sentiment analysis and textual entailment. Hendrycks et al. (2020) make similar observations for more downstream applications. Landgrebe and Smith (2021) seem underwhelmed by the performance of Transformer architectures, in general. They note that Vaswani et al. (2017) report so-called BLEU scores of 28.4 for English-German and 41.8 for English-French and write how *75-85 could be achieved in theory* and *would correspond to the translation abilities of an average bilingual speaker*. The Transformer scores are, in contrast, 'low', in their view. Human translators do not exhibit much better BLEU scores, however. In the original paper introducing the BLEU metric (Papineni et al., 2002), the BLEU scores reported for two human translators were 19.3 and 25.7, respectively.

Transformers are uninterpretable (Boge, 2021), or that they tell us nothing about linguistic competencies (Dupre, 2021). Landgrebe and Smith (2021) argue that language understanding is unlearnable for Transformers, even *with* supervision. They are not interested in interpretability or the ability to distill linguistic theories of competence, merely the learning of inferential and referential semantics.[17] This section briefly discusses the alternative proposed by Landgrebe and Smith (2021) to such deep neural learning architectures: handwritten grammars mapping sentences to logical form. I will argue that if language understanding *is* out of reach for deep neural network architectures, it must also be out of reach for handwritten grammars with logical form.

The approach of Landgrebe and Smith (2021) is a *pipeline* approach. They first use a shallow form of syntactic analysis called *part-of-speech tagging* to induce the syntactic categories of the input words in context. The authors then rely on a 'proprietary AI-algorithm chain that uses world knowledge in the form of a dictionary of lexemes and their word forms along with associated rules, relating, for example, to the transitivity and intensionality of verbs'.[18] This proprietary algorithm chain maps the input to logical form, a process which 'requires world knowledge, for example about temporal succession, which is stored in the computer using ontologies'.

How would this approach to text processing or text generation be *more* meaningful than Transformer models? One argument that perhaps it is not, runs as follows: Assume a handwritten grammar g, following the pipeline approach of Landgrebe and Smith (2021). Assume also g 'understands' language. The Transformer architecture is Turing-complete (Pérez et al., 2019). This means that there is a translation function τ from any handwritten grammar that can be implemented as a Turing machine into an isomorphic Transformer, i.e., $\tau(g) = t$. If g 'understands' language, so does t. So for any handwritten grammar that understands language, there exists a Transformer model that also understands language. Q.E.D.

In fact, the steps of the pipeline approach in Landgrebe and Smith (2021) *have* (all) been modelled by Transformer architectures.[19] Probing experiments suggest that even moderate-sized Transformer-based language models learn similar pipelines from just doing masked language modeling at scale (Tenney et al., 2019). Transformers could also be trained specifically to *simulate* the pipeline approach of Landgrebe and Smith (2021). Since this form of teacher-student training (Fan et al., 2018) can be done on raw text, the Transformer model would in the limit become functionally indistinguishable from the pipeline system. For Searle (1980), none of these steps would capture linguistic meaning. For Landgrebe and Smith (2021), it seems the

[17] Interpretability is one of the advantages of handwritten grammars. Such grammars arguably also map more directly on to what (some) linguists consider linguistic competence.

[18] It is unfortunate that the main step in this pipeline is a proprietary algorithm chain. Progress in artificial intelligence is often attributed to the fact that the vast majority of learning algorithms are made publicly available upon publication, for public scrutiny and adaptation.

[19] The chain's first step, i.e., part-of-speech tagging, has been successfully solved with Transformers (Tsai et al., 2019). Moreover, part of speech is generally considered orthogonal to meaning. How about the ontologies? The ontologies used by Landgrebe and Smith (2021) are similar to the knowledge bases used to ground Transformer models (Zhang et al., 2019). So it seems it cannot be the mere use of ontologies either. We are left with the mapping into logical form. However, Transformer models are also used for mapping text input into logical form, e.g., Babu et al. (2021). In sum, Transformers have been successfully used to model all the components of the pipeline advanced by Landgrebe and Smith (2021).

trouble is that you cannot have it both ways: If you think a grammar mapping sentences into logical form can capture linguistic meaning, you have to admit that the same is possible for Transformer models and other forms of deep neural networks.

