Synthese (2024) 203:15
https://doi.org/10.1007/s11229-023-04435-5

ORIGINAL RESEARCH

The ambiguity of BERTology: what do large language models represent?

Tommi Buder-Gröndahl[1]

Received: 22 June 2023 / Accepted: 17 November 2023 / Published online: 26 December 2023

Abstract
The field of "BERTology" aims to locate linguistic representations in large language models (LLMs). These have commonly been interpreted as representing structural descriptions (SDs) familiar from theoretical linguistics, such as abstract phrase-structures. However, it is unclear how such claims should be interpreted in the first place. This paper identifies six possible readings of "linguistic representation" from philosophical and linguistic literature, concluding that none has a straight-forward application to BERTology. In philosophy, representations are typically analyzed as cognitive vehicles individuated by intentional content. This clashes with a prevalent mentalist interpretation of linguistics, which treats SDs as (narrow) properties of cognitive vehicles themselves. I further distinguish between three readings of both kinds, and discuss challenges each brings for BERTology. In particular, some readings would make it trivially false to assign representations of SDs to LLMs, while others would make it trivially true. I illustrate this with the concrete case study of structural probing: a dominant model-interpretation technique. To improve the present situation, I propose that BERTology should adopt a more "LLM-first" approach instead of relying on pre-existing linguistic theories developed for orthogonal purposes.

Keywords Linguistic representation · Language model · Deep learning · BERTology

1 Introduction

Improving the explainability of *large language models* (LLMs) such as *BERT* (Devlin et al., 2019) is a pressing concern in natural language processing (NLP). A reaseach program titled "BERTology" aims to discover linguistic representations in LLMs' activation patterns, with techniques like *structural probing* (Rogers et al., 2020). In a notable departure from traditional connectionist NLP, BERTology makes heavy use

✉ Tommi Buder-Gröndahl
tommi.grondahl@helsinki.fi

[1] Department of Digital Humanities, University of Helsinki, Yliopistonkatu 3, 00014 Helsinki, Finland

of abstract *structural descriptions* (SDs) derived from theoretical linguistics, such as hierarchical phrase-structures.

Despite frequent proclamations about the linguistic capacities of LLMs, experimental results have been deemed unclear (Kulmizev & Nivre, 2022). This is partly traceable to technical challenges, such as the influence of superficial heuristics (McCoy et al., 2019), differences between probing methods (Immer et al., 2022), or the impact of labeling formalism (Kulmizev et al., 2020). However, I propose that a major ambiguity can be traced to a more fundamental source: the theoretical notion of "linguistic representation" itself. While related matters have been extensively discussed in linguistic and philosophical literature, their effects on BERTology have so far not been addressed.

In particular, there is a discrepancy in how representations have been analyzed across fields. Customary accounts in contemporary philosophy individuate them by *intentional content*, as determined by e.g. informational or teleological relations (Dretske, 1981; Fodor, 1990; Millikan, 2017; Neander, 2017). In contrast, theoretical linguists commonly adopt a *mentalist* reading where linguistic analyses ultimately concern cognitive architecture itself (Chomsky, 1965, 1986; Fodor, 1981; Laurence, 2003; Smith, 2006; Collins, 2014, 2023; Adger, 2022). This doctrine cuts across most branches of the generative tradition, as well as many other frameworks (e.g. Langacker, 1987; Goldberg, 2006). I focus on generativism in this paper due to the centrality of phrase-structure in BERTology.

A natural interpretation of mentalism is that SDs characterize *vehicles* internal to the cognitive system. However, this still remains ambiguous, as it does not specify their grounds of individuation. Even if linguistics is about representations *qua* vehicles, they could still be individuated by contents. Alternatively, they could be individuated by intrinsic (narrow) *vehicle-properties* that only characterize their "shape" within the system. I suggest that ongoing debates in the philosophy of linguistics can be clarified via these two readings: some authors take SDs to be contents of linguistic representations (e.g. Rey, 2020), while others assimilate SDs to vehicle-properties on a high level of abstraction (e.g. Adger, 2022). My purpose is not to defend either reading as such, or to address various exogetic debates (c.f. Collins & Rey, 2021). Instead, I point to challenges in *both* readings when applied to BERTology.

I argue that the content-reading would essentially mark a return to linguistic structuralism (Bloomfield, 1933; Harris, 1951), since it would require SDs to be recoverable from the input data. This directly contrasts the generative analysis of syntactic phrases having an *autonomous* computational status (Chomsky, 1975; Adger, 2022; Collins, 2023). Revising this conception of SDs would thus also require withdrawing many proclaimed results of BERTology. To avoid this, an initial possibility could be to save autonomous SDs by elevating them to a realm of abstract objects (Katz, 1981; Postal, 2003). However, this would effectively yield indeterminacy between representations of SDs generable via *weakly equivalent* grammars (c.f. Quine, 1970), and thereby undermine the whole premise of BERTology.

The vehicle-reading, in turn, succumbs to the problem of relating abstract SDs to their concrete realizers. To avoid logical category errors with their direct assimilation (Postal, 2009; Behme, 2015), SDs can be treated as *mathematical contents* assigned to vehicles via a separate *interpretation function*—in line with the general explanatory framework laid out by Egan (2010, 2014, 2018). But, as recently observed by Facchin (2022), such mathematical contents are vulnerable to well-known *triviality problems* for mapping-theories of physical computation (Sprevak, 2018). The upshot is that the mere availability of mathematical content cannot ground its explanatory relevance.

Following (Egan, 2017), I maintain that mathematical contents are nevertheless valuable due to their use as *explanans* for generalizations that would otherwise be overlooked. By acting as "proxies" of the underying cognitive states, they allow *surrogative reasoning* (Swoyer, 1991). In this interpretation, the justification of stipulating linguistic representations is based on its explanatory value: is it needed for covering some relevant generalizations that would otherwise be left unaddressed? In effect, this order of explanation is *reversed* in BERTology, where the goal is simply to find *some* mapping from LLM-states to target labels interpreted via SDs. This alone is insufficient to ground the explanatory relevance of those SDs, since indefinitely many other mappings would be available as well. I illustrate this dilemma with Hewitt and Manning's (2019) structural probing algorithm as a concrete case-study.

The problem is not restricted to representationalist interpretations of LLMs: the anti-representationalist should also know what she is denying in the first place. Of course, the notion of "representation" might turn out to be too vague to be theoretically useful, in which case anti-representationalism would be justified by default. But this negative conclusion could only be reached after proper engagement with the representationalist claims. By the same token, I retain agnosticism about the final verdict on LLMs; my present aim is to clarify the conditions for both representationalism and its rejection to be informative hypotheses.

While I do not see any easy way out of the present predicament, my overall contention is that BERTology has been overly reliant on linguistic theories developed for very different purposes than connectionist NLP—often decidedly antagonistic with it. *Prima facie*, it would be surprising if these happened to coincide in some deep way. I propose that their seeming convergence is instead an artifact of meta-theoretical interpretation via pre-determined SDs. The central question itself—*why choose these SDs and not others*—has remained insufficiently addressed.

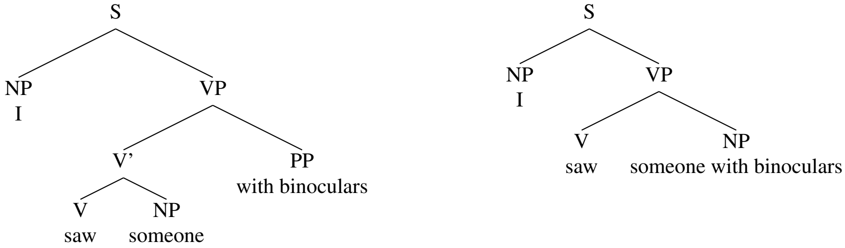
As an alternative methodology for future work, I suggest that BERTology should be approached more bottom-up, with the goal of formulating appropriate SDs for capturing the LLM-pipeline in an explanatorily robust human-readable manner. The *transformer* architecture used in LLMs (Vaswani et al., 2017) has been shown to be amendable for a high-level computational analysis captured in a symbolic meta-language (Weiss et al., 2021). While this research still only pertains to simpler models than LLMs, it points to a promising direction of comparing high-level analyses based on independently grounded links to lower-level algorithms. This is still not guaranteed to rule out all indeterminacy between the higher and lower levels, but at least it would base the selection of the former on the latter.

The paper is organized as follows. In Sect. 2 I argue that BERTology is committed to linguistic representations in a strong sense that deviates from traditional connectionist NLP. This motivates further investigation of how such a committment should be interpreted. Section 3 delineates the vehicle-content distinction in computational systems, and uses it to ground different readings of "linguistic representation". Sections 4–5 introduce three versions of both the content- and vehicle-reading, and demonstrate their challenges for BERTology with a focus on structural probing. Section 6 further elaborates on the practical ramifications of these problems. Section 7 summarizes the paper and discusses prospects for future work.

2 BERTology: the return of representations to NLP

Deep neural networks (DNNs) have traditionally been contrasted with representations of abstract SDs such as phrase-structures. For present purposes, it suffices that a phrase is a hierarchical object that *dominates* its immediate constituents and all of their constituents. This can explain why expressions like *I saw someone with binoculars* are ambiguous, as shown in (1):

(1)



DNNs are not pre-programmed to represent such SDs. This invites two alternative interpretations of how they can attain linguistic performance: *eliminative connectionism* and *implementational connectionism* (Marcus, 1998). The first of these would discard abstract representations altogether. In their seminar paper on learning the English past tense, Rumelhart and McClelland (1986) write:

> (...) a reasonable account of the acquisition of past tense can be provided without recourse (...) to the notion of a "rule" as anything more than a *description* of the language. (...) The child need not figure out what the rules are, nor even that there are rules.
> (Rumelhart and McClelland 1986, p. 267; emphasis in the original)

That is, inputs, outputs, or their relations being described in a certain way does not justify inferring that the model represents them in that way. Rumelhart and McClelland's *anti-representationalist* position reduces the role of SDs to the description of the data or task, leaving them out when explaining model-internal computation.

In contrast, implementational connectionism takes a DNN's sub-symbolic states and their transitions to realize rules and representations on a high level of description. This is succinctly put by Pinker and Price (1988):

> PDP[1] models would occupy an intermediate level between symbol processing and neural hardware: they would characterize the elementary information processes provided by neural networks that serve as the building blocks of rules or algorithms. Individual PDP networks would compute the primitive symbol

[1] "PDP" stands for *parallel distributed processing*, i.e. connectionist architectures.

> associations (such as matching an input against memory, or pairing the input and output of a rule), but the way the overall output of one network feeds into the input of another would be isomorphic to the structure of the symbol manipulations captured in the statements of rules.
> (Pinker and Price 1988, p. 76)

For building NLP applications, the distinction between eliminative and implementational connectionism is not immediately relevant. As long as models work in end-to-end settings, agnosticism can be maintained about the representations involved. If no manually programmed rules are needed, linguistic questions can be set aside. This attitude is captured in the famous quip attributed to Frederic Jelinek: "Whenever I fire a linguist our system performance improves" (for elaboration, see Jelinex, 2005).

However, increasing *model explainability* has become a central goal in recent years (Danilevsky et al., 2020). This is motivated both by the scientific aim of understanding DNNs better, as well as practical concerns such as model biases (Nadeem et al., 2020) or adversarial data (Li et al., 2020). Explaining the structure of DNNs in human-understandable terms can no longer remain agnostic about model-internal representations, since those are precisely what it aims to uncover.

