LLMs as a synthesis between symbolic and distributed approaches to language

Gemma Boleda
Universitat Pompeu Fabra / ICREA
gemma.boleda@upf.edu

Abstract

Since the middle of the 20th century, a fierce battle is being fought between symbolic and distributed approaches to language and cognition. The success of deep learning models, and LLMs in particular, has been alternatively taken as showing that the distributed camp has won, or dismissed as an irrelevant engineering development. In this position paper, I argue that deep learning models for language actually represent a synthesis between the two traditions. This is because 1) deep learning architectures allow for both distributed/continuous/fuzzy and symbolic/discrete/categorical-like representations and processing; 2) models trained on language make use of this flexibility. In particular, I review recent research in interpretability that showcases how a substantial part of morphosyntactic knowledge is encoded in a near-discrete fashion in LLMs. This line of research suggests that different behaviors arise in an emergent fashion, and models flexibly alternate between the two modes (and everything in between) as needed. This is possibly one of the main reasons for their wild success; and it makes them particularly interesting for the study of language. Is it time for peace?

1 Introduction

Since the middle of the 20th century, a fierce battle is being fought between two antagonistic approaches to language and cognition. Although the specifics vary, they can be broadly characterized as follows. Symbolic approaches typically work with discrete, interpretable categories (like "noun", "verb" for parts of speech) and discrete, interpretable rules to combine them (such as those of formal grammars).[1] Distributed approaches instead couple uninterpretable high-dimensional continuous representations, such as vectors, with continuous functions to combine them, such as those defined in the different components of a Transformer.[2]

The debate between the two approaches has taken different forms in different fields: classicism vs. connectionism in in cognitive science (Buckner and Garson, 2019), symbolic / rule-based vs. data-driven / Machine Learning-based approaches in AI (Russell and Norvig, 2020), formalism / generativism vs. functionalism / cognitivism in linguistics (Harris, 2021).[3] The crux of the debate is that, across all these fields, some researchers focus on the rule-like behavior of language and cognition and others on its slippery nature.

The advent of deep learning has added fuel to the scientific fire. In some circles, the success of deep learning models has been alternatively taken as showing that the distributed camp has won, or dismissed as an irrelevant engineering development. A prime example is Steve Piantadosi's 2024 provocatively titled article "Modern language models re-

[1] In early work in NLP, these approaches were paired with top-down processing of linguistic data, through rule-based systems defined by hand. In later work, the processing part has instead been data-driven: data is manually annotated according to a given representation system, and a processing algorithm is induced from the data via machine learning. The latter includes modern neural networks trained for, e.g., dependency parsing. This means that the border between symbolic and distributed approaches is, quite fittingly with this papper, blurry. Relatedly, within formal linguistics different approaches have started softening the discreteness of both categories and rules (see e.g. Erk, 2022, for a comprehensive discussion of probabilistic approaches to semantics and pragmatics). Still, even in these cases the most common approach is to add probabilities or constraints to symbolically-defined rules and categories.

[2] I use "symbolic" and "distributed" as umbrella terms, with related notions being discrete/categorical/localist for the former, and continuous/fuzzy/sub-symbolic for the latter. The different terms touch on different properties that for the purposes of this paper can be lumped together; I will make nuances explicit when needed.

[3] Functionalists do not use distributed representations, but the issues underlying the divide between formalists and functionalists are very related to the general debate, as will become clear during the article, so I am including functionalism in the distributed camp. Also note that the respective positions are rooted in the philosophical traditions of rationalism and empiricism (Markie and Folescu, 2023).

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 9365–9379
November 4-9, 2025

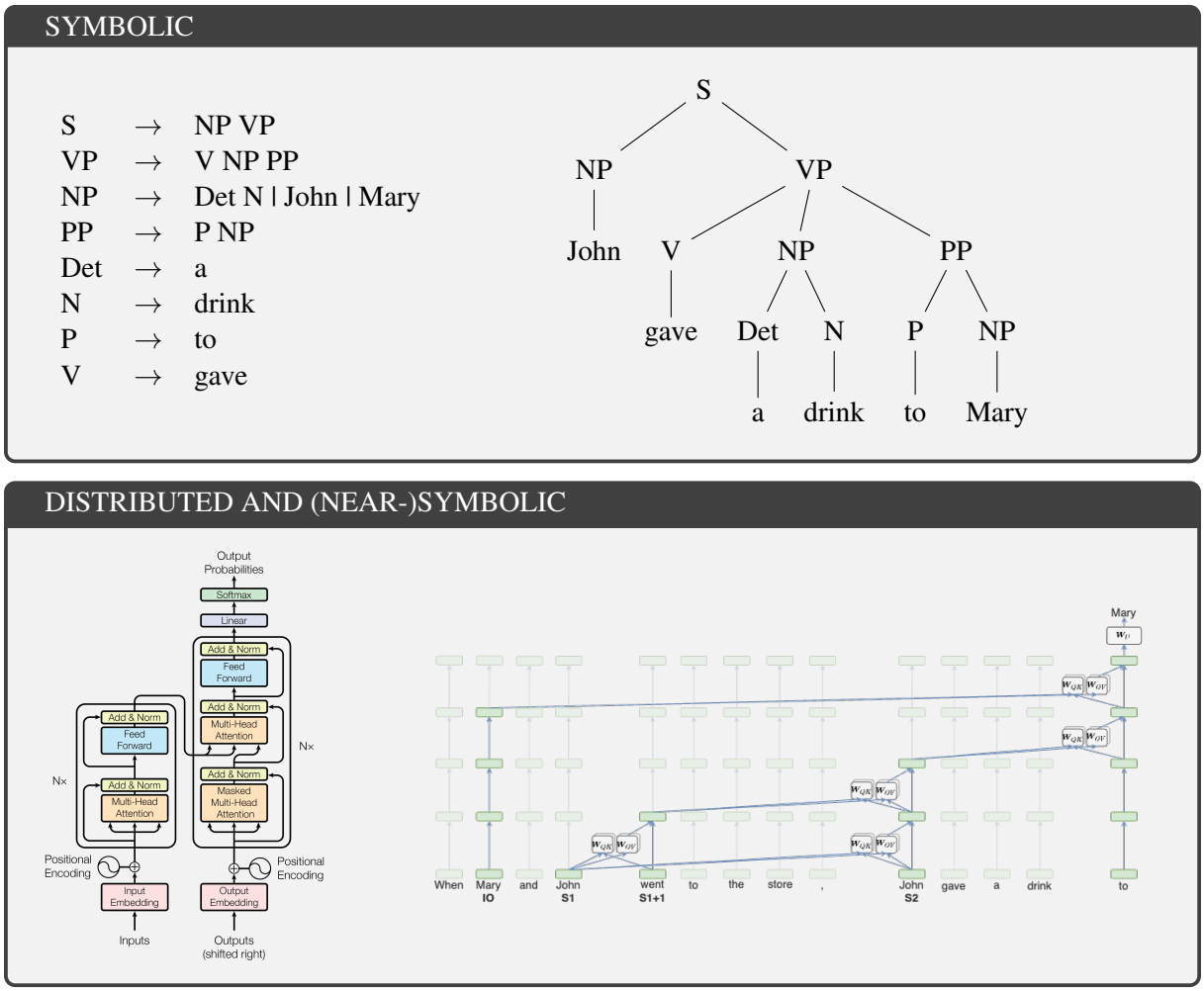

Figure 1: Schematic illustration of the contrast between symbolic formalisms and deep learning. Top: context-free grammar and parse tree for the sentence "John gave a drink to Mary". Bottom: transformer architecture and circuit for the fragment "When Mary and John went to the store, John gave a drink to", with prediction “Mary” (adapted from Vaswani et al. (2017) and Ferrando et al. (2024), with permission). In the circuit, the representations are continuous (vectors), but the different components function together in an interpretable algorithm, with attention heads carrying operations such as copying (see text for details).

fute Chomsky’s approach to language” (Piantadosi, 2024) and the answers it has received, some as heated as “Modern Language Models Refute Nothing” (Rawski and Baumont, 2023); another is the also provocative squib by the late Felix Hill, titled “Why transformers are *obviously* good models of language” (Hill, 2024, emphasis in the original). I believe that these maximalist positions are sterile.