Somewhat surprisingly, Landgrebe and Smith (2021) do not discuss the fact that the classical arguments of Searle and Dreyfus against the possibility of machine understanding of language were presented with such handwritten grammars in mind. I think Transformers and related neural architectures present real advantages over handwritten grammars. These advantages have nothing to do with expressivity, word-word interactions, and context-sensitivity, but with their *explanatory power*. Transformers can be used to make theories of learning testable, while handwritten grammars cannot. Consider, for example, the hypothesis that the semantics of directionals is *not* learnable from next-word prediction alone. Such a hypothesis can be falsified by training Transformers language models and seeing whether their representation of directionals is isomorphic to directional geometry; see Patel and Pavlick (2022) for details. Transformers and related architectures, in this way, provide us with practical tools for evaluating hypotheses about the learnability of linguistic phenomena.

5 Concluding remarks

I have argued that Transformers and related architectures seem able to learn both inferential and referential semantics. Clearly, you can do more with language than inferential and referential semantics, and some of these things are well beyond what you can ask a language model to do. If I ask you to walk like a penguin, I ask you to do something that language models cannot do. What we do with language is to many an important part of its meaning, and if so, language models learn only part of the meaning of language. Many linguists and philosophers have tried to distinguish between referential semantics and such embedded practices. Wittgenstein (1953), for example, would think of referential semantics—or the ability to point—as a non-privileged practice. While Wittgenstein does not give special attention to this 'pointing game', it has played an important role in psycholinguistics and anthropology, for example. Language models play many languages better than us, e.g., writing poetry or jokes, translating or summarizing texts, or spotting grammatical errors—but the pointing game has been the litmus test for machine understanding of language since Searle's Chinese Room, and it is widely used to probe for lexical semantics.

Language models have other limitations: You cannot encode the precise semantics of second-order quantifiers like *most of* in vector space. For a finite set of pairs of sets, it can learn the right inferences, e.g., that *most* members of A are also members of B, but only for a limited set of cases. So what do we make of a language model that can do the pointing game, as well as the other games just mentioned, but only decide whether most members of A are also in B, if A and B are sufficiently small? My answer is: Well, what would me make of, say, a 14-year old child with the same skills? If a 14-year old child can point to the referents of Italian nouns, translate Italian sentences into another language, summarize documents written in Italian, but only decide whether *la maggior parte delle A sono B* for small A and B, would you not say this child still speaks Italian? The requirement that you can apply all words

correctly in all cases is a very high bar for saying someone understands a language. Just like knowing a strawberry is a nut, is not generally seen as a test of one's ability to understand English. Also, recall that Landgrebe and Smith (2021) are not claiming that Transformers have insufficient levels of referential semantics. Rather, they claim Transformers have *no* referential semantics. In other words, any signs of referential semantics would challenge their claims.

I have other, more serious quarrels with Transformers: They are slow and costly to train, with terrible carbon footprints, and they exhibit slow inference times. They generally require GPUs, which are inaccessible in many parts of the world. The word segmentation algorithms and positional encoding schemes typically used in conjunction with the Transformer architecture are biased toward fusional (mostly Indo-European) languages. Each of these points is reason to consider alternatives to Transformer models. The arguments put forward by Landgrebe and Smith (2021) against Transformer models, however, are problematic.

One contribution of this work was the defense of Transformers and related neural architectures against a series of false claims, i.e., that they exhibit limited expressivity, are unable to capture word-word dependencies, are not sensitive to context, and do not generalize well. Another contribution was an in-depth discussion of another claim presented by Landgrebe and Smith (2021), namely that Transformer models are incapable of understanding language, in the sense of 'hatching on' to the world. I introduced a distinction between inferential and referential semantics, originally presented by Marconi (1997), making it clear that this argument only concerns referential semantics. I then pointed to a recent finding in the artificial intelligence literature: The observation that unsupervised alignment of isomorphic representations enables grounding of language models in mental representations or representations of the physical world. This observation makes referential semantics in neural networks possible, under very permissive assumptions. All that such grounding requires is learning a linear projection into the mental or physical space. This is sufficient, since language model vector spaces have been shown to be near-isomorphic to mental, perceptual, and physical spaces. Projections into such spaces can easily be learned when supervision is available, using point-set-registration or graph alignment algorithms, but it has also been shown that this can even be done in the absence of supervision, e.g., with generative adversarial networks. I provided a thought experiment called the Color Radio to provide some intuition how such grounding could be obtained in practice.