Contemporary NLP is built around *pre-trained LLMs*, such as BERT (Devlin et al., 2019) or GPT (OpenAI, 2023). These are trained on massive datasets for generic linguistic tasks such as predicting masked tokens (for BERT) or upcoming text (for GPT), and can then be *fine-tuned* for domain-specific tasks. Pre-training is thus aimed to give them generic linguistic competence on which to build in subsequent tasks.

Interpreting pre-trained LLMs has reached such a central status that "BERTology" is now recognized as a dedicated subfield of NLP (Rogers et al., 2020). Notably, it makes heavy use of linguistic theory. BERT has been suggested to represent phrase-structures (Coenen et al., 2019), dependency relations (Jawahar et al., 2019), semantic roles (Kovaleva et al., 2019), constructions (Madabushi et al., 2020), and lexical semantics (Soler & Apidianaki, 2020), among others. While LLMs still do not use explicitly programmed representations, BERTology aims to *find* representations in them (see Fig. 1).

In stark contrast to Rumelhart and McClelland's (1986, p. 267) contention that the language-learner "need not figure out what the rules are, nor even that there are rules", BERTology is thus founded on the premise that LLMs *do* in fact figure out what the rules are, and achieve this via internal representations. Table 1 collects representative quotes from literature, which further display this dedication.

The leading model-interpretation technique is *structural probing*, where a classifier ("probe") is trained to map LLM-states to linguistic target labels. For example, Hewitt and Manning (2019) use matrix B for enacting the linear transformation in equation (2), where $\mathbf{h}_i$ and $\mathbf{h}_j$ are encodings for the i:th and j:th word in the input sentence:

(2) $d_B(\mathbf{h}_i, \mathbf{h}_j) = (B(\mathbf{h}_i - \mathbf{h}_j))^T(B(\mathbf{h}_i - \mathbf{h}_j))$

The metric $d_B(\mathbf{h}_i, \mathbf{h}_j)$ is trained to recreate the hierarchical distance between each word pair in the input's parse tree (obtained from a pre-existing treebank). This is a representative example of BERTology and its connection to linguistic theory. My purpose here is not to evaluate this particular probe in further technical detail or

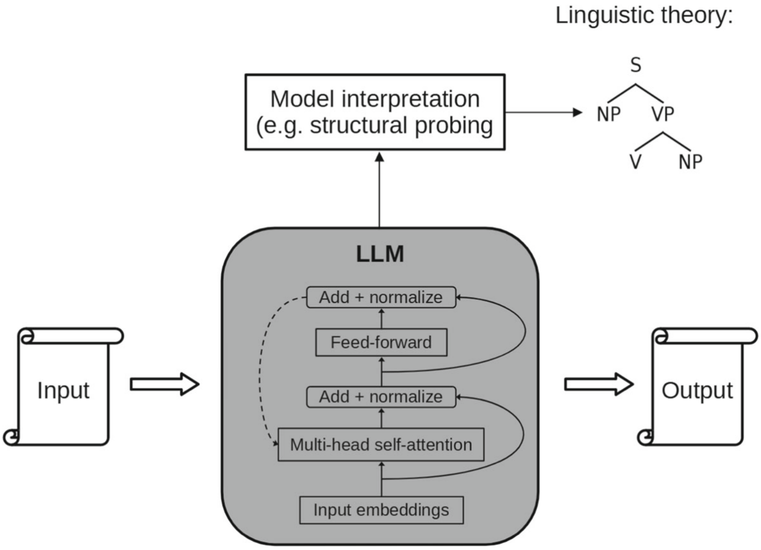


Fig. 1 Linguistically driven model interpretation in BERTology

compare it to other contenders (c.f. Immer et al., 2022). Rather, I will use it as a concrete case-study for illustrative purposes.

Despite the prominence of BERTology, no consensus has been reached on the linguistic representations present in LLMs. On the one hand, impressive model performance has been attained across many linguistic tasks (McCoy et al., 2020; Lasri et al., 2022), and structural probing has uncovered systematic correlations between DNN layers and linguistic classes (Jawahar et al., 2019; Tenney et al., 2019; Immer et al., 2022). On the other hand, models often rely on *superficial heuristics* (McCoy et al., 2019). As Rogers et al. (2020, p. 854) put it: “if there is a shortcut in the data, we have no reason to expect BERT to not learn it”. It remains unclear how to draw the line between genuine linguistic representations and mere complex heuristics.

Pater (2019, p. e61–e62) observes that the interpretation of experiments often hinges on theoretical assumptions: DNNs’ partial success illustrates their strength to some, while their partial failure illustrates their deficiencies to others. This is problematic, given that the purpose of BERTology is precisely to evaluate underlying theoretical hypotheses empirically. If interpreting results hinges on the prior acceptance of some hypotheses over others, then the results do not genuinely help decide between them. The predicament is concisely summarized by Kulmizev and Nivre (2022):

> (...) hypotheses, methodologies, and conclusions comprise many conflicting insights, giving rise to a paradoxical picture reminiscent of Schrödinger’s cat – where syntax appears to be simultaneously dead and alive inside the black box models.
> (Kulmizev & Nivre 2022, p. 02)

Table 1 Quotes from BERTology literature (bolded parts show dedication to representations)

“Our goal is to design a simple method for testing whether **a neural network embeds each sentence’s dependency parse tree** in its contextual word representations—a structural hypothesis.” (Hewitt & Manning, 2019, pp.4129–4130)
“Investigating how **BERT represents syntax**, we describe evidence that attention matrices **contain grammatical representations.**” (Coenen et al., 2019, p. 8592)
“In this work, we investigate the **linguistic structure implicitly learned by BERT’s representations.**” (Jawahar et al., 2019, p. 3652)
“Another theme that emerges in several studies is the **hierarchical nature of the learned representations.**” (Belinkov & Glass, 2019, p. 52)
“We propose a methodology and offer the first detailed analysis of BERT’s capacity to **capture different kinds of linguistic information by encoding it in its self-attention weights.**” (Kovaleva et al., 2019, p. 4365)
“We find that **the model represents the steps of the traditional NLP pipeline in an interpretable and localizable way,** and that the regions responsible for each step appear in the expected sequence: POS tagging, parsing, NER, semantic roles, then coreference.” (Tenney et al., 2019, p. 4593)
“Neural networks can and do improve on this task by **inducing their own representations of sentence structure** which capture many of the notions of linguistics” (Manning et al., 2020, p. 30047)
“(...) **representation of higher-level semantic phenomena** follows the **encoding of syntax and predicate semantics.**” (Kuznetsov & Gurevych, 2020, p. 177)
“To correctly predict the number of the verb, the **DNN must derive an implicit analysis of the structure of the sentence**” (Linzen & Baroni, 2021, p. 198)
“In our experiments, we focus on answering two questions: (i) How is number information **encoded in BERT’s representations?** and (ii) How is number information **transferred from a noun to its head verb** for the model to use it on the behavioral task?” (Lasri et al., 2022, p. 8822)
“(...) models are largely able to capture long-range syntactic dependencies that **require hierarchical representations of sentences.**” (Mueller et al., 2022, p. 1352)
“Due to their strong performance on many language-based tasks that require some linguistic understanding, it is natural to hypothesize that the **models must implicitly encode some linguistic knowledge.**” (Li et al., 2022, p. 1144)

Kulmizev and Nivre draw attention to a number of technical aspects that should be better taken into account, and it is easy to agree with this call for further clarification.[2] However, I take a further step in proposing that a major ambiguity can be traced to an even more fundamental source: the notion of *linguistic representation* itself.

[2] They emphasize four venues for improvement. First, linguistic properties need to be distinguished from *coding properties* from which they can be inferred. Second, linguistic data should be differentiated from its theoretical interpretation, as illustrated by the influence of labeling frameworks on probing (Kulmizev et al., 2020). Third, it is important to tease apart the effect of each variable involved in the choice of model architecture, hyperparameters, training protocol, data, and the end-to-end task. Finally, research questions need to be clarified based on whether they concern what a model *does* learn in a particular experiment setting, what it *can* learn in principle, or what it *must* learn under certain conditions. Especially the first two considerations will recur in my discussion as well.

3 The ambiguity of "linguistic representation"

The *mentalist* view that language arises from cognition is ubiquitous in linguistic theory, especially in the generative framework (Chomsky, 1965, 1986; Gleitman, 2021) but also elsewhere, as in cognitive grammar (Langacker, 1987) and construction grammar (Goldberg, 2006). I focus on generativism due to the centrality of phrase-structure in BERTology. Despite the formal nature of Chomsky's initial linguistic work, the mentalist commitment is already indicated in his early writings:

> A language *L* is understood to be a set (in general infinite) of finite strings of symbols drawn from a finite "alphabet." (...)
> A grammar of *L* is a system of rules that specifies the set of sentences of *L* and assigns to each sentence a structural description. (...)
> It is appropriate, in my opinion, to regard the grammar of *L* as a representation of fundamental aspects of the knowledge of *L* possessed by the speaker-hearer who has mastered *L*.
> (Chomsky, 1975, p. 5)

The quote above recognizes four kinds of entities: (i) *languages* in the sense of formal language theory (i.e. sets of strings); (ii) SDs such as phrase-structures; (iii) *grammars* as abstract generative systems; and (iv) *knowledge of language* as a cognitive state. This taxonomy further grounds the distinction between *weak* and *strong generative capacity*, where the former concerns the ability of a grammar to generate strings, and the latter concerns its ability to generate SDs (Chomsky, 1980; Miller, 1999). All weakly equivalent grammars generate the same expressions, but generative theory concerns strong generative capacity instead (Ott, 2017). A central meta-theoretical question is, thus: what is the relation between concrete linguistic expressions, abstract SDs, and concrete cognitive systems?

Linguistic representations provide the crucial link for connecting SDs to cognitive states. The basic idea is that SDs receive their cognitive relevance by being represented in the cognitive system. Thereby, SDs can be used to *individuate representations*: e.g. NP and VP are SDs that have cognitive relevance by virtue of the system containing NP- and VP-representations.

However, "representation" is polysemous. Especially in the AI-literature, it is often treated as roughly synonymous with "model-internal state". For example, states that arise from word-inputs can be called "word representations" (e.g. Pennington et al., 2014). This needs to be clearly distinguished from more theoretically committing interpretations adopted in philosophy, linguistics, or cognitive science.

A stricter notion of representation can be clarified with the *vehicle-content distinction* (Dennett, 1991; Millikan, 1993). Vehicles are objects operated on by a cognitive/computational system by virtue of their intrinsic properties and interrelations (e.g. Piccinini, 2015). Contents are semantic interpretations assigned to vehicles—i.e. properties/entities to which they *refer*. In veridical representation, the stimulus that triggers the vehicle also realizes its content. In misrepresentation, the content and stimulus diverge. In addition, vehicles have *vehicle-properties* that only characterize their intrinsic "shape" within the system. Figure 2 displays this overall schema.