In this position paper, I join more constructive voices exploring what deep learning models might contribute to linguistic theory (Manning, 2015; Warstadt and Bowman, 2022; Futrell and Mahowald, 2025). In particular, I propose that LLMs are actually a **synthesis** between symbolic and distributed approaches to language (Figure 1).[4]

This view rests on two theses. The first is that the debate exists precisely because language is **both** symbolic (or discrete) and distributed (or fuzzy)—and everything in between. Indeed, the sustained debate between the two approaches suggests that neither is able to capture language on its own (Boleda and Herbelot, 2016); and it is necessary to move towards integrated models that capture the full spectrum of language. The second thesis is that modern LLMs are one such kind of model, because they support both distributed and (near-)symbolic representations and processing. In my view, this is one of the main reasons for their amazing success at language.

[4]I center the discussion on LLMs as the most widely adopted type of model, but I will also touch on other kinds of models, such as neural machine translation models.

Now, the synthesis view of LLMs may come as a surprise, since neural networks undoubtedly fall in the distributed camp; however, what is often overlooked in the debate is the fact that neural networks do have the potential for near-symbolic representations and processing (Smolensky and Legendre, 2006, see Section 3 for discussion). Crucially, however, this potential still leaves open what newer-generation neural network models will do with it in practice. My argument is based on recent results in the interpretability literature which suggest that, when deep learning models are exposed to language data and are asked to do a predictive task like language modeling, they develop near-discrete representations and quasi-symbolic processes in addition to distributed ones.

Figure 1 schematically illustrates the contrast between symbolic formalisms and deep learning architectures as I see it, with the example of syntax: while symbolic formalisms are discrete, neural networks afford both distributed and near-discrete representations and processes.

The contributions of the paper are as follows:

- summarizing for the CL/NLP community the ways in which language, as a phenomenon, exhibits both regularity and messiness (Thesis 1; Section 2);
- appraising recent interpretability work that suggests that LLMs deploy near-symbolic representations and processes in addition to distributed ones (Thesis 2; Section 3);
- explaining how this situates LLMs in a debate that has permeated the study of cognition since the 1950s (remainder of the paper).

2 Language is both regular and messy

Regarding Thesis 1, let's start by exemplifying clear cases of regularity. In morphosyntax, for instance, it is common to posit that words belong to different parts of speech (such as noun or verb).[5] Languages mark morphosyntax formally, and the combinatorics of linguistic units are governed by morphosyntactic properties. For instance, the English suffix *-ed* marks tense, and only verbs inflect for tense (*follow/followed*, but *before/*befored*). Similarly, syntactic phrases can stand in different syntactic relations (such as subject, object, or indirect object), which can also be formally marked. For instance, the indirect object in English is marked by the preposition *to*, as in example (1). In many languages, different units standing in a given syntactic relation display agreement (Wechsler and Zlatić, 2003). For instance, in English, subjects and verbs agree in number; in example (2), the singular subject (*A student*) must appear with a singular verb (*is*). In Spanish, there is gender and number agreement also within the noun phrase: in example (3), the highlighted suffix *-a* on the determiner and adjective mark feminine gender, in agreement with the noun's lexical gender.

(1) John gave a drink to/*for Mary

(2) A student is/*are crossing the street

(3) L**a**s partes interesad**a**s
the.F.PL party.PL interested.F.PL
'The interested parties.'

In the syntax-semantics interface, a classic phenomenon is anaphora, with syntactic constraints determining the shape of anaphoric pronouns: for instance, in (4), the pronoun *him* cannot refer to Mark (Chomsky, 1981). As an example from compositional semantics, it is well known that adding negation in a sentential context reverses polarity (Zeijlstra, 2007, see example (5)).

(4) Mark$_i$ combs himself$_i$/*him$_i$

(5) I will/will not come to lunch

All of these phenomena are categorical or discrete, in that there is no "in between" state: verbs inflect for tense, prepositions don't; *is* is right and *are* wrong in the context of singular subjects; *not* is a like a binary switch for polarity; etc. Moreover, in all of them, we find a systematic relationship between form and function, or grammar and meaning, such as *-ed* marking past tense.

This kind of data is what spurred symbolic approaches, where discrete symbols are combined via discrete rules relying on formal features, across domains as different as phonology (Chomsky and Halle, 1968; Prince and Smolensky, 1993), syntax (Chomsky, 1957; Kaplan and Bresnan, 1982; Langacker, 1987; Gazdar et al., 1985; Pollard and Sag, 1994), semantics (Montague, 1974; Kamp and Reyle, 1993; Partee et al., 1990; Pustejovsky, 1995), and pragmatics (Sperber and Wilson, 1995;

[5] The exact shape that this takes depends on the theory, with some theories placing more strength in the grammar and others on the lexicon (see Borer, 2017, for discussion). The difficulties discussed in this section surface in both kinds of theories, though in different ways.

Roberts, 2012; Webber, 2016).[6]

However, one needs only scratch the surface for regularity to break down. The border between parts of speech is notoriously fuzzy (Croft, 2001; Evans and Levinson, 2009); there is no universal agreed upon set of syntactic relations (Napoli, 1993; Van Valin Jr, 2005); negation is far from being a binary switch in many contexts (e.g., *not unhappy* does not mean *happy*), and is hugely complex from a semantic point of view (Zeijlstra, 2007); and even agreement can break down (Wechsler and Zlatić, 2003).

In my view, messiness comes from two main sources. First, **fuzzy borders between categories** like those of parts of speech are pervasive across linguistic domains (Croft, 2001; Dowty, 1991; Haspelmath, 2007).[7] Continuing with the example of parts of speech (see Appendix A for other examples), in many Indo-European languages there is much fuzziness between adjectives and nouns, nouns and verbs, and adjectives and verbs; so much so that, when manually POS-tagging a corpus, a common recourse is to allow for multiple tags (Marcus et al., 1993). An example is shown in (6), where *frightened* could be either a verbal participle, interpreted as in (6-a), or an adjective denoting an emotional state, as in (6-b) (analogous to *sad*, *happy*).[8]

(6) The frightened child
  a. The child who was frightened by something/someone
  b. The child feeling fright

The other way in which languages resist symbolic treatment is by breaking down the systematic **relationship between form and function**. Clear examples are irregular or semi-regular morphological forms, arising from historical processes (Matthews, 1991). For instance, in many English verbs the past tense is not marked by *-ed*, but by an irregular form (*went*, *was*) or a semi-regular pattern (e.g. the so-called ablaut pattern in forms such as *sang*, *drank*, *began*).

While these are purely formal irregularities, most form-function mismatches actually result from an interaction between grammar and meaning. Example (7) showcases agreement *ad sensum*: In sharp contrast to (2) above, here a plural verb is allowed despite the fact that the subject is headed by a singular noun. Agreement *ad sensum* usually happens with singular head nouns that denote sets or pluralities, such as *group* —i.e., cases where there is a mismatch between grammatical and semantic features.