In Sect. 4, I addressed the hybrid pipeline approach to natural language understanding advanced by Landgrebe and Smith (2021), showing that any of the components of their pipeline could be replaced by Transformers without changing the underlying function. Finally, I discussed other limitations of modern-day language models: They are slow and costly to train, have terrible carbon footprints, exhibit slow inference, and require costly GPUs. This is all orthogonal to my discussion of Landgrebe and Smith (2021), of course. There are also obvious limitations to what you can cram into a vector, e.g., the semantics of second-order quantifiers. The question is whether *this* is

relevant for language understanding. I leave this question—as well as the question of how such semantics would be represented in *our* long-term memory—open for now.

Declarations

Conflicts of Interest The author received no funding for this work and has no conflicts of interest to declare.

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), Online. Association for Computational Linguistics.

Aldarmaki, H., Mohan, M., & Diab, M. (2018). Unsupervised word mapping using structural similarities in monolingual embeddings. *Transactions of the Association for Computational Linguistics, 6*, 185–196.

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 (Vol. 1: Long Papers)* (pp. 451–462), Vancouver, Canada. Association for Computational Linguistics.

Babu, A., Shrivastava, A., Aghajanyan, A., Aly, A., Fan, A., & Ghazvininejad, M. (2021). Non-autoregressive semantic parsing for compositional task-oriented dialog. In: *Proceedings of the 2021 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies* (pp. 2969–2978), Online. Association for Computational Linguistics.

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* (pp. 5185–5198), Online. Association for Computational Linguistics.

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

Boge, F. (2021). Two dimensions of opacity and the deep learning predicament. *Minds and Machines*.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., Agarwal, S., Herbert-Voss, A., Krueger, G., Henighan, T., Child, R., Ramesh, A., Ziegler, D., Wu, J., Winter, C., & Amodei, D. (2020). Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 1877–1901). Curran Associates Inc.

Caucheteux, C. & King, J.-R. (2021). Language processing in brains and deep neural networks: Computational convergence and its limits. *bioRxiv*.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *The Behavioral and Brain Sciences, 36*(3), 181–204.

Conneau, A., Lample, G., Ranzato, M., Denoyer, L., & Jégou, H. (2017). Word translation without parallel data. *CoRR*, abs/1710.04087.

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 (Vol 1: Long and Short Papers)* (pp. 4171–4186), Minneapolis, Minnesota. Association for Computational Linguistics.

Dupre, G. (2021). (what) can deep learning contribute to theoretical linguistics? *Minds and Machines, 31*(4), 617–635.
Fan, Y., Tian, F., Qin, T., Li, X.-Y., & Liu, T.-Y. (2018). Learning to teach. In: *International conference on learning representations*.
Ging, S., Zolfaghari, M., Pirsiavash, H., & Brox, T. (2020). Coot: Cooperative hierarchical transformer for video-text representation learning. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 22605–22618). Curran Associates Inc.
Gold, E. M. (1967). Language identification in the limit. *Information and Control, 10*(5), 447–474.
Gu, J., Wang, C., & Zhao, J. (2019). Levenshtein transformer. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d' Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.
Harnad, S. (1990). The symbol grounding problem. *Physica D: Nonlinear Phenomena, 42*(1), 335–346.
Hendrycks, D., Liu, X., Wallace, E., Dziedzic, A., Krishnan, R., & Song, D. (2020). Pretrained transformers improve out-of-distribution robustness. In: *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 2744–2751), Online. Association for Computational Linguistics.
Hsieh, Y.-L., Cheng, M., Juan, D.-C., Wei, W., Hsu, W.-L., & Hsieh, C.-J. (2019). On the robustness of self-attentive models. In: *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 1520–1529), Florence, Italy. Association for Computational Linguistics.
Jackson, S. A., & Sharkey, N. E. (1996). Grounding computational engines. *Artificial Intelligence Review, 10*(1–2), 65–82.
Landgrebe, J., & Smith, B. (2021). Making ai meaningful again. *Synthese, 198*(March), 2061–2081.
Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., & Zettlemoyer, L. (2020). BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: *Proceedings of the 58th annual meeting of the Association for Computational Linguistics* (pp. 7871–7880), Online. Association for Computational Linguistics.
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), Punta Cana, Dominican Republic. Association for Computational Linguistics.
Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., & Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692.
Marconi, D. (1997). *Lexical competence*. Bradford Book: A Bradford book.
Mohammadshahi, A., & Henderson, J. (2021). Recursive non-autoregressive graph-to-graph transformer for dependency parsing with iterative refinement. *Transactions of the Association for Computational Linguistics, 9*, 120–138.
Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics* (pp. 311–318), Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
Patel, R., & Pavlick, E. (2022). Mapping language models to grounded conceptual spaces. In: *International Conference on Learning Representations*.
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. J., Kanwisher, N. G., Botvinick, M. M., & Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications, 9*.
Pérez, J., Marinkovic, J., & Barceló, P. (2019). On the turing completeness of modern neural network architectures. In: *7th international conference on learning representations, ICLR 2019, New Orleans, LA, USA, May 6–9, 2019*. OpenReview.net.
Piantadosi, S. T., & Hill, F. (2022). Meaning without reference in large language models.
Rapaport, W. J. (1994). Chapter 10 - syntactic semantics: Foundations of computational natural-language understanding. In E. Dietrich (Ed.), *Thinking computers and virtual persons* (pp. 225–273). Academic Press.
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*. W H Freeman.