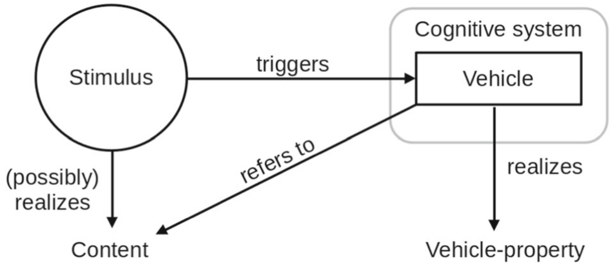


Fig. 2 Vehicle-content distinction in cognitive systems

From now on, I take (stipulated) linguistic representations to be *linguistically individuated vehicles* in the system under investigation. SDs play the part of individuating such vehicles: for instance, a VP-representation is a vehicle belonging to the equivalence class identified by the SD "VP". Linguistic mentalism hinges on the presence of such SD-individuated representations in human cognition, and BERTology extends this commitment to LLMs.[3]

But dedication to linguistic representations (*qua* vehicles) does not yet tell us how their *linguistic status* is grounded. One option is to classify vehicles by content. An example of this is 'words that denote a type of fish', which individuates a class of words (i.e. vehicles) by their semantic interpretation (i.e. content). Alternatively, vehicles can also be classified by vehicle-properties, such as 'capitalized' for written words. Consequently, treating linguistic representations as vehicles individuated by SDs is ambiguous with respect to whether this individuation is based on content or vehicle-properties. This yields two distinct readings of "linguistic representation":

- **Content-reading:** vehicles individuated by SDs as contents
- **Vehicle-reading:** vehicles individuated by SDs as vehicle-properties

The content-reading takes SDs to characterize contents of cognitive vehicles, which in turn makes those vehicles linguistic representations. In contrast, the vehicle-reading takes vehicles themselves to realize intrinsic properties characterized by SDs. My purpose is not to defend or oppose either reading as such. Instead, I raise problems for *both* when applying them to BERTology. Section 4 discusses the content-reading and Sect. 5 the vehicle-reading.

4 Content-reading

I identify three candidates for the content-reading: *directly referential* (Sect. 4.1), *fictionalist* (Sect. 4.2), and *Platonist* (Sect. 4.3). Each has trouble uniting BERTology with a foundational generative notion: the *autonomy of levels*.

[3] Conversely, I interpret *anti-representationalism* about language as rejecting the presence of linguistically identifiable cognitive vehicles (see Sect. 6 for further discussion).

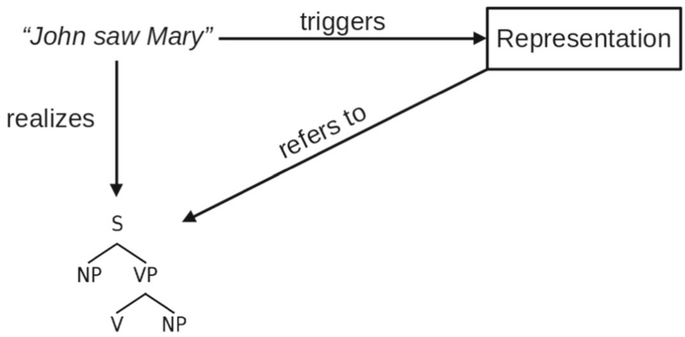


Fig. 3 Directly referential content-reading

4.1 Directly referential reading

An initial option would be to treat SDs as properties of the input data itself, which the model ($\mathcal{M}$) somehow picks out. This general idea (not applied to language specifically) is pursued by Cappelen and Dever (2021), who propose a causal theory of reference for DNNs following a Kripkean line of semantic analysis (Kripke, 1980). Figure 3 summarizes this *directly referential* content-reading.

Here, the SD is realized by the input, which allows grounding the reference relation in the causal pipeline between it and $\mathcal{M}$. For example, according to *informational semantics*, a representation refers to the entity/property that it carries information of (Dretske, 1981; Fodor, 1990). *Teleosematics* adds the requirement that such information-carrying must be the proper function of the representation, based on natural selection in biological systems or design in artificial systems (Millikan, 2017; Neander, 2017). Such naturalistic theories of reference are united by their reliance on causal relations between the representation and its content.

This also suggests a simple and *prima facie* appealing interpretation of BERTology techniques like structural probing: they aim to find parts of $\mathcal{M}$ that refer to SDs as per some causal theory of reference. Applied to Hewitt and Manning's (2019) probe (see Sect. 2): the parse tree distance between tokens w_i and w_j is already present in the input, and their respective encodings $\mathbf{h}_i$ and $\mathbf{h}_j$ refer to this information in a way that is reliably captured by the metric $d_B(\mathbf{h}_i, \mathbf{h}_j)$. This way, $d_B(\mathbf{h}_i, \mathbf{h}_j)$ picks out those aspects of $\mathbf{h}_i$ and $\mathbf{h}_j$ that refer to phrase-structural properties in the input.

The referential interpretation thus requires SDs to be directly present in the input. Consequently, they should be definable via linear information. But now a problem arises: phrase-structures as characterized in generative syntax are precisely *not* definable in this way. In a recent paper outlining this perspective, Collins (2023, p. 7) takes the following principle to be a "a core aspect any theory must respect":

> syntax determines units of combined lexical items that are *not identifiable or individuated in terms of linear order or any other perceptible property* associated with morphophonemic form.
> (Collins 2023, p. 7; my emphases)

The pre-generative structuralist tradition defined categories *distributionally*: two utterances belong to the same class if they have the same distribution with respect to other (similarly defined) classes (Bloomfield, 1933; Harris, 1951). Built around phonology, structuralism treated morphemes as phoneme sequences and syntactic phrases as morpheme sequences, each stipulated for allowing more concise distributional generalizations (Harris, 1951, p. 151). A central motivation for early generativism was the rejection of this approach for syntax. In stark contrast to it, Chomsky (1957, 1965, 1975) elevated each level of grammar (phonology, morphology, syntax, etc.) to an *autonomous* status, where higher levels were not definable via elements on lower levels. Instead, each level had its own vocabulary of computational primitives to be combined, and levels were linked via additional mapping rules. This *autonomy of levels* has major repercussions on linguistic representations, as explained by Adger (2022):

> (...) the perspective in [Chomsky (1975)] is top-down rather than bottom up. Each level is an independently specified concatenation algebra consisting of a set of primes (symbols) and relations at that level, and the algebra specifies certain strings of symbols as well-formed, so the *"representations" are not derived from the utterance.* Rather, each string of symbols at one level can be converted into a lower level of structure through a specified set of mappings
> (Adger, 2022, p. 251; my emphases)

In particular, phrase-structures are not derived from linear concatenations of units (words, morphemes, phonemes, graphemes, etc.). Syntactic information is fundamentally novel in kind by exhibiting non-linear hierarchical relations, which requires a wholly different set of computational units and operations.[4] As (Katz 1981, p. 38) recounts: "the potential for highly abstract generative grammars could be realized only if a new and far less concrete interpretation of grammars was found".

Of course, the autonomy of levels could be rejected. I am not assessing its merits as such, nor do I presume that it must be correct (see Sect. 6 for discussion of alternative frameworks). Instead, I examine a *conditional* question: given that BERTology literature contains numerous claims about LLMs representing abstract SDs as defined in generative linguistics, how *could* this be possible in the first place? If SDs are autonomous with respect to the linear input and thus not present in it, they cannot be represented via causal relations as required by naturalistic theories of reference.

In fact, the problem can already be predicted based on a viable candidate for information processing in DNNs. Buckner (2018) proposes that they enact *transformational abstraction*, where higher layers discard, combine, and alter features from lower layers. The basic idea is shown in Fig. 4.

Transformational abstraction results in increasing levels of *transformational invariance*: the ability to detect features that remain stable across other changes. This yields

[4] As an anonymous reviewer correctly notes, generative theory has commonly treated phonological and semantic features as belonging to lexical items and thereby to the syntactic derivation. Interface representations that map syntax to phonology ("PF") and semantics ("LF") have also played a vital role. Such interactions between levels somewhat complicate proclamations of their autonomy. That notwithstanding, it remains a foundational principle of generative syntax that syntactic phrase-structure is not grounded in linear concatenations of words. This is the crux of my present argument.

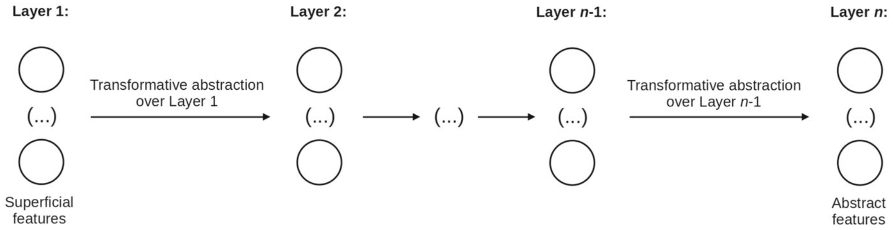

Layer 1:
Layer 2:
Layer n-1:
Layer n:
Transformative abstraction over Layer 1
Transformative abstraction over Layer n-1
(...)
Superficial features
Abstract features

Fig. 4 Transformative abstraction in DNNs

the potential to represent many abstract properties. As Buckner (2018) discusses, the concept 'triangle' was problematic for classical empiricism because it seems that sensory information cannot correspond to triangularity in general; only specific triangles. However, higher layers of an image-detector DNN could become sensitive to just those properties that make up triangular inputs, by combining and removing lower-level information from prior layers. These properties are still fully based on low-level sensory information, derivable from it via transformative abstraction.

Likewise, Nefdt (2023, pp. 92–95) proposes that DNNs can extract abstract linguistic structures from data. Crucially, this idea relies on the assumption that the structures are *in the data* to begin with:

> What neural networks are especially good at is picking up *patterns hidden in complex sets of data.* (...) The result is a hyper-empiricist framework for capturing the *real patterns of complex systems in reality.*
> (Nefdt, 2023, p. 93; my emphases)

If this is indeed how DNNs work, they might well represent something like high-level distributional properties as originally envisioned in the structuralist paradigm. But the generative enterprise was specifically founded upon the *rejection* of such distributional approaches to SDs. In the directly referential reading, assigning representations of autonomous SDs to LLMs would therefore become *trivially false*. This invites an alternative interpretation of the relation between inputs and SDs.

4.2 Fictionalist reading

In an original and innovative account, Rey (2020) aims to combine the content-based individuation of linguistic representations with the lack of SDs having real existence in the input.[5] He proposes that linguistic contents are *intentional inexistents*, adopting the term from Brentano (1874/1911).[6] Linguistic representation is thus *misrepresentation*, as depicted in Fig. 5.

It is helpful to consider an analogy from the philosophy of perception. A much-discussed example of a misrepresentation is seeing a stick as bent in water. This

[5] Rey calls his account "folieism", but I adopt the more familiar term "fictionalism" for convenience.

[6] In spite of terminology, Rey's account of content departs from Brentano's, which is non-naturalist in treating intentional objects as *sui generis* entities. Whatever its general merits, applying this view to LLMs is immediately problematic for similar reasons as the direct "grasping" of Platonic contents (Sect. 4.3).

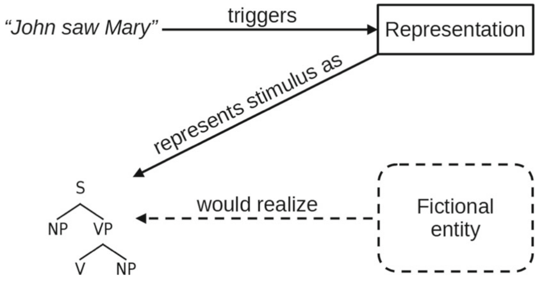


Fig. 5 Fictionalist content-reading

constitutes a problem for simple versions of direct realism, since the stick looks bent without actually being bent. The representationalist solution is to say that the stick is mistakenly represented as bent, where 'bent' acts as the represented content without being realized by the actual stick that triggers the representation (Jackson, 1977).