(7) A group of students from New Zealand is/are crossing the street

This kind of semantic leakage into syntax poses a hurdle to symbolic approaches based solely on formal features. Within a symbolic framework, it is still possible to add constraints that take into account the semantics of the head noun in computing agreement, for instance by adding a DENOTES-PLURALITY feature to the representation of the noun. And, while approaches to this phenomenon in formal linguistics are highly sophisticated, they involve integrating semantics along these lines (Wechsler and Zlatić, 2003). This is an apparently easy fix, which however opens a path fraught with difficulties, as encoding conceptual aspects of meaning in a discrete way is arguably unfeasible (see below).

In formal linguistics, the difficulty has been handled by strictly distinguishing semantic features that are grammatically relevant from those that "merely" constitute world knowledge, and circumscribing the empirical scope of linguistic theory to the former (Jackendoff, 1990; Levin, 1993). However, as pointed out in functional approaches like cognitive linguistics, interactions between grammar and conceptual aspects of meaning are pervasive in language; and there is no clear dividing line between semantic properties that are relevant vs. irrelevant for grammar (Langacker, 1987; Fillmore et al., 1988; Goldberg, 1995). Thus, the strict division betweeen linguistic-semantics and other-

[6] I will mainly discuss morphosyntax and semantics, for two reasons: Because these domains are representative of the issues that underlie the debate between symbolic and distributed approaches, and because most of the work on LLM interpretability is in these domains (and the latter is the literature that provides the empirical basis for the synthesis view). However, in Appendix A I also briefly discuss phonology and morphology. Also note that I will also mostly use English examples for space reasons. Nothing in my argument hinges on this choice.

[7] To the point that scholars have often questioned the existence of many categories (see e.g. Croft, 2001, for parts of speech). There is an important theoretical distinction between ascertaining the existence of a given theoretical construct (e.g. in the mind/brain) and gauging its usefulness as a scientific tool. The discussion in this paper is aimed at the former; but the data I discuss cannot distinguish between the two levels.

[8] While context often disambiguates, discussing the manual annotation of the Penn Treebank, Marcus et al. (1993, p. 316) note that "even given explicit criteria for assigning POS tags to potentially ambiguous words, [sometimes] the word's part of speech *simply cannot be decided*" (my emphasis).

semantics is questionable. Moreover, it narrows the empirical scope of linguisticy theory, delegating many language phenomena to other disciplines. Conversely, the risk in functional approaches, given the difficulties involved in encoding the relevant factors, is to forfeit the predictive power of linguistic theories, thus staying at a descriptive level.

And so we enter the ultimate messy place in language: conceptual aspects of meaning. The clearest example is word meaning, which is notoriously fuzzy, vague, and slippery (Wittgenstein, 1953; Kilgarriff, 1997; Boleda, 2020). For instance, in contrast to cases like (1) and (2) above, the similarities and differences between *fast* and *swift* are subtle, and there is no hard and fast rule to determine when to use one and when to use the other. Trying to delimit a word's meaning is similarly challenging; Wittgenstein (1953) famously discussed the case of *game*, concluding that there are no necessary and sufficient conditions determining what counts as a game, and all we can ask for is some kind of "family resemblance". This is why dealing with lexical semantics in terms of discrete features, such as DENOTES-PLURALITY, is fraught with difficulties.

For these reasons, if a fully symbolic approach to parts of speech is problematic, a fully symbolic approach to lexical semantics has been argued to be ultimately unfeasible; very prominently in our community (Boleda, 2020), but also in other traditions from philosophy (Wittgenstein, 1953; Gardenfors, 2014) to lexicography (Kilgarriff, 1997; Hanks, 2000). And, indeed, despite monumental efforts building lexical resources like WordNet, or developing systems for tasks like Word Sense Disambiguation, our community could not model word meaning at scale until distributed methods came along.

That being said, even within this messy domain we still find categorical distinctions. For instance, while different word senses are often impossible to delineate precisely (Kilgarriff, 1997), in some cases the distinction is very clear, e.g. the ANIMAL and COMPUTER DEVICE senses of *mouse*; and some concepts are crisper than others (e.g. FIVE vs. FAST). And other aspects of semantics are largely discrete and symbolic, notably reference (Frege, 1892). We use language to talk about the world and, from a linguistic point of view, there is nothing fuzzy in the distinction between, say, two people with the same name. Thus, whether *Elizabeth Blackburn won the Nobel prize* is true will depend on which Elizabeth Blackburn we're talking about in the given context.[9]

To sum up, this overview suggests that language is indeed both discrete and fuzzy; and that there is no neat discrete/fuzzy divide, nor any area of language that is completely discrete or completely fuzzy. At the same time, there *are* clearly areas that are more discrete (such as morphosyntax) and areas that are fuzzier (such as word meaning).

Crucially, no scholar questions any of the empirical data I have discussed so far; what changes is the way they are appraised. Some traditions focus on the regularities and consider the rest as either special cases or phenomena outside the purview of linguistic theory; whereas others sustain that the ubiquity of these "special cases" makes the regularities an epiphenomenon at best (Weissweiler et al., 2025). These are conscious choices that are based on carefully considered theoretical positions. The clearest example of this dichotomy is the aforementioned generative vs. cognitive divide, with generative linguistics tending towards the former (Chomsky, 1957; Kaplan and Bresnan, 1982; Montague, 1974, among many others) and cognitive linguistics towards the latter (Langacker, 1987; Fillmore et al., 1988; Goldberg, 1995, again among many others). My tenet here is that **both properties are fundamental**, and we cannot reduce language to one or the other. Therefore, we need models that natively encompass both regularity and messiness.

3 LLMs and regularity

To my knowledge, it has not been contested that neural networks in general, and LLMs in particular, can do fuzzy processing of the sort required for e.g. lexical semantics. Therefore, here I will place my emphasis on regularity (Appendix B briefly discusses non-symbolic and non-compositional processing in LLMs).

Indeed, the main criticism of neural networks has historically been their inadequacy in handling rule-like linguistic behavior (see e.g. Manning, 2015 and Pinker and Prince, 1988). However, the distributed camp has long argued that the architecture of neural networks affords symbolic-like processing (Rumelhart and McClelland, 1986; Minsky and Papert, 1988; Smolensky and Legendre, 2006; Smolensky, 2012). Recall from above that distributed approaches couple high-dimensional rep-

[9] As of 2025, according to the internet there are at least two Elizabeth Blackburns: a Nobel laureate and a judge in Florida.

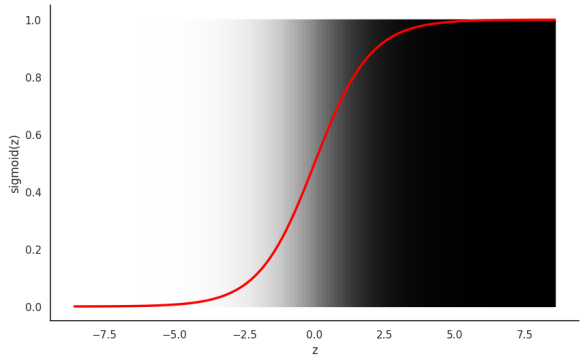

Figure 2: Non-linear functions such as the sigmoid provide the potential for both continuous and near-discrete behavior.

resentations with continuous functions to combine them; importantly, however, some of these functions are nonlinear, and this is what gives neural networks the potential for rule-like behavior (Minsky and Papert, 1988). Take the sigmoid as an example (Figure 2): when its input falls near 0, the value passed on will be continuous; but when its input is larger or smaller, it will be quasi-binary. This allows networks to learn to combine their inputs in a way that leverages non-linearities to build more or less distributed representations and processing, as needed.