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., & Fedorenko, E. (2021). The neural architecture of language: Integrative modeling converges on predictive processing. *bioRxiv*.

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

Serre, T., Kouh, M., Cadieu, C., Knoblich, U., Kreiman, G., & Poggio, T. (2007). A theory of object recognition: Computations and circuits in the feedforward path of the ventral stream in primate visual cortex. *AI Memo 2005-036, CBCL Memo*.

Shiv, V., & Quirk, C. (2019). Novel positional encodings to enable tree-based transformers. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc.

Søgaard, A. (2013). Semi-supervised learning & domain adaptation in natural language processing. *Synthesis Lectures on Human Language Technologies, 6*(2), 1–103.

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), Berlin, Germany. Association for Computational Linguistics.

Søgaard, A., Vulić, I., Ruder, S., & Faruqui, M. (2019). *Cross-lingual word embeddings*. Synthesis lectures on human language technologies (2nd Ed.). Morgan & Claypool Publishers.

Tenney, I., Das, D., & Pavlick, E. (2019). BERT rediscovers the classical NLP pipeline. In: *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 4593–4601), Florence, Italy. Association for Computational Linguistics.

Tsai, H., Riesa, J., Johnson, M., Arivazhagan, N., Li, X., & Archer, A. (2019). Small and practical BERT models for sequence labeling. In: *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)* (pp. 3632–3636), Hong Kong, China. Association for Computational Linguistics.

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

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates Inc.

Wang, Z., Duan, Z., Zhang, H., Wang, C., Tian, L., Chen, B., & Zhou, M. (2020). Friendly topic assistant for transformer based abstractive summarization. In: *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 485–497), Online. Association for Computational Linguistics.

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

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS ONE, 9*(11), e112575.

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.

Zaheer, M., Guruganesh, G., Dubey, K. A., Ainslie, J., Alberti, C., Ontanon, S., Pham, P., Ravula, A., Wang, Q., Yang, L., & Ahmed, A. (2020). Big bird: Transformers for longer sequences. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, & H. Lin (Eds.), *Advances in neural information processing systems* (Vol. 33, pp. 17283–17297). Curran Associates Inc.

Zhang, M., Liu, Y., Luan, H., & Sun, M. (2017). Adversarial training for unsupervised bilingual lexicon induction. In: *Proceedings of the 55th annual meeting of the Association for Computational Linguistics (Vol. 1: Long Papers)* (pp. 1959–1970), Vancouver, Canada. Association for Computational Linguistics.

Zhang, Z., Han, X., Liu, Z., Jiang, X., Sun, M., & Liu, Q. (2019). ERNIE: Enhanced language representation with informative entities. In: *Proceedings of the 57th annual meeting of the Association for Computational Linguistics* (pp. 1441–1451), Florence, Italy. Association for Computational Linguistics.