By the same token, Rey's account of SDs—or, more generally, *standard linguistic entities* (SLEs)—takes them not to exist in the actual input to perception. Instead, human cognition contains representations that have SLEs as their intentional contents. These representations are activated via a process that matches perceptual data to the most appropriate representation based on some kind of hypothesis testing procedure, such as Bayesian inference (Rey, 2020, pp. 373–377). This account is motivated by the classical approach to generative phonology outlined in Chomsky and Halle (1968), where phonological representations have idealized phonetic content.[7] For example, a phoneme such as /p/ is a phonological representation, the intentional content of which is a set of ideal acoustic and/or articulatory properties. Representing a piece of acoustic data as /p/ involves testing different hypotheses with respect to how well this data would be predicted via different phonological representations, and reaching /p/ as the most appropriate candidate. Crucially, the data rarely (if ever) satisfies the criteria of actually instantiating /p/, which would require unrealistically ideal circumstances. This is why the representation is strictly a *mis*representation.

As an analogy to SLEs, Rey raises geometrical concepts such as 'cube', which lack actual physical manifestations. The basic idea is that while no real cubes are present in the environment, our minds are nevertheless primed to see approximately cubic objects *as* cubes; i.e. to apply a representation with the intentional content 'cube'. For Rey, SLEs are similar ideal entities/properties that function only as intentional contents, being (at least mostly) absent from the data itself.

Rey's account might well be suitable for phonological representation, although some challenges may arise with further details.[8] However, phonology has no direct relevance for interpreting LLMs, which take readily individuated orthographic words as inputs.[9] Unlike ideal phonetic contents, orthographic words are straight-forwardly

[7] Chomsky and Halle (1968, p. 65) maintain that a phonological representation "can be interpreted as a set of instructions to the physical articulatory system, or as a refined level of perceptual representation".

[8] For instance, it is unclear how the approach would fare with the possibility of *substance-free phonology* where phonological features lack acoustic or articulatory content (Blaho, 2007; Odden, 2013; Iosad, 2017).

[9] Strictly speaking, LLMs use "subword" tokens, which reduce the vocabulary size while increasing coverage (Sennrich et al., 2016) This technical detail has no impact on the present discussion.

present in the data itself. Instead, problems arise when extending the analysis to abstract SDs such as syntactic phrase-structures.

Treating SLEs as fictitious ideal properties does not yet remove the problem raised in Sect. 4.1: the autonomy of levels makes SDs undefinable via lower-level linear information (Adger, 2022). Linear strings do not yield hierarchical phrase-structure *even in idealized contexts* (Collins, 2023, p. 110). Therefore, some further strategy is required for obtaining their representations. For this, Rey advocates *Ramsification*, where each theoretical predicate is replaced with an existentially quantified second-order variable such that the properties they designate become defined by how the theory relates them to each other and to observable stimuli (Lewis, 1970).[10] SDs would thus be those properties that serve appropriate roles in the overall linguistic theory in relation to other linguistic properties, relevant cognitive processes (e.g. parsing), and relevant observational data (e.g. grammaticality judgements).

But since the theoretical roles of stipulated SDs arise from their putative computational roles in the cognitive system, it is unclear if Ramsification can yield a "content" over and above vehicle-properties. This problem is noted by Dupre (2022):

> The inferentialist proposal says: when a psychological type is treated in these sorts of ways by psychological processes, it represents. But the representational story then just seems like a third wheel. Nothing is gained by the stipulation that such-and-such computational system is, purely in virtue of these computational properties, also a representational system. All the causal and explanatory work is done by the computational story.
> (Dupre, 2022)

As a possible rejoinder, the intentional status of SD-representations could be traced to their relations to phonological (and perhaps also semantic) representations, which in turn are individuated by content. Their theoretical roles (recognized in Ramsification) would thus be at least indirectly content-laden. But the gist of Dupre's point still stands: relations between SDs and other linguistic representations are exhausted by "the computational story": i.e. vehicle-properties and relations between vehicles. Without a prior understanding of these, Ramsification cannot get off the ground.

4.3 Platonist reading

A possible candidate for avoiding the problems I have raised could be to treat linguistic contents as abstract, Platonic objects (Katz, 1981; Postal, 2003). This allows the autonomy of levels, since abstract linguistic objects on different levels do not need to be mutually constitutive. However, it remains unclear how these would be fixed as contents of linguistic representations, especially in LLMs.

Katz(1981, pp. 193–200) ends up proposing a special "faculty of intuition" to account for how we can be in *a priori* contact with the abstract realm of SDs. While this assumption is already controversial about human cognition (Benacerraf, 1973), it does not even get off the ground with LLMs, which are straight-forwardly mechanistic systems. Discarding such direct epistemic "grasping", the remaining option seems to

[10] I thank an anonymous reviewer for raising this point.

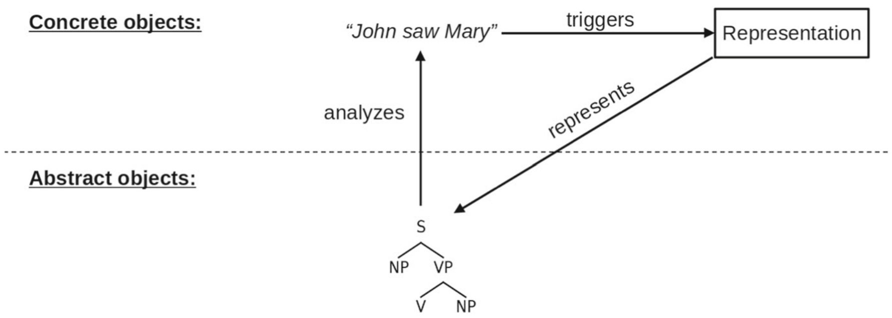


Fig. 6 Platonist content-reading

Table 2 Summary of content-readings

Reading	Problem for BERTology
Directly referential	Autonomy of levels
Fictionalist	Autonomy of levels
Platonist	Quinean indeterminacy between SDs

be that abstract SDs constitute contents of linguistic representations when they can *analyze* the input. This is illustrated in Fig. 6.

However, all weakly equivalent grammars can analyze the same expressions even if they strongly generate distinct SDs (see Sect. 3). This interpretation would thus result in *Quinean indeterminacy* about linguistic representation (Quine, 1970): there is no fact of the matter which of the (mutually incompatible) SDs that could analyze the input is the content. The problem clearly arises from the fact that the abstract SDs are limited to only analyzing the input, discarding the cognitive system itself. In short, unless the system's *internal structure* is considered, Quinean indeterminacy looms. This indicates that the vehicle-reading might fare better.

4.4 Summary

To recap, neither real nor fictional physical properties of the linear input constitute viable candidates for SDs as contents of linguistic representations in LLMs. They cannot theoretically ground BERTology without sacrificing the autonomy of levels assumed in generative linguistics. This threatens to make LLM-interpretations that rely on autonomous SDs *trivially false*. If a Platonic conception of SDs is adopted instead, representation-claims are again in danger of becoming trivial, this time due to Quinean indeterminacy. Table 2 summarizes the three readings and their problems.

5 Vehicle-reading

Despite the prevalence of the content-reading in philosophical literature, its specific troubles with linguistic representations invite considering the vehicle-reading instead.[11] I identify three candidates for what the relation between vehicles and SDs could be: *identity* (Sect. 5.1), *direct realization* (Sect. 5.2), or *indirect realization* (Sect. 5.3). After rejecting the first two, I observe a challenge in the last, once again related to triviality concerns.

5.1 Identity reading

As mentioned in Sect. 3, the word "representation" is sometimes used for any internal states. While this technically fits the vehicle-reading, it is obviously unfalsifiable and hence uninformative. The vehicle-properties for individuating linguistic representations should be *non-trivial*.

5.2 Direct realizational reading

A more substantive idea would be that linguistic explanations are *descriptive abstractions* over cognitive states and processes. Descriptive abstraction is a theoretical process of attaining a high-level analysis by omission of information (Boone & Piccinini, 2016; Kuokkanen, 2022). This seems to be what Adger (2022) is after in assimilating a mental representation of a grammar to a brain-state:

> "A mental representation of the grammar of the language" is just the mental structure (brain state) which is, at the relevant level of abstraction from physiological mechanisms, the grammar of the language.
> (Adger, 2022, p. 252)

That is, by performing descriptive abstraction of an appropriate kind, one should reach an analysis of those aspects of cognition that *realize* the grammar—i.e. "represent" it in the vehicle-reading.[12] Fig. 7 shows this realizational interpretation.

Descriptive abstraction is based on the specification of *equivalence classes*, where each member of a class has the same role. Assuming the system to be computational, it transforms vehicles into others based on their intrinsic form and interrelations

[11] A possible initial objection could arise that representations must *ipso facto* have content, since otherwise they would not be "representations" in the first place. For two reasons, I am not worried about such an *a priori* argument. First, the vehicle-reading is agnostic about whether linguistic representations have contents; it only maintains that their *linguistic status* is based on vehicle-properties and not contents (if such exist). Second, terminology is rarely a dependable guide to ontology. The word "atom" was originally used as something *ipso facto* non-decomposable; but was later adopted for decomposable entities in scientific practice. An *a priori* case against splitting atoms based on etymology would evidently not work. By the same token, the vehicle-reading needs to be assessed based on its theoretical and empirical merits, mere conceptual analysis being unreliable.

[12] Chomskyan linguists (such as Adger) would, of course, likely deny that LLMs have similar linguistic representations as humans. My point here is that the schema of descriptive abstraction gives criteria for what *would* be required for linguistic representation in a physical system. Whether they are present in the human brain, LLMs, etc. is a further question that can be evaluated only when the criteria are settled.

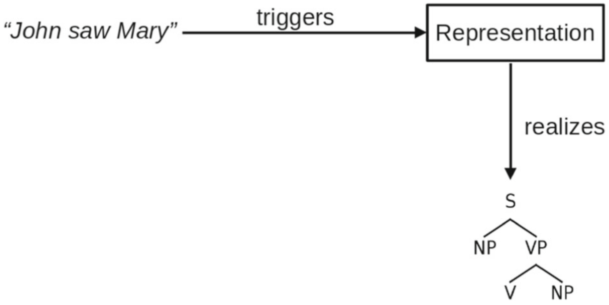


Fig. 7 Realizational interpretation of the vehicle-reading

(Piccinini, 2015). Hence, the equivalence classes should be determined by some combination of (i) intrinsic properties of vehicles, (ii) relations between vehicles, and (iii) input–output transformations between states determined by the distribution of vehicles. While some of these are relational, they still remain fully system-internal and independent of content. They are thus vehicle-properties in the present sense.

Notably, the realizational vehicle-reading can incorporate the autonomy of linguistic levels. If SDs individuate equivalence-classes of computational vehicles, their autonomy only requires that they do not bear constitutive relations to each other and are instead related via separate links. This has a straight-forward correlate in DNNs: *layers*. Even if the information generalized between layers is obtained via transformational abstraction over information encoded in previous layers (see Sect. 4), the computational vehicles themselves (i.e. nodes of the DNN) still remain separate. The autonomy of levels could thus be attained by assigning representations of different levels to dedicated layers—as has indeed been argued on empirical grounds in BERTology (e.g. Hewitt & Manning, 2019; Tenney et al., 2019; Manning et al., 2020).