If we put this potential together with the properties of language discussed in the previous section, we can expect LLMs to exploit this potential when trained on language. And this is indeed what recent literature on interpretability suggests.

However, what counts as near-symbolic behavior in the context of neural networks? While this is a very difficult notion to pin down, in this paper I count as near-symbolic the existence of small sub-units of the network that are causally involved in encoding or processing a single linguistic construct.[10] This sub-unit can be at different levels of description, from single neurons to larger components like attention heads.

It has been known for close to a decade that neural LMs encode non-trivial knowledge of syntax, including its hierarchical nature (Linzen et al., 2016; Gulordava et al., 2018; Futrell et al., 2019; Rogers et al., 2021). However, most earlier work used techniques such as probing, which could show THAT they encode syntactic knowledge, but not HOW. Newer methods in interpretability (see Ferrando et al., 2024, for a survey) focus on precisely this question, and it is these methods that have provided the clearest evidence for near-discreteness in some aspects of linguistic processing in deep learning models.[11] Most studies focus on morphosyntactic properties or syntactic relations.

Neurons. Several studies have identified neurons that selectively respond to morphosyntactic properties such as part of speech, number, and tense (Bau et al., 2019; Durrani et al., 2023; Gurnee et al., 2023, 2024). For instance, Durrani et al. (2023) find neurons sensitive to part of speech in three multi-lingual LLMs (BERT, RoBERTa, and XL-Net), such as neuron 624 in layer 9 of RoBERTa responding to verbs in the simple past tense and neuron 750 in layer 2 to verbs in the present continuous tense. Moreover, some morphosyntactic neurons are "universal" (Gurnee et al., 2024), in the sense that they can be found across different instantiations of the same auto-regressive LLM. This suggests that language data provide a strong pressure for neurons encoding morphosyntactic properties to arise.

Other studies look at the effects of specific neurons on the output (Geva et al., 2022a,b; Ferrando et al., 2023). Geva et al. (2022b) identified neurons that drastically promote the prediction of tokens with specific features, some of which are morphosyntactic in nature; for instance, neuron 1900 in layer 8 of GPT2 increased the probability of WH words (e.g. "which", "where", "who") and neuron 3025 in layer 6 of WikiLM the probability of adverbs (e.g. "largely", "rapidly", "effectively"). Relatedly, Ferrando et al. (2023) identified a small set of neurons that are functionally active in making grammatically correct predictions (for instance in subject-verb agreement) in models of the GPT2, OPT, and BLOOM families.

My favorite example regarding neurons is Bau et al. (2019), who analyzed neurons associated to morphosyntactic properties in a neural Machine Translation model from the pre-transformer era. Altering the values of these neurons changed the morphosyntactic properties of the translations. For example, altering the activation of a single encoder neuron changed the translation of the whole phrase *The interested parties* into Spanish, switching its gender from feminine to masculine (cf. (8), with the highlighted feminine *-a* vs. masculine *-o* gender

[10] This definition does not imply that this sub-unit need be the only one involved in the relevant behavior; see Section 4 for discussion.

[11] The vast majority of results in this literature concerns English; in what follows, I'll refer to results for English.

suffixes). Remarkably, both translations are correct, but they convey different meanings: the feminine noun *parte* is a general equivalent of *party*, and the masculine *partido* in this context implies specifically a political party.

(8) The interested parties
Original: L**a**s partes interesad**a**s
Modified: L**o**s partid**o**s interesad**o**s

Attention heads. Attention heads with specialized syntactic functions have also been widely found in LLMs and neural MT models (Raganato and Tiedemann, 2018; Clark et al., 2019; Htut et al., 2019; Voita et al., 2019; Krzyzanowski et al., 2024). Figure 3(a) shows the activations of BERT's head 7 in layer 6 for the sentence *many employees are working at its giant Renton, Wash., plant*. This head specializes in the possessive construction; in the example, the possessive determiner (*its*) sharply attends to its head noun (*plant*), in a dependency relation that has 5 intervening tokens in the surface structure. Other heads highlighted in this literature correspond to a wide range of syntactic relations such as subject, object, prepositional complement, adjectival modifier, or adverbial modifier. Note that all heads are near-discrete; Figure 3(b) depicts a head with a broad attention pattern. The existence of these broad heads again suggests the need for distributed processing of other properties, which are however more difficult to interpret.

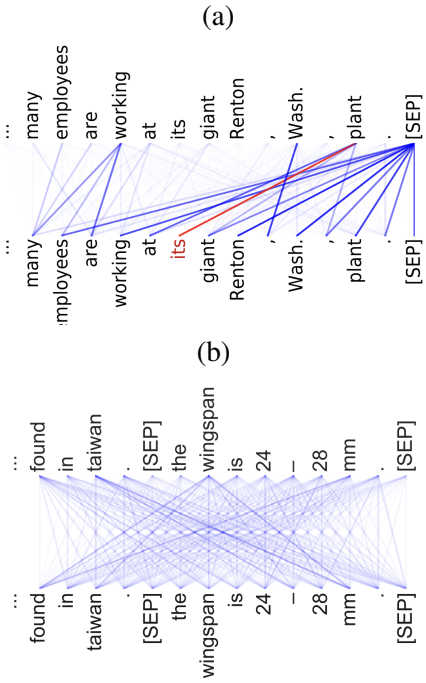

Figure 3: Near-discrete and continuous attention heads in BERT (adapted from Clark et al. (2019), CC-BY license; line thickness is proportional to amount of attention). (a) Head 7 in layer 6 tracks dependencies between possessive determiners and their head nouns dependency in a near-discrete fashion: the determiner "its", highlighted in red, sharply attends to its head noun "plant". (Note that most tokens have near-discrete attention to the [SEP] token. Clark et al. (2019) interpreted this as a no-op signal.) (b) Head 1 in layer 1 instead presents a broad attention pattern with no clear interpretation.

Circuits. In recent years, more evidence has emerged around the notion of "circuit", or subgraphs within neural networks (Cammarata et al., 2020).[12] A particularly relevant example for us is Wang et al. (2023), which describes in detail a circuit in GPT2-small governing the prediction of the indirect object of a sentence. Figure 1 (bottom right) contains a schematic depiction of the circuit for the sentence *When John and Mary went to the store, John gave a drink to* __, where the LLM predicts *Mary*. This interpretable circuit corresponds to an algorithm that identifies the names in the sentence (in the example, *John* and *Mary*), removes the names that appear in the second sentence (*John*), and outputs the remaining name (*Mary*). The model does this through different attention heads with specialized functions.

Merullo et al. (2024) further provide evidence that this circuit is robust (they identify the same circuit in a larger GPT2 model) and generalizes: some of its individual components are reused for a task that is different both semantically and syntactically (it involves the generation of a word denoting the color of an object described among other objects in the preceding context). This suggests that the uncovered circuit is at a quite high level of abstraction in terms of linguistic knowledge. Ferrando and Costa-jussà (2024) contribute further evidence of abstract generalization in circuits. They show that one and the same circuit is responsible for solving subject-verb agreement in English and Spanish in the multi-lingual LLM Gemma 2B.

To sum up, the interpretability literature provides evidence for near-symbolic morphosyntactic processing in different sub-units of LLMs (neurons,

[12] More specifically a circuit is "A subgraph of a neural network. Nodes correspond to neurons or directions (linear combinations of neurons). Two nodes have an edge between them if they are in adjacent layers. The edges have weights which are the weights between those neurons [...]" (Olah et al., 2020).

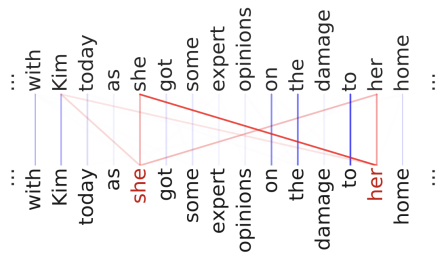

Figure 4: BERT's attention head tracks co-reference dependencies (head 5 in layer 4); adapted from Clark et al. (2019). The anaphoric pronoun "her" sharply attends to antecedent "she".

attention heads, circuits). Much less attention has been devoted to other domains, such as compositional semantics and the syntax-semantic interface, but the existing evidence points in the same direction. For instance, BERT has attention heads specializing in co-reference, in which anaphoric mentions sharply attend to their antecedent (Clark et al., 2019, see Figure 4); and one of the aforementioned "universal neurons" in Gurnee et al. (2024) selectively responds to negation.[13]

4 Discussion

The preceding section has explored the near-symbolic encoding and processing of linguistic information within LLMs. However, as mentioned in the introduction, deep learning models can flexibly switch between discrete and distributed modes, and everything in between. In this, they are very different from formalisms and representations used in theoretical linguistics.

Indeed, as emphasized throughout this paper, while representations in theoretical linguistics are discrete, in LLMs they are at most *near*-discrete. Moreover, there is wide variation in the degree of discreteness (!) exhibited with respect to different phenomena, or even within a phenomenon. For instance, in the work cited above, Durrani et al. (2023) found drastically fewer neurons responding to the POS of function words (like determiners or numerals) than to the POS of content words (like nouns and verbs). They conjectured that the representation of POS in the networks may be more distributed in the latter than in the former case. Similarly, Bau et al. (2019) find that gender and number are represented in a more distributed fashion than tense in the NMT model they analyze.

Another crucial difference with classical formalisms in linguistics is the fact that in neural networks there is a high degree of redundancy (Durrani et al., 2023). For instance, when Wang et al. (2023) ablated some of the heads that they identified in the indirect object circuit explained above, they found that the circuit still worked to some extent. They subsequently went on to identify back-up heads that replaced the role of the initially identified heads. Redundancy is a well-known property of neural networks, and one crucial for their functioning, as it allows for graceful as opposed to catastrophic degradation in behavior (LeCun et al., 1989).

The flip side of redundancy is polysemanticity, that is, the fact that units respond to different properties (Rumelhart et al., 1986). For instance, in many (but not all) cases a neuron that responds to, say, tense, will also respond to some other unrelated property. In a fine-grained analysis of GPT2-small attention heads including manual annotation, Krzyzanowski et al. (2024) found that around 90% are polysemantic. There are advantages to polysemanticity, such as the fact that it allows networks to represent more features than they have dimensions (Elhage et al., 2022, call this "superposition").

If we put the two features together (redundancy and polysemanticity), we see that each feature is represented across different individual neurons and neurons are responsible for different features. By definition, this is what makes a representation distributed (Hinton et al., 1986). So why am I arguing that LLMs are a synthesis between continuous and discrete approaches? Because, as a matter of fact, even if they could represent and process everything in a distributed fashion, they do not. They learn to process some aspects of language in a near-symbolic manner, to the point that specific interpretable algorithms can be reverse-engineered (Ferrando and Costa-jussà, 2024). The 90% figure just mentioned, from Krzyzanowski et al. (2024), implies that 10% of the attention heads analyzed are monosemantic —when they would not need to be, and in fact *poly*semanticity has advantages, as mentioned above. Similarly, most of the "universal neurons" identified by Gurnee et al. (2024) are monosemantic, and they have clear functional roles

[13] The emergence of discrete behavior, and prominently circuits, has been related to what has been called "grokking" (Power et al., 2022), that is, the sudden appearance of generalization capabilities in symbolic tasks. See e.g. Nanda et al. (2023) and Varma et al. (2023) for discussion. Here I focus on symbolic behavior in linguistic representations and processing, but of course its emergence in learning is an exciting topic for further study.

in circuits, such as deactivating attention heads. This stands in stark contrast to, for instance, the much more distributed representation of words in static or contextualized word embeddings. And, indeed, the evidence for near-discrete behavior overwhelmingly comes from domains where symbolic formalisms have been the most successful, such as grammar.

In the context of this paper, it is important to distinguisth between *symbolic* and *interpretable*. This paper's metareviewer remarked that "the paper proposes symbolic representations that lack the properties that make symbolic representations appealing to most researchers, namely that they are interpretable by humans. The authors need to either make the point that the semi-discrete representations and rules in LLMs are actually interpretable in the way that traditional symbols and rules are, or make a case for symbolic representation separate from interpretability." My view falls squarely on the latter side. I do not think we can expect LLMs to ever amount to a complete symbolic framework; nor that they *should*, because in my view language is not completely symbolic either. Therefore, the implication of my paper in this regard is that, if we aim at obtaining more understandable and transparent model architectures, we cannot simply aim at reducing LLMs to symbolic systems.[14] Instead, we need to devise methods that embrace the full symbolic-to-distributed spectrum. Since (as far as I can tell) these methods do not exist yet, the only roadmap I can offer at present is to point out that we need to find new roads.

Relatedly, in using models to elucidate how language works, we should remember that the ultimate testing ground for theories of language and cognition is the brain. Recent work suggests that there may be analogies between LLM and brain encoding of language (Tuckute et al., 2024). However, while research in neuroscience has yielded quite robust results on the different brain regions where language encoding takes place, and some of their roles, it has made much less progress on the properties of linguistic representations and the computations that are carried out during processing (Tuckute et al., 2024), namely, on the topic of this paper. This is another very exciting avenue for further work.

[14] We can of course still extract symbolic knowledge for specific phenomena, and this can be very useful, the same way that symbolic frameworks are very useful in many domains.

5 Conclusion

I started this piece by pointing out that a fierce battle is being fought, since the second half of the 20th century, between symbolic and distributed approaches to language and cognition. And I actually find it worrying that much of this discussion is being led by scholars outside the CL/NLP community. Since we know the most about LLMs, we should participate in ascertaining what they tell us (and what they can't tell us yet) about how language works. One of the motivations of my paper is precisely to foster this kind of debate within our community.

The view I have put forth in this paper is that LLMs are a synthesis between the two approaches; they allow us to integrate regularity and messiness into a single modeling tool, thus overcoming the difficulties faced by symbolic-only or distributed-only systems. Importantly, more and less distributed representations and processing arise in an emergent fashion; LLMs **learn** to behave in a quasi-symbolic fashion at times, in a highly fuzzy and distributed fashion at others, because that allows them to perform better at linguistic tasks, that is, they do so responding to pressures from language data.

So, may it be time for peace? The research I have surveyed has only scratched the surface, and we need everyone on board to continue to make progress in our collective understanding of how language works. In particular, we need methods that go beyond specific, cherry-picked phenomena and allow a systematic exploration of the models (Ferrando and Voita, 2024, is a relevant step in this direction); a better systematization of the empirical landscape to be explored (e.g., Weissweiler et al., 2025, propose to broaden the benchmarks by which we evaluate the linguistic abilities of LLMs); and a stronger engagement with theory when evaluating the implications of deep learning models of language.

Limitations

I am aware that my definition of what counts as near-symbolic in LLMs is, ironically, fuzzy. I think that, given the present state of the art (mechanistic interpretation of deep learning models is still in its infancy), the best I can do is offer an initial definition and many examples of the kind of behavior that I think provides support for my position. Delineating it more precisely is a pressing need for

the future.