This reading also supports a straight-forward interpretation of structural probing: it finds equivalence-classes of LLM-states based on the probing task. States that the probe classifies in the same way constitute an equivalence-class, which is further assimilated to the linguistic category assigned to the probe's target label. For example, the equivalence-class of states from which a probe outputs the target "NP" would *be* the model's NP-representation. By the same token, Hewitt and Manning's (2019) probe captures an equivalence-class of distance relations between two encodings, and this simply *is* the representation of parse tree distance in BERT.

However, there are fundamental differences between lower-level realizing structures and higher-level abstract structures, which forbids their direct assimilation. As a prominent example, consider the set-theoretical definition of the operation *Merge* in the *minimalist* variant of generative theory. Merge is an operation that puts two syntactic objects (words or phrases) together, resulting in a complex phrase with two constituents. Chomsky (1995) further maintains that a set-theoretic formulation should be adopted as its simplest possible formalization. This yields axiom (3):

(3) Merge(A, B) = {A, B}

At the same time, such analyses are intended to explain human linguistic cognition. But clearly (3) does not directly denote a brain-state even on a high level of description: it is an abstract set. On behalf of linguistic Platonism (see Sect. 4.3), it has been asserted that such discrepancies between abstract SDs and concrete brain-states make linguistic mentalism incoherent (Katz, 1981; Postal, 2003, 2009; Behme, 2015). However, as Levine (2018, p. 53) observes, Chomsky has also acknowledged this and explicitly denied their direct assimilation:

> We don't have sets in our heads. So you have to know that when we develop a theory about our thinking, about our computation, internal processing and so on in terms of sets, that it's going have to be translated into some terms that are neurologically realizable.
> (Chomsky 2012, p. 91; also cited in Levine 2018, p. 53)

That is, if Merge is defined set-theoretically, understanding its physical realization would first require translating it into something else—presumably a genuine descriptive abstraction over brain-states. The idea that linguistic theories are *directly* interpretable as descriptive abstractions of brain-states would indeed be incoherent, and is not Chomsky's position either. Instead, the mentalist contention is that linguistic formalisms can somehow aptly capture relevant properties of the cognitive system under discussion.[13] This amounts to an *indirect* reading, as covered in Sect. 5.3.

Initially, it might seem that the situation is different for LLMs, which are not specified in physical terms (akin to brain-states) but are already mathematically individuated in the DNN pipeline (via vectors, matrices, and operations across them). Nevertheless, a comparable problem arises here as well. In Marr's (1982) taxonomy, linguistic theory belongs to the highest *computational* level; whereas the specification of LLMs belongs to the *algorithmic* level that spells out explicit steps for realizing a computation but abstracts away from the lowest *implementational* level describing the physical substrate. The challenge for linguistic mentalism concerns linking the computational and implementational levels. LLM-interpretation, in turn, aims to link the computational level to lower-level algorithms. Similar problems arise for both, given distinct meta-theoretical vocabularies on different levels (c.f. Dunbar, 2019).

In sum, the realizational vehicle-reading would commit a *category error* by directly equivocating abstract SDs with equivalence classes of model-states. This problem has been raised to undermine linguistic mentalism on the whole, but such an accusation goes against Chomsky's own acknowledgement that their relation must be indirect. That being said, it is not trivial how this indirect relation should be construed. For this purpose, I append the meta-theoretical taxonomy with *mathematical contents*.

[13] For example, central properties of sets include non-associativity and the irrelevance of linear order: $\{x, \{y, z\}\} \neq \{\{x, y\}, z\}$ and $\{x, y\} = \{y, x\}$. The set-theoretical formalism thus conveys that syntactic representations behave non-associatively (marking hierarchical distinctions) and do not take linear information into account. The representations are not "sets in the head", but instead behave in a set-like manner in certain computational respects. (I thank an anonymous reviewer for highlighting this point.)

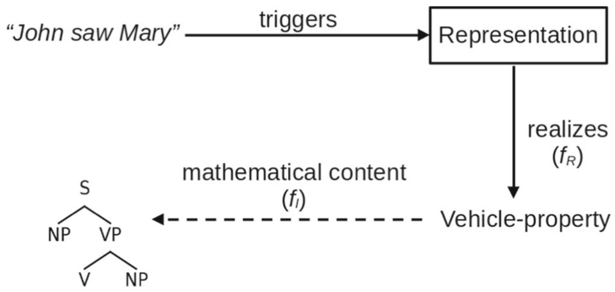


Fig. 8 Indirect realizational vehicle-reading

5.3 Indirect realizational reading

Egan (2010, 2014, 2018) distinguishes between two functions that determine the computation implemented by a physical system. The *realization function* (f_R) maps its states to equivalence classes of vehicles. The *interpretation function* (f_I) maps vehicle-types (determined by f_R) to contents, of which there are two kinds. *Cognitive contents* assimilate to what I have called "contents"—i.e. external referents. *Mathematical contents* are independent of the system's environment, and provide abstract interpretations of vehicles and their relations.[14]

Mathematical contents have a somewhat intermediate position between contents and vehicle-properties in my taxonomy (see Sect. 3). As terminology already suggests, Egan treats them as a kind of content. However, they are "narrow"— i.e. independent of the system's environment—and are thus restricted to explaining vehicle-properties. As an example, Egan (2010) provides Marr's (1982, p. 337) analysis of early vision involving the computation of the Laplacean of a Gaussian. The computation allows the system to detect light intensity changes in its typical ecology; but this is not part of the mathematical content, which only concerns operations internal to the system itself.[15] Nevertheless, it provides an abstract description (rather than e.g. a neural one), and is thus irreducible to mere equivalence classes of vehicles.

Similarly, linguistic SDs could be treated as specifying mathematical contents linked to computational vehicles. Figure 8 displays this *indirect realizational* schema.

Unlike in other readings of "linguistic representation" discussed so far, here the linguistic status of vehicles is *meta-theoretical*. This allows maintaining the irreducible abstractness of SDs without succumbing to the problems of the Platonist content-reading (see Sect. 4.3). Even though mathematical contents are abstract, there is no mysterious "grasping" involved any more than between numbers and calculators, or the visual system and Gaussian functions. Figure 9 clarifies this: only vehicle-types (determined by f_R) directly concern the system, and mathematical contents are mapped to them 1–1 via a separate function (f_I) that only has a theory-internal status.

[14] Egan further maintains that mathematical contents suffice for computational explanation proper. My present discussion is independent of her pragmatism about cognitive contents.

[15] Egan's reading of Marr contrasts others that take intentional content to be essential for computational explanation (e.g. Burge, 1986). My purpose here is not to evaluate it as a Marr-interpretation, but instead to assess its applicability for explaining linguistic representation.

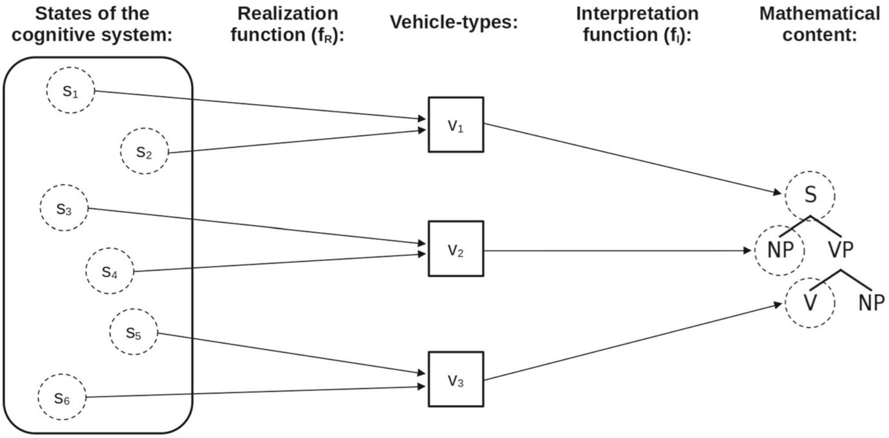

States of the cognitive system:
Realization function (f_R):
Vehicle-types:
Interpretation function (f_I):
Mathematical content:
s_1
s_2
s_3
s_4
s_5
s_6
v_1
v_2
v_3
S
NP
VP
V
NP

Fig. 9 Realization and interpretation functions

However, if linguistic representations are based on separately chosen interpretation functions, their empirical status is once again endangered. As Facchin (2022) shows, mathematical contents are vulnerable to similar *triviality problems* as structural mapping accounts of physical computation. These are based on the possibility of mapping any physical states to any abstract structures, as long as there are at least as many of the former as the latter. This was proven for inputless finite-state automata by Putnam (1988), and has since been broadened to cover all computational systems with finite storage (Sprevak, 2018).[16] Consequently, linguistic representation cannot be secured simply by there being *some* f_R and f_I, the successive application of which to model-states yields SDs. If the model is sufficiently complex and the SDs are restricted to a finite set (e.g. by maximum tree-depth), such functions can *always* be devised.

In BERTology, structural probing can be treated as discovering a feasible f_R based on the probing task: it finds a way to group model-states based on their correlation with the probe's targets. Using again Hewitt and Manning's (2019) probe for illustration, $d_B(\mathbf{h}_i, \mathbf{h}_j)$ is trained to approximate parse-tree distance between w_i and w_j (respectively encoded as $\mathbf{h}_i$ and $\mathbf{h}_j$). This allows grouping encodings based on their effects on d_B. But the resulting equivalence-classes do not wear linguistic interpretations on their sleeves. Suppose d_B is used to group encodings to vehicle-classes $\{v_1, ..., v_n\}$. What determines the correct f_I for these?

Initially, it might seem that f_I can be based on the probe's target labels, which are transparently related to information in the training set (such as pre-determined parses). However, it is crucial not to conflate the probe's interpretation with the model's interpretation. I am not concerned with the interpretation of the probe's targets; this is simply assumed to begin with. My problem is the interpretation of vehicle-classes in the model itself, which has never seen those targets.

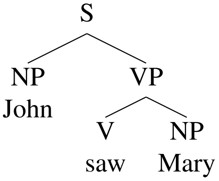
[16] I am targeting a more specific case than typical charges against structural mapping accounts: I am not concerned with restricting the set of computing systems to avoid pancomputationalism, or evaluating whether every physical system computes something (c.f. Chalmers, 1995; Piccinini, 2015). Instead, I aim to restrict mathematical contents in a class of systems already known to compute (LLMs).

Suppose we said that vehicle-class v_i represents linguistic property P if v_i is an equivalence-class discovered by the probe, the target classes of which are interpreted as P. By making the interpretation probe-dependent, this would be backwards with respect to the goal of BERTology: the probe is supposed to find *pre-existing* representations. It is meant to be an experimental tool for discovering vehicle-classes that *already* represent P, not a component for *grounding* the representation itself. To avoid such confusion in explanatory order, the status of the probe should be more modest. But now the triviality problem arises again: if the probe does not fix f_I, what does?

To deal with this problem, we need to look further into the nature of explanation with mathematical contents in cognitive science and linguistics. *Prima facie*, triviality problems should arise here as much as in BERTology. However, I highlight a notable difference in explanatory strategies employed in these discipilines. This also grounds my main critique of BERTology: so far, it has focused on how LLMs *can* be interpreted as opposed to what *must* be included to capture their central properties.