As a reviewer pointed out, the interpretability literature "has not yet shown that these localized mechanisms truly function as symbolic components in a larger sense, nor has it demonstrated how they can scale to capture the kinds of generalizable rules or logical inferences that symbolic systems have historically handled". While this is falls outside the scope of the article, it is worth stating that there is much less research in these kinds of phenomena. Interestingly, however, there is at least some tentative evidence for parts of models working as symbolic components. The study of Merullo et al. (2024) discussed above suggests systematic component reuse; similarly, Lindsey et al. (2025) show how Claude 3.5 Haiku performs multi-hop reasoning re-using components across a range of phenomena. One of the cases they discuss is "addition circuitry [that] generalizes between very different contexts". This circuitry selectively activates in contexts where it's useful to perform implicit addition, and includes mechanisms to "represent and store intermediate computations for later use". These are just preliminary findings, and this is certainly another area where much more research is needed, as is research on the interplay between linguistic and reasoning abilities more generally.

Acknowledgments

Thank you to Marco Baroni, Rafael Gutiérrez, Louise McNally and the members of the COLT research group for precious feedback . This research is an output of grant PID2020-112602GB-I00/MICIN/AEI/10.13039/501100011033, funded by the Ministerio de Ciencia e Innovación and the Agencia Estatal de Investigación (Spain).

References

Marco Baroni. 2001. The representation of prefixed forms in the italian lexicon: Evidence from the distribution of intervocalic [s] and [z] in northern italian. In Geert Booij and Jaap van Marle, editors, *Yearbook of Morphology 1999*, pages 121–152. Springer, Dordrecht.

Anthony Bau, Yonatan Belinkov, Hassan Sajjad, Nadir Durrani, Fahim Dalvi, and James Glass. 2019. Identifying and controlling important neurons in neural machine translation. In *International Conference on Learning Representations*.

Gemma Boleda. 2020. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234.

Gemma Boleda and Aurélie Herbelot. 2016. Formal distributional semantics: Introduction to the special issue. *Computational Linguistics*, 42(4):619–635.

Hagit Borer. 2017. Morphology and syntax. In Andrew Spencer and Arnold M. Zwicky, editors, *The Handbook of Morphology*. Wiley-Blackwell.

Cameron Buckner and James Garson. 2019. Connectionism. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Fall 2019 edition. Metaphysics Research Lab, Stanford University.

Nick Cammarata, Shan Carter, Gabriel Goh, Chris Olah, Michael Petrov, Ludwig Schubert, Chelsea Voss, Ben Egan, and Swee Kiat Lim. 2020. Thread: Circuits. *Distill*. Https://distill.pub/2020/circuits.

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Noam Chomsky. 1957. *Syntactic Structures*. Mouton & Co.

Noam Chomsky. 1981. *Lectures in Government and Binding: The Pisa lectures*. Number 9 in Studies in Generative Grammar. Foris, Dordrecht.

Noam Chomsky and Morris Halle. 1968. *The Sound Pattern of English*. Harper & Row, New York.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Maria Copot, Timothee Mickus, and Olivier Bonami. 2022. Idiosyncratic frequency as a measure of derivation vs. inflection. *Journal of Language Modelling*, 10(2):193–240. Number: 2.

William A. Croft. 2001. *Radical Construction Grammar: Syntactic Theory in Typological Perspective*. Oxford University Press, Oxford.

David Dowty. 1991. Thematic proto-roles and argument selection. *language*, 67(3):547–619.

Nadir Durrani, Fahim Dalvi, and Hassan Sajjad. 2023. Discovering salient neurons in deep nlp models. *Journal of Machine Learning Research*, 24(362):1–40.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. 2022. Toy models of superposition. *Transformer Circuits Thread*.

Katrin Erk. 2022. The probabilistic turn in semantics and pragmatics. *Annual Review of Linguistics*, 8:101–121.

Nicholas Evans and Stephen C. Levinson. 2009. The myth of language universals: Language diversity and its importance for cognitive science. *Behavioral and Brain Sciences*, 32(5):429–448.

Javier Ferrando and Marta R. Costa-jussà. 2024. On the similarity of circuits across languages: a case study on the subject-verb agreement task. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10115–10125, Miami, Florida, USA. Association for Computational Linguistics.

Javier Ferrando, Gerard I. Gállego, Ioannis Tsiamas, and Marta R. Costa-jussà. 2023. Explaining how transformers use context to build predictions. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5513, Toronto, Canada. Association for Computational Linguistics.

Javier Ferrando, Gabriele Sarti, Arianna Bisazza, and Marta R. Costa-Jussà. 2024. A primer on the inner workings of transformer-based language models. *Preprint*, arXiv:2405.00208.

Javier Ferrando and Elena Voita. 2024. Information flow routes: Automatically interpreting language models at scale. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17432–17445, Miami, Florida, USA. Association for Computational Linguistics.

Charles Fillmore, Paul Kay, and Catherine O'Connor. 1988. Regularity and idiomaticity in grammatical constructions: The case of let alone. *Language*, 64:501–538.

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

Richard Futrell and Kyle Mahowald. 2025. How linguistics learned to stop worrying and love the language models. *Preprint*, arXiv:2501.17047.

Richard Futrell, Ethan Wilcox, Takashi Morita, Peng Qian, Miguel Ballesteros, and Roger Levy. 2019. Neural language models as psycholinguistic subjects: Representations of syntactic state. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 32–42, Minneapolis, Minnesota. Association for Computational Linguistics.

Peter Gardenfors. 2014. *The geometry of meaning: Semantics based on conceptual spaces*. MIT press.

Aina Garí Soler and Marianna Apidianaki. 2021. Let's play mono-poly: BERT can reveal words' polysemy level and partitionability into senses. *Transactions of the Association for Computational Linguistics*, 9:825–844.

Aina Garí Soler, Marianna Apidianaki, and Alexandre Allauzen. 2019. Word usage similarity estimation with sentence representations and automatic substitutes. In *Proceedings of the Eighth Joint Conference on Lexical and Computational Semantics (*SEM 2019)*, pages 9–21, Minneapolis, Minnesota. Association for Computational Linguistics.

Gerald Gazdar, Ewan Klein, Geoffrey K Pullum, and Ivan A Sag. 1985. *Generalized phrase structure grammar*. Harvard University Press.

Mor Geva, Avi Caciularu, Guy Dar, Paul Roit, Shoval Sadde, Micah Shlain, Bar Tamir, and Yoav Goldberg. 2022a. LM-debugger: An interactive tool for inspection and intervention in transformer-based language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 12–21, Abu Dhabi, UAE. Association for Computational Linguistics.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022b. Transformer feed-forward layers build predictions by promoting concepts in the vocabulary space. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 30–45, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Adele E. Goldberg. 1995. *Construction grammar: a construction grammar approach to argument structure*. University of Chicago Press.

Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. 2018. Colorless green recurrent networks dream hierarchically. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1195–1205, New Orleans, Louisiana. Association for Computational Linguistics.

Wes Gurnee, Theo Horsley, Zifan Carl Guo, Tara Rezaei Kheirkhah, Qinyi Sun, Will Hathaway, Neel Nanda, and Dimitris Bertsimas. 2024. Universal neurons in GPT2 language models. *arXiv preprint arXiv:2401.12181*.

Wes Gurnee, Neel Nanda, Matthew Pauly, Katherine Harvey, Dmitrii Troitskii, and Dimitris Bertsimas. 2023. Finding neurons in a haystack: Case studies with sparse probing. *Preprint*, arXiv:2305.01610.

Coleman Haley, Edoardo M. Ponti, and Sharon Goldwater. 2024. Corpus-based measures discriminate inflection and derivation cross-linguistically. *Journal of Language Modelling*, 12(2):477–529. Number: 2.

Patrick Hanks. 2000. Do word meanings exist? *Computers and the Humanities*, 34(1/2):205–215.

Randy Allen Harris. 2021. *The linguistics wars: Chomsky, Lakoff, and the battle over deep structure*. Oxford University Press.

Martin Haspelmath. 2007. Pre-established categories don't exist: Consequences for language description and typology. *Linguistic Typology*, 11(1).

Martin Haspelmath. 2011. The indeterminacy of word segmentation and the nature of morphology and syntax. *Folia Linguistica*, 45(1):31–80.

Felix Hill. 2024. Why transformers are obviously good models of language. *Preprint*, arXiv:2408.03855.

G. E. Hinton, J. L. McClelland, and D. E. Rumelhart. 1986. Distributed representations. In D. E. Rumelhart and J. L. McClelland, editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition. Volume 1: Foundations*, pages 77–109. MIT Press, Cambridge, MA.

Phu Mon Htut, Jason Phang, Shikha Bordia, and Samuel R. Bowman. 2019. Do attention heads in bert track syntactic dependencies? *Preprint*, arXiv:1911.12246.

Yusuke Ide, Joshua Tanner, Adam Nohejl, Jacob Hoffman, Justin Vasselli, Hidetaka Kamigaito, and Taro Watanabe. 2025. CoAM: Corpus of all-type multiword expressions. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27004–27021, Vienna, Austria. Association for Computational Linguistics.

Ray Jackendoff. 1990. *Semantic Structures*. MIT Press, Cambridge, Massachusetts.

Hans Kamp and Uwe Reyle. 1993. *From Discourse to Logic: Introduction to Model-theoretic Semantics of Natural Language, Formal Logic and Discourse Representation Theory*. Kluwer, Dordrecht.

Ronald M. Kaplan and Joan Bresnan. 1982. Lexical-functional grammar: A formal system for grammatical representation. In Joan Bresnan, editor, *The Mental Representation of Grammatical Relations*, chapter 4, pages 173–281. MIT Press, Cambridge, Massachusetts.

Adam Kilgarriff. 1997. I don't believe in word senses. *Computers and the Humanities*, 31(2):91–113.

Robert Krzyzanowski, Connor Kissane, Arthur Conmy, and Neel Nanda. 2024. We inspected every head in gpt-2 small using saes so you don't have to. Alignment Forum.

Ronald W. Langacker. 1987. *Foundations of Cognitive Grammar Volume I*. Stanford University Press, Stanford, California.

Yann LeCun, John Denker, and Sara Solla. 1989. Optimal brain damage. In *Advances in Neural Information Processing Systems*, volume 2. Morgan-Kaufmann.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press, Chicago and London.

Rochelle Lieber. 2004. *Morphology and lexical semantics*, volume 104. Cambridge University Press.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, and 8 others. 2025. On the biology of a large language model. *Transformer Circuits Thread*.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Shervin Malmasi, Anjie Fang, Besnik Fetahu, Sudipta Kar, and Oleg Rokhlenko. 2022. MultiCoNER: A large-scale multilingual dataset for complex named entity recognition. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3798–3809, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Christopher D. Manning. 2015. Computational linguistics and deep learning. *Computational Linguistics*, 41(4):701–707.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Peter Markie and M. Folescu. 2023. Rationalism vs. Empiricism. In Edward N. Zalta and Uri Nodelman, editors, *The Stanford Encyclopedia of Philosophy*, Spring 2023 edition. Metaphysics Research Lab, Stanford University.

Peter Hugoe Matthews. 1991. *Morphology*. Cambridge university press.

Jack Merullo, Carsten Eickhoff, and Ellie Pavlick. 2024. Circuit component reuse across tasks in transformer language models. In *The Twelfth International Conference on Learning Representations*.

Marvin L Minsky and Seymour A Papert. 1988. Perceptrons: expanded edition.

Richard Montague. 1974. English as a formal language. In Richmond H. Thomason, editor, *Formal philosophy: Selected Papers of Richard Montague*, chapter 6, pages 188–221. Yale University Press, New Haven.

Neel Nanda, Lawrence Chan, Tom Lieberum, Jess Smith, and Jacob Steinhardt. 2023. Progress measures for grokking via mechanistic interpretability. In *The Eleventh International Conference on Learning Representations*.

Donna Jo Napoli. 1993. *Syntax: Theory and problems*. Oxford University Press, New York.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020. Zoom in: An introduction to circuits. *Distill*. Https://distill.pub/2020/circuits/zoom-in.

Barbara H. Partee, Alice Meulen, and Robert E. Wall. 1990. *Mathematical Methods in Linguistics*. Kluwer, Dordrecht.

Steven T. Piantadosi. 2024. Modern language models refute chomsky's approach to language. In Edward Gibson and Moshe Poliak, editors, *From Fieldwork to Linguistic Theory: A Tribute to Dan Everett*, pages 353–414. Language Science Press.

Janet B. Pierrehumbert. 2016. Phonological representation: Beyond abstract versus episodic. *Annual Review of Linguistics*, 2(Volume 2, 2016):33–52.

Steven Pinker and Alan Prince. 1988. On language and connectionism: Analysis of a parallel distributed processing model of language acquisition. *Cognition*, 28(1-2):73–193.

Carl Pollard and Ivan A Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

Alethea Power, Yuri Burda, Harri Edwards, Igor Babuschkin, and Vedant Misra. 2022. Grokking: Generalization beyond overfitting on small algorithmic datasets. *CoRR*, abs/2201.02177.

Alan Prince and Paul Smolensky. 1993. Optimality theory: Constraint interaction in generative grammar. Technical Report 2, Rutgers University Center for Cognitive Science.

James Pustejovsky. 1995. *The Generative Lexicon*. The MIT Press, Cambridge, MA (etc.).

Alessandro Raganato and Jörg Tiedemann. 2018. An analysis of encoder representations in transformer-based machine translation. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 287–297, Brussels, Belgium. Association for Computational Linguistics.

Jon Rawski and Joshua Baumont. 2023. Modern language models refute nothing. `https://lingbuzz.net/lingbuzz/007203`.

Craige Roberts. 2012. Information structure: Towards an integrated formal theory of pragmatics. *Semantics and pragmatics*, 5:6–1.

Simon Roessig, Doris Mücke, and Martine Grice. 2019. The dynamics of intonation: Categorical and continuous variation in an attractor-based model. *PLoS ONE*, 14(5):e0216859.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the Association for Computational Linguistics*, 8:842–866.

David E. Rumelhart and James L. McClelland. 1986. On learning the past tenses of english verbs. In *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 2, pages 216–271. MIT Press.

David E Rumelhart, James L McClelland, PDP Research Group, and 1 others. 1986. *Parallel distributed processing, volume 1: Explorations in the microstructure of cognition: Foundations*. The MIT press.

Stuart Russell and Peter Norvig. 2020. *Artificial Intelligence: A Modern Approach*, 4th edition. Pearson.

Paul Smolensky. 2012. Symbolic functions from neural computation. *Philosophical Transactions of the Royal Society A*, 370:3543–3569.

Paul Smolensky and Géraldine Legendre. 2006. *The Harmonic Mind: From Neural Computation to Optimality-Theoretic Grammar, Volume 1*. MIT Press, Cambridge, MA.

Dan Sperber and Deirdre Wilson. 1995. *Relevance: Communication and Cognition*, 2nd edition. Blackwell Publishing.

Harish Tayyar Madabushi, Edward Gow-Smith, Marcos Garcia, Carolina Scarton, Marco Idiart, and Aline Villavicencio. 2022. SemEval-2022 task 2: Multilingual idiomaticity detection and sentence embedding. In *Proceedings of the 16th International Workshop on Semantic Evaluation (SemEval-2022)*, pages 107–121, Seattle, United States. Association for Computational Linguistics.