(Mentalist) linguists assert that certain theories of SDs (i.e. abstract grammars) should be favored over others due to their value for explaining observable linguistic capacities. As Egan (2017, p. 155) maintains: “Computational models are proposed to explain our manifest success at some cognitive task”. As a toy example, consider the observation that English-speakers can process the sentence *John saw Mary*. A generative linguist (of an early generation) could analyze this sentence as (4):

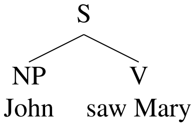
(4)



This should also tell something about the cognitive structures underlying the speakers’ linguistic competence. In the present interpretation, (4) is the mathematical content of a vehicle-class consisting of cognitive states (see Fig. 8). While we do not know what these are, we can still use the placeholder f_R of the function that determines them. They are subsequently mapped by f_I to parts of (4) in some manner that respects its syntactic structure (see Fig. 9). Conversely, the linguist stipulates that some vehicle-classes in English-speakers’ cognition are mappable to (4) in this way.

Consider, now, the dilemma that the same vehicle-types (determined by f_R) could also be mapped to different SDs by another interpretation function f'_I. For example, f'_I could assign *saw Mary* to a single non-decomposable verb, as in (5):

(5)

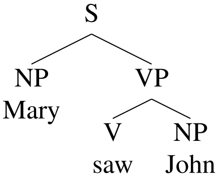


Given that f'_I is not ruled out by any *a priori* principle, the triviality problem arises: why would English-speakers’ linguistic cognition be better analyzed by (4) than (5)? It seems that the relevant factor lies in its superior explanatory power. Typical candidates for rules generating (4) are (6a–b):

(6) a. S → NP VP
b. NP → N
c. VP → V NP
d. N → John | Mary
e. V → saw

This automatically grounds further predictions that can be used to formulate linguistic hypotheses. In particular, another SD (7) is expected to be grammatical:

(7)



On the other hand, (5) can only be generated via rules such as (8), which have no implications for the grammaticality of *Mary saw John*.

(8) a. S → NP V
b. NP → N
c. N → John
d. V → saw Mary

Hence, (4) makes a correct prediction that (5) fails to make. Empirical experimentation on speakers' competence with *Mary saw John* can thereby be used to compare them. Furthermore, this prediction is made fully within abstract linguistic theory, and could not be replicated simply by e.g. looking directly at brain-states. It allows *surrogative reasoning* (Swoyer, 1991) by functioning as an explanatory "proxy" for the underlying cognitive states.[17]

The choice of mathematical content now becomes incorporated into the overall task of theory formation, with customary *desiderata* including empirical scope, lack of incorrect predictions, internal coherence, simplicity, etc. Given the general underdetermination of theory by data, it is expected that multiple candidates may often have equal support at any given time. Theories can also sometimes tap into the same underlying structures in different ways (c.f. Nefdt, 2023, pp. 138–145). But there is nothing special about mathematical contents in either respect: these considerations apply to theory-construction in general, and no detrimental triviality problems arise.

If the only criterion for representing mathematical contents was that *some* interpretation function exists that maps vehicle-types to them, the triviality problem indeed seems insurmountable. However, this does not threaten the actual employment of mathematical contents in linguistics or cognitive science. In fact, the same argument could also target Marr's (1982) account of the computation involved in the human visual system. Suppose that clear equivalence-classes of brain-states were found that mapped to each step in computing the Laplacean of the Gaussian in early vision. Now,

[17] See Matthews (2007) for a partly related approach to propositional attitude ascriptions.

take those classes of brain-states and devise another interpretation function that maps them to some completely different mathematical contents. Is this a threat to Marr's theory of vision? Surely not: Marr's point is not merely that there is *some* mapping from brain-states to computing the Laplacean of a Gaussian; he maintains that this formulation has *special explanatory relevance* for understanding the visual system.

Essentially, the triviality problem arises when mathematical contents are assigned simply because they *can* be. Avoiding it requires investigating which kinds of mathematical contents are actually *needed* to capture central properties of the system. As one final example for further illustration, we can consider Karlsson's (2006) objection to Chomsky's (1957) recursive treatment of language, on the grounds of a corpus study indicating that languages like English, Finnish, and Russian are limited to three clauses in center-embedding. Based on this, he proposes that these languages are "finite-state, type 3 in the Chomsky hierarchy" (Karlsson, 2006, p. 1). Without taking a stance on the dispute as such, I maintain that it must concern the *minimum requirements* for capturing essential properties of language; not simply possible formalisms.

Suppose that Karlsson is correct and language is sufficiently formalizable as a linear, non-hierarchical system. Would this mean that using phrase-strutural SDs as mathematical contents of cognitive states involved in language-processing was *impossible*? No: we could just take those states and devise f_I to map them to such SDs. Clearly, this would not falsify Karlsson's account, which instead concerns what is needed at minimum. If linear SDs suffice for this, the mere possibility of hierarchical SDs is theoretically moot. Conversely, if Chomsky was right instead, using hierarchical SDs would be *necessary*—not merely possible— for linguistic explanation proper.

Likewise, the challenge for BERTology is to show that certain SDs are not only available in principle but needed for capturing non-negligible properties of LLMs left out by other SDs. This would correspond to how cognitive science avoids the triviality problem of mathematical contents: they should manifest some properties that allow uniquely robust surrogative reasoning about the underlying structures and operations. By merely finding some subset of LLM-states that can be mapped to target labels interpreted as SDs, structural probing falls short of this goal: it does not establish that those states (determined by f_R) or those SDs (determined by f_I) have any special explanatory relevance compared to indefinitely many other candidates.

5.4 Summary

Aside of the uninformative identity-reading, the vehicle-reading allows direct or indirect realizational variants. The first commits a category error in directly assimilating distinct levels of analysis. The second is the most promising reading in my estimation, but is challenged by triviality problems familiar from structural mapping theories of physical computation. I argued that these are not detrimental to the use of mathematical contents in cognitive science and linguistics overall, where the choice between alternatives can be grounded in generic considerations of theory-construction. However, they have been insufficiently addressed in BERTology. Instead of merely finding *some* mappings between model-states and SDs, focus should be on evaluating which SDs are actually *needed*. Table 3 summarizes the three vehicle-readings.

Table 3 Summary of vehicle-readings

Reading	Problem for BERTology
Identity	Uninformative
Direct realizational	Category-errors in assimilating levels
Indirect realizational	Triviality problems in mapping

6 Ramifications for BERTology

To recap, BERTology is dedicated to linguistic representations in LLMs, and commonly interprets these via SDs formulated in the generative framework (Sect. 2). My main contention is that such claims are not as theoretically innocuous as they might initially seem. I now turn to some possible ways forward.

Initially, a simple solution would be to abandon the contentious theoretical assumptions of generative linguistics, such as the autonomy of levels. Other linguistic frameworks have rejected these (e.g. Langacker, 1987; Croft, 2001; Goldberg, 2006), and perhaps BERTology should as well. Without denying that this approach might be the right direction overall, I raise three challenges that arise with it.

First, it contradicts claims manifestly made in the experimental literature, as captured in Table 1 (Sect. 2). LLMs have been explicitly analyzed via hierarchical phrase-structures (Belinkov & Glass, 2019; Hewitt & Manning, 2019; Mueller et al., 2022) and as representing “the steps of the traditional NLP pipeline in an interpretable and localizable way” (Tenney et al., 2019, p. 4593). This has been directly contrasted with the idea that an LLM would merely be a “giant associational learning machine” (Manning et al., 2020, p. 30046). In my reading, such claims deserve to be taken seriously at face value. At least, a more metaphorical reading should not be the default starting-point; it could be defended if literal readings turn out to be untenable.

Second, leading non-generative linguistic theories—such as cognitive grammar (Langacker, 1987) and construction grammar (Goldberg, 2006)—heavily rely on semantics as replacing formal syntax in driving linguistic analyses. Moreover, these perspectives are typically connected to *embodied* conceptions of meaning as contiguous with sensory-motor cognition. Since LLMs only use textual information, they lack similar processes. Their semantic representations— if such exist—must either be derived from the input text or arise endogenously from the model’s internal structure (these alternatives corresponding to the content- and vehicle-reading, respectively). Possible future LLMs with additional sensory-motor grounding might well be fruitfully analyzable via embodied approaches; but current ones do not use sensory-motor information. Hence, whatever the overall linguistic merits of embodied semantics-driven frameworks may be, they cannot be directly transported to BERTology at the moment.

Third, the remaining option would be going back to the structuralist notion of SDs as distributional generalizations. This idea has indeed gained much recent attention in NLP (e.g. Mickus et al., 2020; Brunila & LaViolette, 2022). I have already emphasized its inherent discrepancy with the autonomy of levels. Another crucial disunity is found between structuralism's *externalist* treatment of linguistic categories on the one hand, and linguistic representations as stipulated *internal* vehicles on the other hand. As Gastaldi and Pellissier (2021) note:

> (...) the distributional hypothesis imparts a radically different direction to linguistic research [from generativism], where the knowledge produced is not so much about cognitive agents than about the organization of language. It follows that, understood as a hypothesis, *distributionalism constitutes a statement about the nature of language itself*, rather than about the capacities of linguistic agents. (Gastaldi and Pellissier 2021, p. 570; emphasis in the original)

Distributionalism does not concern internal representations, and has traditionally shunned them explicitly and not at all subtly:

> It remains for linguists to show, in detail, that the speaker has no 'ideas', and that the noise is sufficient
> (Bloomfield, 1936, p. 93)

This fits well with *anti-representationalist* approaches to language, such as the eliminative connectionism of Rumelhart and McClelland (1986). But in order to have scientific import, hypotheses of internal linguistic representations should *contrast* anti-representationalism. Instead, distributionalism ends up effectively conflating them.

Be that as it may, the content-reading could be salvaged at the expense of underlying generative assumptions about phrase-structure, most notably the autonomy of levels. This may well be the right way to go. After all, it would be surprising if linguistic structures formulated in an explicitly anti-connectionist rule-based framework happened to closely match what data-driven LLMs are doing. That notwithstanding, I emphasize that this direction comes at a cost: many proclaimed results of BERTology would need to be re-interpreted as something other than their what manifest literal reading suggests. It is unclear what this should be, if the contrast between genuine linguistic representation and mere "associational learning" is still to be retained.

Moving on to the vehicle-reading, here the main challenge is tackling the triviality problem of interpretational mapping without begging the question. Given that a literal assimilation between LLM-states and abstract SDs is ruled out as a category error, a less direct relation between these is needed, as provided by Egan's (2010; 2018) notion of mathematical content. The challenge then becomes to demonstrate why certain SDs are better mathematical contents of LLMs than others. Since structural probing alone only shows that *some* mapping can be found between LLM-states and SDs, it does not yet establish that those SDs have any special explanatorily value.

BERTology has, in effect, started out by already assuming that the relevant SDs are abstract phrase-structures, and then moved on to find the best correlates for these in LLMs. This methodology leaves out the investigation of whether those SDs are even needed in the first place. Despite having set out to discover what kinds of linguistic representations are present in LLMs, it ends up presupposing much of the answer.[18]

[18] This matter is partly related to debates in theoretical linguistics concerning the use of pre-established categories in analyzing languages beyond those that originally motivated their stipulation. In particular, Haspelmath (2010, 2020) has proposed treating each language on its own terms without imposing prior

While I see no easy way out of this predicament, my overall contention is that BERTology should adopt a more "LLM-first" approach, instead of using SDs pre-defined for different theoretical purposes (within a decidedly non-connectionist framework). As an analogy, Lakoff (1990) characterized the "cognitive commitment" as grounding linguistic theory in what is independently known about human cognition in other disciplines, particularly cognitive psychology.[19] A corresponding commitment would be useful for LLMs as well: their high-level analysis should be built around well-established facts about their algorithmic nature.