Greta Tuckute, Nancy Kanwisher, and Evelina Fedorenko. 2024. Language in brains, minds, and machines. *Annual Review of Neuroscience*, 47:277–301.

Robert D Van Valin Jr. 2005. *Exploring the syntax-semantics interface*. Cambridge University Press.

Vikrant Varma, Rohin Shah, Zachary Kenton, János Kramár, and Ramana Kumar. 2023. Explaining grokking through circuit efficiency. *Preprint*, arXiv:2309.02390.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, page 6000–6010.

Aline Villavicencio, Francis Bond, Anna Korhonen, and Diana McCarthy. 2005. Introduction to the special issue on multiword expressions: Having a crack at a hard nut. *Computer Speech & Language*, 19(4):365–377.

Elena Voita, David Talbot, Fedor Moiseev, Rico Sennrich, and Ivan Titov. 2019. Analyzing multi-head self-attention: Specialized heads do the heavy lifting, the rest can be pruned. In *Proceedings of the*

57th Annual Meeting of the Association for Computational Linguistics, pages 5797–5808, Florence, Italy. Association for Computational Linguistics.

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

Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. 2023. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *ICLR - The Eleventh International Conference on Learning Representations*.

Alex Warstadt and Samuel R. Bowman. 2022. What artificial neural networks can tell us about human language acquisition. In *Algebraic Structures in Natural Language*, pages 17–60. CRC Press.

Bonnie Lynn Webber. 2016. *A formal approach to discourse anaphora*. Routledge.

Stephen Wechsler and Larisa Zlatić. 2003. *The Many Faces of Agreement*. Stanford Monographs in Linguistics. CSLI Publications.

Leonie Weissweiler, Kyle Mahowald, and Adele Goldberg. 2025. Linguistic generalizations are not rules: Impacts on evaluation of lms. *arXiv preprint arXiv:2502.13195*.

Ludwig Wittgenstein. 1953. *Philosophical Investigations*. Basil Blackwell, Oxford.

Hedde Zeijlstra. 2007. Negation in natural language: On the form and meaning of negative elements. *Language and Linguistics Compass*, 1(5):498–518.

A Regularities and messiness across domains

As mentioned in the main section of the paper, we find regularity and messiness throughout the different aspects and domains of language. Syntax and semantics were discussed in the main text; here we add some brief notes about phonology and morphology.

In phonetics and phonology, "there is accumulating evidence that the categorical and continuous aspects of speech are deeply intertwined" (Roessig et al., 2019). A basic aspect is a language's phonological inventory, i.e. the set of phonemes that constitute it. Phonemes are "the smallest contrastive sound unit[s] in a language that can distinguish meaning"; for instance, in Catalan, /l/ and /r/ are phonemes (e.g. *cara* means 'face', *cala* 'small bay'), but in Japanese they are allophones, that is, different phonetic realizations or pronunciations of the same phoneme. There is ample evidence of categorical conceptualization of continuous speech by speakers into phonemes, but also equally ample evidence for challenges to a purely symbolic treatment of phonology, analogous to those discussed in the main text for parts of speech (Pierrehumbert, 2016). Similarly, different phonemes combine according to phonotactic rules or constraints (e.g. in English, but not Catalan, the sequence of phonemes /sp/ can begin a syllable, e.g. in the word *spa*), and those face difficulties analogous to those of syntax, with semantics leaking into phonotactics (e.g. Baroni, 2001).[15]

As for morphology, a basic notion such as that of word is as common as it is controversial and challenging to delimit (Haspelmath, 2011). Similarly, inflection and derivation are considered fundamentally different kinds of processes, but their border is again fuzzy (Copot et al., 2022; Haley et al., 2024). Moreover, we find that derivational morphology, like inflectional morphology, presents pervasive regularities together with irregularities and semi-regularities (Matthews, 1991). For instance, the English suffix *-ion* selects for verbal roots and produces nouns referring to actions, processes, or results (create/creation, operate/operation, donate/-donation). However, this pattern has many exceptions due to historical borrowing, primarily from Latin and French. Many verbs and their corresponding nouns were borrowed into English as separate words, preserving irregularities from the original language (*destroy/*destroyion/destruction*, from Latin *destructio admit/*admition/admission*, from Latin *admissio*). Remember that we discussed an analogous case with irregular verbs in English in the main text. Furthermore, derivational morphology also displays the semi-regular match between form and meaning that we found in morphosyntax (Lieber, 2004; Boleda, 2020).

B Non-symbolic and non-compositional linguistic processing in LLMs

In the main text I have taken for granted that LLMs can do non-symbolic and non-compositional linguistic processing. Here I am presenting evidence for the latter, for completeness. The realm with the richest evidence of non-symbolic processing is that of conceptual aspects of meaning, which as

[15] In Catalan, *spa* is pronounced /əs'pa/.

discussed in the paper defy symbolic treatment. I discuss two representative examples, lexical semantics and sentential semantics, but the evidence is vast.

As for lexical semantics, recall that, to account for a word's meaning and usage, symbolic methods like those in traditional Word Sense Disambiguation define a set of senses for each word and assign each use of a word in context to one of the senses. This has long been known to be problematic, as many sense boundaries are blurry and word usages can be more and less similar to each other (Kilgarriff, 1997). LLMs provide instead graded representations for words in context, and this has been linked to their leap in success. Among the many papers about this, let me point to two specific analyses using BERT: Garí Soler et al. (2019) show that BERT-estimated similarity between word usages corresponds to human similarity scores; and Garí Soler and Apidianaki (2021) show that "BERT representations offer good estimates of the partitionability of words into senses", that is, to how easy or difficult it is to define different senses for a given word.

As for sentential semantics, similarly, while logic-based relations between sentences like entailment are more discrete in nature, similarity relations between sentences are clearly on a continuum. LLMs are good at modeling this continuum, as measured in the Semantic Textual Similarity task (STS). One of the tasks in the GLUE benchmark is STS, using data from (Cer et al., 2017), which consists of pairs of sentences and human-annotated similarity scores. All top 20 models in the leaderboard of GLUE achieve a correlation of 0.91 or more with the human data (both Spearman and Pearson);[16] human correlation is 0.93. Note that GLUE evaluates models simultaneously on a range of linguistic tasks; these models perform well at STS while at the same time performing well on a range of other natural language tasks, including entailment (NLI, RTE). This is further evidence for the synthesis view, this time from the point of view of model behavior rather than internal representations and processing.

Turning to non-compositional linguistic processing, a phenomenon that has received a lot of attention in NLP are so-called Multi-Word Expressions (Villavicencio et al., 2005) like *United Arab Emirates*, which have syntactic structure but often function as a single linguistic unit. LLMs perform well at MWE-related tasks like MWE detection (Tayyar Madabushi et al., 2022); moreover, Ide et al. (2025) show that a fine-tuned LLM outperforms the previously best system at MWE identification in a varied MWE corpus. This previous system included a rule-based component and a specifically trained neural network component; again, this suggests that LLMs implicitly implement more symbolic and more distributed processing, and this is benefitial for non-compositional and semi-compositional phenomena. Similarly, LLMs show strong performance at Named Entity Recognition (Malmasi et al., 2022), another well-studied non-compositional phenomenon. Taken together, this suggests that LLMs do non-trivial processing of non-compositional aspects of language, too.

[16] https://gluebenchmark.com/leaderboard, retrieved Sept 19 2025. GLUE is described in (Wang et al., 2018).