It is as of yet unclear what exactly an "LLM-first linguistics" should look like, but one prospect would be to draw from techniques for mapping DNN-algorithms to a human-readable symbolic format. As an example, Weiss et al. (2021) present a programming language called *Restricted Access Sequence Processing Language* (RASP) which models the transformer architecture used in LLMs (Vaswani et al., 2017). Outside of technical details, the main idea is as follows:

> RASP abstracts away low-level operations into simple primitives, allowing a programmer to explore the full potential of a transformer without getting bogged down in the details of how these are realized in practice. At the same time, RASP enforces the information-flow constraints of transformers, preventing anyone from writing a program more powerful than they can express.
> (Weiss et al., 2021, p. 11083)

Weiss et al. provide a RASP-solution to multiple computational problems. These include the recognizion of *Dyck* languages consisting of balanced brackets, with clear relevance for recursion. They further show that training a transformer with the number of layers and attention heads predicted by the RASP-solution attains high task performance, which generally decreases with fewer layers. In addition, RASP can be used to predict attention patterns for each input token (i.e. which tokens give the most information for its encoding). This allows comparing the attention patterns of trained models to such predictions, as well as inducing the learning of RASP-type patterns by directly supervising this via the loss function during training.

From the perspective of Marr's (1982) levels, RASP could be seen as either a low-level computational description or a high-level algorithmic description. In either case, it bears a 1–1 relation to lower-level algorithms enacted by a transformer and can thus act as its *explanatory mechanism* by tapping into its underlying causal structure (Kaplan, 2011; Levy, 2013). As Egan (2017) emphasizes, computational explanation

Footnote 18 continued
assumptions about putatively universal categories (for a critique, see Newmeyer, 2010). But despite manifest affinities between these concerns and the interpretation of LLMs, they also differ in important respects. Even if SDs differed drastically between languages, this would not yet resolve how LLMs represent these language-specific SDs. My present question is not whether LLMs represent different languages as falling under universal SDs, but how they represent *any* SDs in the first place. That being said, the universality of categories in multi-lingual LLMs is an important question in BERTology (Chi et al., 2020), and further research is needed to better understand it in light of the issues I have discussed here.

[19] A potential rejoinder to Lakoff could be that since so little is still known about fundamental properties of cognition, even "language-first" approaches could help uncover some aspects of it, at least in conjunction with other disciplines (c.f. Nefdt, 2023, pp. 186–196). But even if this is true about human cognition, it does not apply to LLMs in the same way: unlike humans, they *are* already algorithmically understood.

Table 4 Summary of six readings of "linguistic representation" and their problems

Type of reading	Variant	Challenge
Content-reading	Realizational	Autonomy of levels
	Fictionalist	Autonomy of levels
	Platonist	Quinean indeterminacy between SDs
Vehicle-reading	Identity	Uninformative
	Direct realizational	Category errors in assimilating levels
	Indirect realizational	Triviality problems in mapping

often departs from such strict mechanistic requirements. Still, when these requirements are well-established on independent grounds, they can be used in the comparison of different candidates for high-level descriptions. RASP provides an example of how this can be done for transformers, including in tasks with linguistic relevance (such as recognizing *Dyck* languages). It grounds concrete restrictions about which tasks are possible in principle and which are precluded on architectural grounds (the latter including e.g. arbitrary loops). Through further connections to empirically assessable hypotheses, it can ground inferences about the computation implemented.

Weiss et al.'s (2021) experiments only concern simple transformer architectures and are far from being directly applicable to LLMs. Nevertheless, their framework provides at least a glimpse of an alternative to the standard BERTology paradigm exemplified by structural probing. Rather than simply finding a mapping from model-states to high-level analyses, Weiss et al. begin with a theoretically grounded computational description that transparently maps to the model. Since links between levels of explanation are explicit and clear, triviality problems never arise.

7 Conclusions and future work

One prominent notion in contemporary NLP is that LLMs bring to question the generative approach to linguistic theory.[20] Somewhat surprisingly, this is not the driving idea behind BERTology. Instead, LLMs are readily interpreted via abstract SDs hypothetically represented by model-internal states. BERTology thus embodies a newfound representational realism in connectionist NLP, shifting from an eliminative to an implementational perspective. Nevertheless, I have argued that ambiguities concerning the interpretation of "linguistic representation" pose major difficulties. Table 4 collects all six readings and their main challenges.

The content-reading flies in the face of the autonomy of linguistic levels, which is a foundational principle of generative linguistics. The vehicle-reading allows retaining the autonomy of levels but brings about different complications. Assimilating LLM-states directly with abstract SDs would be a logical category error; but a more indirect reading raises triviality problems familiar from the philosophy of computation. In sum,

[20] This was recently forcefully argued in a manuscript by Steven Piantadosi: "Modern language models refute Chomsky's approach to language" (Lingbuzz, 2023: https://lingbuzz.net/lingbuzz/007180).

some readings threaten to make representational hypotheses trivially true, while others threaten to make them trivially false. The challenge is, thus, how to fix a non-trivial middle-ground without begging the question.

By the same token, the problem also has repercussions for anti-representationalist interpretations of LLMs. If the notion of "representation" turned out to be irresolvably unclear, this would motivate an anti-representationalist fallback position that simply removes the notion from the theoretical vocabulary altogether.[21] In contrast, a stronger form of anti-representationalism would rely on "representation" having a clear interpretation, but reject its applicability to LLMs. As with representationalist accounts, evaluating these alternatives hinges on scrutinizing the underlying conditional question: what *would* be required for LLMs to represent SDs?

On the face of it, the situation is not unique to BERTology but plagues all applications of linguistics to cognitive science: it remains unclear how different levels of analysis should be linked (c.f. Poeppel & Embick, 2005). Nevertheless, I maintain that Egan's (2010, 2014, 2018) account of computational explanation allows a reasonable interpretation of linguistic mentalism, where abstract SDs constitute mathematical contents that describe a system's computational properties on a high level.[22] The explanatory value of such mathematical contents resides in their use as explanatory "proxies" for surrogative reasoning (Swoyer, 1991). Triviality problems are not a particular threat from this perspective: even though data always underdetermines theory, this is a general aspect of all analysis and not specific to mathematical contents.

In contrast, BERTology has so far focused on methods that merely aim to find *some* mapping from LLM-states to SDs. This leaves a central question unexamined: are those SDs actually *needed* for describing the LLM on a high level of abstraction? The main basis for selecting between mathematical contents is thus missing.

To alleviate the problem, I propose that BERTology should make more use of low-level computational (or high-level algorithmic) frameworks that reliably capture the DNN-pipeline (e.g. Weiss et al., 2021). Here, the choice of mathematical content can be based on already-known facts about the LLM, given the explicit nature of the mapping between different levels of abstraction. While comparable techniques are usually unavailable in the study of human cognition, the algorithmic transparency of LLMs allows building such mappings in a reliable way—at least in principle. The goal should not be to find those LLM-states that best map to some SDs, but the converse: to find those SDs that best explain the LLM.

Acknowledgements I thank Anna-Mari Rusanen, Otto Lappi, and Jami Pekkanen for valuable discussions, and anonymous reviewers for helpful comments.

Funding Open Access funding provided by University of Helsinki (including Helsinki University Central Hospital). This project was funded by the Academy of Finland (decision number 350775). There are no financial or non-financial conflicts of interest related to this work.

[21] See Favela and Machery (2023) for an evaluation of related matters in psychology and neuroscience.

[22] While this account gives linguistic mentalism a reasonable theoretical standing (*pace* Postal, 2009; Behme, 2015), it remains uncommitted to the *truth* of any particular linguistic theory—mentalist or other.

References

Adger, D. (2022). What are linguistic representations? *Mind & Language, 37*(2), 248–260.

Behme, C. (2015). Is the ontology of biolinguistics coherent? *Language Sciences, 47*, 32–42.

Belinkov, Y., & Glass, J. (2019). Analysis methods in neural language processing: A survey. *Transactions of the Association for Computational Linguistics, 7*, 49–72.

Benacerraf, P. (1973). Mathematical truth. *Journal of Philosophy, 70*(19), 661–679.

Blaho, S. (2007). The syntax of phonology: A radically substance-free approach (PhD Thesis). University of Tromsø.

Bloomfield, L. (1933). *Language*. Henry Holt.

Bloomfield, L. (1936). Language or ideas. *Language, 12*(2), 89–95.

Boone, W., & Piccinini, G. (2016). Mechanistic abstraction. *Philosophy of Science, 83*(5), 686–697.

Brentano, F. (1874/1911). Psychology from an empirical standpoint. Routledge and Kegan Paul.

Brunila, M., & LaViolette, J. (2022). What company do words keep? Revisiting the distributional semantics of J.R. Firth & Zellig Harris. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 4403–4417).

Buckner, C. (2018). Empiricism without magic: Transformational abstraction in deep convolutional neural networks. *Synthese, 195*(12), 5339–5372.

Burge, T. (1986). Individualism and psychology. *The Philosophical Review, 95*(1), 3–45.

Cappelen, H., & Dever, J. (2021). *Making AI intelligible: Philosophical foundations*. Oxford University Press.

Chalmers, D. J. (1995). On implementing a computation. *Minds and Machines, 4*, 391–402.

Chi, E.A., Hewitt, J. & Manning, C.D. (2020). Finding universal grammatical relations in multilingual BERT. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 5564–5577).

Chomsky, N. (1957). *Syntactic structures*. Mouton.

Chomsky, N. (1965). *Aspects of the theory of syntax*. MIT Press.

Chomsky, N. (1975). *The logical structure of linguistic theory*. Plenum press.

Chomsky, N. (1980). *Rules and representations*. Columbia University Press.

Chomsky, N. (1986). *Knowledge of language*. Praeger Publications.

Chomsky, N. (1995). *The minimalist program*. MIT Press.

Chomsky, N. (2012). *The science of language*. Cambridge University Press.

Chomsky, N., & Halle, M. (1968). *The sound pattern of English*. Harper & Row.

Coenen, A., Reif, E., Yuan, A., Kim, B., Pearce, A., Viégas, F. & Wattenberg, M. (2019). Visualizing and measuring the geometry of BERT. In Proceedings of the 33rd Conference on Neural Information Processing Systems (pp. 8592–8600).

Collins, J. (2014). Representations without representa: Content and illusion in linguistic theory. In P. Stalmaszczyk (Ed.), *Semantics and beyond: Philosophical and linguistic inquiries* (p. 2764). De Gruyter.

Collins, J. (2023). Internalist priorities in a philosophy of words. *Synthese, 201*(3), 110.

Collins, J., & Rey, G. (2021). Chomsky and intentionality. In N. Allott, T. Lohndal, & G. Rey (Eds.), *A companion to Chomsky* (pp. 488–502). Wiley.

Croft, W. A. (2001). *Radical construction grammar: Syntactic theory in typological perspective*. Oxford University Press.

Danilevsky, M., Qian, K., Aharonov, R., Katsis, Y., Kawas, B. & Sen, P. (2020). A survey of the state of explainable AI for natural language processing. In Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing (pp. 447–459).

Dennett, D. C. (1991). *Consciousness explained.* Little Brown and Company.

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 (pp. 4171–4186).

Dretske, F. I. (1981). *Knowledge and the flow of information*. The MIT Press.

Dunbar, E. (2019). Generative grammar, neural networks, and the implementational mapping problem: Response to Pater. *Language, 95*(1), e87–e98.

Dupre, G. (2022). Georges Rey's representation of language. BJPS Review of Books, , Retrieved from https://www.thebsps.org/reviewofbooks/dupre-on-rey/

Egan, F. (2010). Computation models: A modest role for content. *Studies in History and Philosophy of Science, 41*(3), 253–259.

Egan, F. (2014). How to think about mental content. *Philosophical Studies, 170*(1), 115–135.

Egan, F. (2017). Function-theoretic explanation and the search for neural mechanisms. In D. Kaplan (Ed.), *Explanation and integration in mind and brain science* (pp. 145–163). Oxford University Press.

Egan, F. (2018). The nature and function of content in computational models. In M. Sprevak & M. Colombo (Eds.), *The Routledge handbook of the computational mind* (pp. 247–258). Routledge.

Facchin, M. (2022). Troubles with mathematical contents. *Philosophical Psychology, 5*, 1–24.

Favela, L. H., & Machery, E. (2023). Investigating the concept of representation in the neural and psychological sciences. *Frontiers in Psychology, 5*, 14.

Fodor, J.A. (1981). Some notes on what linguistics is about. N. Block (Ed.), Readings in philosophy of psychology, vol. II (pp. 197–207).

Fodor, J. A. (1990). *A theory of content and other essays*. MIT Press.

Gastaldi, J. L., & Pellissier, L. (2021). The calculus of language: Explicit representation of emergent linguistic structure through type-theoretical paradigms. *Interdisciplinary Science Reviews, 46*(4), 569–590.

Gleitman, L. (2021). Language as a branch of psychology: Chomsky and cognitive science. In N. Allott, T. Lohndal, & G. Rey (Eds.), *A companion to Chomsky* (pp. 109–122). Wiley.

Goldberg, A. E. (2006). *Constructions at work: The nature of generalization in language*. Oxford University Press.

Harris, Z. S. (1951). *Methods in structural linguistics*. The University of Chicago Press.

Haspelmath, M. (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language, 86*(3), 663–687.

Haspelmath, M. (2020). Human linguisticality and the building blocks of languages. *Frontiers in Psychology, 10*, 3056.

Hewitt, J., & Manning, C.D. (2019). A structural probe for finding syntax in word representations. In Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers) (pp. 4129–4138).

Immer, A., Hennigen, L.T., Fortuin, V. & Cotterell, R. (2022). Probing as quantifying inductive bias. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (pp. 1839–1851).

Iosad, P. (2017). *A substance-free framework for phonology: An analysis of the Breton dialect of Bothoa*. Edinburgh University Press.

Jackson, F. (1977). *Perception: A representative theory*. Cambridge University Press.

Jawahar, G., Sagot, B. & Seddah, D. (2019). What does BERT learn about the structure of language? In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3651–3657).

Jelinek, F. (2005). Some of my best friends are linguists. *Language Resources and Evaluation, 39*(1), 25–34.

Kaplan, D. (2011). Explanation and description in computational neuroscience. *Synthese, 183*(3), 339–373.

Karlsson, F. (2006). Recursion in natural languages. In Advances in Natural Language Processing, 5th International Conference on NLP, FinTAL 2006 (p. 1).

Katz, J. (1981). *Language and other abstract objects*. Rowman and Littlefield.

Kovaleva, O., Romanov, A., Rogers, A. & Rumshisky, A. (2019). Revealing the dark secrets of BERT. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 4365–4374).

Kripke, S. (1980). *Naming and necessity*. Harvard University Press.

Kulmizev, A., & Nivre, J. (2022). Schrödinger's tree-on syntax and neural language models. *Frontiers in Artificial Intelligence, 5*, 85.

Kulmizev, A., Ravishankar, V., Abdou, M. & Nivre, J. (2020). Do neural language models show preferences for syntactic formalisms? In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (pp. 4077–4091).

Kuokkanen, J. (2022). Vertical-horizontal distinction in resolving the abstraction, hierarchy, and generality problems of the mechanistic account of physical computation. *Synthese, 200*(3), 247.

Kuznetsov, I., & Gurevych, I. (2020). A matter of framing: The impact of linguistic formalism on probing results. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 171–182).

Lakoff, G. (1990). The invariance hypothesis: Is abstract reason based on imageschemas? *Cognitive Linguistics, 1*(1), 39–74.

Langacker, R. W. (1987). *Foundations of cognitive grammar, volume 1, theoretical prerequisites*. Stanford University Press.

Lasri, K., Pimentel, T., Lenci, A., Poibeau, T. & Cotterell, R. (2022). Probing for the usage of grammatical number. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics Volume 1: Long Papers (pp. 8818–8831).

Laurence, S. (2003). Is linguistics a branch of psychology? In A. Barber (Ed.), *Epistemology of language* (pp. 69–106). Oxford University Press.

Levine, R. (2018). 'Biolinguistics': Some foundational problems. In C. Behme & M. Neef (Eds.), *Essays on linguistic realism* (pp. 21–60). John Benjamins Publishing Company.

Levy, A. (2013). Three kinds of new mechanism. *Biology and Philosophy, 28*(1), 99–114.

Lewis, D. (1970). How to define theoretical terms. *Journal of Philosophy, 67*(13), 426–446.

Li, J., Cotterell, R. & Sachan, M. (2022). Probing via prompting. In Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 1144–1157).

Li, L., Ma, R., Guo, Q., Xue, X. & Qiu, X. (2020). BERT-ATTACK: Adversarial attack against BERT using BERT. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 6193–6202).

Linzen, T., & Baroni, M. (2021). Syntactic structure from deep learning. *Annual Review of Linguistics, 7*, 195–212.

Madabushi, H.T., Romain, L., Divjak, D. & Milin, P. (2020). CXGBERT: BERT meets construction grammar. In Proceedings of the 28th International Conference on Computational Linguistics (pp. 4020–4032).

Manning, C. D., Clark, K., & Hewitt, J. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. *PNAS, 117*(48), 30046–30054.

Marcus, G. F. (1998). Rethinking eliminative connectionism. *Cognitive Psychology, 37*(3), 243–282.

Marr, D. (1982). *Vision*. W.H. Freeman and Company.

Matthews, R. J. (2007). *The measure of mind: Propositional attitudes and their attribution*. Oxford University Press.

McCoy, T., Frank, R., & Linzen, T. (2020). Does syntax need to grow on trees? Sources of hierarchical inductive bias in sequence-to-sequence networks. *Transactions of the Association for Computational Linguistics, 8*, 125–140.

McCoy, T., Pavlick, E. & Linzen, T. (2019). Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 3428–3448).

Mickus, T., Paperno, D., Constant, M. & van Deemter, K. (2020). What do you mean, BERT? Assessing BERT as a distributional semantics model. In Proceedings of the Society for Computation in Linguistics (pp. 350–361).

Miller, P. H. (1999). *Strong generative capacity: The semantics of linguistic formalism*. CSLI Publications.

Millikan, R. G. (1993). Content and vehicle. In N. Eilan, R. McCarthy, & B. Brewer (Eds.), *Spatial representation* (pp. 256–268). Blackwell.

Millikan, R. G. (2017). *Beyond concepts: Unicepts, language, and natural information*. Oxford University Press.

Mueller, A., Frank, R., Linzen, T., Wang, L. & Schuster, S. (2022). Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models. In Findings of the Association for Computational Linguistics: ACL 2022 (pp. 1352–1368).

Nadeem, M., Bethke, A. & Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (pp. 5356–5371).

Neander, K. (2017). *A mark of the mental: A defence of informational teleosemantics*. MIT Press.

Nefdt, R. M. (2023). *Language, science, and structure: A journey into the philosophy of linguistics*. Oxford University Press.

Newmeyer, F. (2010). On comparative concepts and descriptive categories: A reply to Haspelmath. *Language, 86*(3), 688–695.

Odden, D. (2013). Formal phonology. *Nordlyd, 40*(1), 249–273.

OpenAI (2023). GPT-4 technical report (Tech. Rep.).

Ott, D. (2017). Strong generative capacity and the empirical base of linguistic theory. *Frontiers in Psychology, 7*, 8.

Pater, J. (2019). Generative linguistics and neural networks at 60: Foundation, friction, and fusion. *Language, 95*(1), e41–e74.

Pennington, J., Socher, R. & Manning, C.D. (2014). GloVe: Global vectors for word representation. In Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP) (pp. 1532–1543).

Piccinini, G. (2015). *Physical computation: A mechanistic account*. Oxford University Press.

Pinker, S., & Price, A. (1988). On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition, 28*(1–2), 73–193.

Poeppel, D., & Embick, D. (2005). Defining the relation between linguistics and neuroscience. In A. Cutler (Ed.), *Twenty-first century psycholinguistics: Four cornerstones* (pp. 1–16). Lawrence and Erlbaum Associates.

Postal, P. (2003). Remarks on the foundations of linguistics. *The Philosophical Forum, 34*(3–4), 233–252.

Postal, P. (2009). The incoherence of Chomsky's 'biolinguistic' ontology. *Biolinguistics, 3*(1), 104–123.

Putnam, H. (1988). *Representation and reality*. MIT Press.

Quine, W. V. O. (1970). Methodological reflections on current linguistic theory. *Synthese, 21*, 386–398.

Rey, G. (2020). *Representation of language: Philosophical issues in a Chomskyan linguistics*. Oxford University Press.

Rogers, A., Kovaleva, O., & Rumshisky, A. (2020). A primer in BERTology: What we know about how BERT works. *Transactions of the Association for Computational Linguistics, 8*, 842–866.

Rumelhart, D. E., & McClelland, J. L. (1986). On learning the past tenses of English verbs. In J. L. McClelland, D. E. Rumelhart, & T. P. R. Group (Eds.), *Parallel distributed processing: Explorations in the microstructure of cognition: Vol. 2. psychological and biological models* (pp. 216–271). MIT Press.

Sennrich, R., Haddow, B. & Birch, A. (2016). Neural machine translation of rare words with subword units. In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers) (pp. 1715–1725).

Smith, B. C. (2006). Why we still need knowledge of language. *Croatian Journal of Philosophy, 6*(3), 431–456.

Soler, A.G., & Apidianaki, M. (2020). BERT knows Punta Cana is not just beautiful, it's gorgeous: Ranking scalar adjectives with contextualized representations. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (pp. 7371–7385).

Sprevak, M. (2018). Triviality arguments about computational implementation. In M. Sprevak & M. Colombo (Eds.), *Routledge handbook of the computational mind* (pp. 175–191). Routledge.

Swoyer, C. (1991). Structural representation and surrogative reasoning. *Synthese, 87*(3), 449–508.

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).

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Polosukhins, I. (2017). Attention is all you need. In Proceedings of the 31st International Conference on Neural Information Processing (pp. 6000–6010).

Weiss, G., Goldberg, Y. & Yahav, E. (2021). Thinking like transformers. In Proceedings of the 38th international conference on machine learning (pp. 11080–11090).