LLMs as a synthesis between symbolic and continuous approaches to language

Anonymous ACL submission

Abstract

Since the middle of the 20th century, a fierce battle is being fought between symbolic and continuous approaches to language and cognition. The success of deep learning models, and LLMs in particular, has been alternatively taken as showing that the continuous camp has won, or dismissed as an irrelevant engineering development. However, 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 continuous/distributed and symbolic/discrete-like representations and computations; 2) models trained on language make use this flexibility. In particular, I review recent research in mechanistic 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 is also what makes them particularly interesting for the study of language and cognition. 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 details vary, they can be broadly characterized as follows. Symbolic approaches use discrete formalisms to represent language. Examples in computational linguistics (CL) are POS tags, parse trees, and discrete word senses.[1] Continuous approaches use distributed representations, in the form of high-dimensional algebraic objects such as vectors. In CL, static word embeddings (à la word2vec; Mikolov et al., 2013) are a prime example.

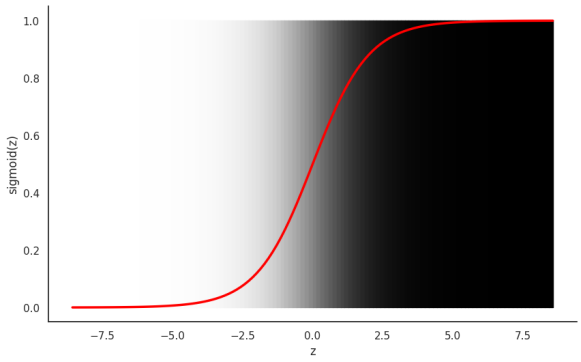

Figure 1: Non-linear functions such as the sigmoid provide the potential for both continuous and near-discrete behavior.

The debate has taken different forms in different fields; in cognitive science, this opposition has been dubbed classicism vs connectionism (Buckner and Garson, 2019); in AI, different terms are used by different authors (Russell and Norvig, 2020); in linguistics, the issues underlying the divide between generative and cognitive linguists are related to this debate Harris, 1993. 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. However, the fact that this debate exists might be a testimony to the fact that language and cognition are **both** symbolic (or discrete) and continuous (or fuzzy) —and everything in between (see Section 2).

Focusing on language, in this position paper I argue that modern LLMs support both continuous and (near-)discrete representations and processing, and thus are a **synthesis** between the two antagonis-

[1] In early work, 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.

tic positions.[2] This may seem a strange position to adopt, since neural networks undoubtedly fall in the continuous camp. However, something that is often overlooked in the debate is the fact that neural networks have the potential for (near-)discrete behavior. This potential comes from the non-linearities in their architecture (Minsky and Papert, 1988). Take the sigmoid as an example (Figure 1): 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 its inputs in a way that leverages the two behaviors. Crucially, while neural network architectures allow for flexibility in behavior, what they will do with this potential in practice is an open question.

The present paper is motivated by the fact that LLMs do seem to indeed exploit the potential for quasi-symbolic behavior with respect to language: A lot of recent work within interpretability provides evidence for near-discrete representations and processes, as discussed in Section 3. What is more, these representations arise in an emergent fashion; LLMs **learn** to behave in a a quasi-symbolic fashion, because that allows them to perform better at linguistic tasks. This, in turn, may be one of the reasons for their amazing success at capturing natural language.

2 How discrete is language?

Linguists have found symbolic formalisms useful across all main domains of language, such as phonology (Chomsky and Halle, 1968; Prince and Smolensky, 1993), morphosyntax (Chomsky, 1957; Bresnan, 1982; Langacker, 1987; Pollard and Sag, 1994; Goldberg, 1995), semantics (Montague, 1974; Partee et al., 1990; Pustejovsky, 1995), and pragmatics (Grice, 1989; Sperber and Wilson, 1995). In this article, I will focus on morphosyntax and semantics.

Work in morphosyntax posits for instance that words belong to different parts of speech (such as determiner, noun, or verb) and can stand in different syntactic relations (such as subject, object, or indirect object). Languages mark morphosyntax formally, and restrictions in the co-occurrence of linguistic units (morpheme, words, clauses) are governed by morphosyntactic properties. For instance, in English only verbs inflect for tense; and, in most, verbs past tense is signaled ty the suffix *-ed* ("follow/followed"). Similarly, only some verbs allow for indirect objects, and the indirect object in English is marked by the preposition *to* (see example (1)). In many languages different units in the sentence display agreement (Wechsler and Zlatić, 2003). Example (1) showcases how, in Spanish, there is gender and number agreement within the noun phrase: the highlighted suffix *-a* on the determiner and adjective mark feminine gender, in agreement with the noun's lexical gender. Similarly, in English, subjects and verbs agree in number; in example (3), the singular subject ("A student") cannot combine with a plural verb ("are").

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

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

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

In compositional semantics and the syntax-semantics interface, we find phenomena such as negation, where, in a sentential context, adding negation reverses polarity (Zeijlstra, 2007, see example (4)), and anaphora, where syntactic constraints determine the shape of anaphoric pronouns: for instance, in (5), the pronoun "him" cannot refer to Mark (Chomsky, 1981).

(4) I will/will not come to lunch

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

All of these phenomena are largely symbolic and discrete, in that there is no "in between" state: the choice between "is" and "are" is determined by the number of the subject; "not" is a like a binary switch for polarity in sentences; etc. However, even in this realm one only needs to scratch the surface for discreteness 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 (Dowty, 1991); 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).

Consider agreement *ad sensum*, exemplified in (6). Here, the syntactic subject is the singular noun "group", but the plural form, forbidden in example (3), is allowed in this case.

[2] I center the discussion on LLMs as the most widely adopted type of model, but in the discussion I will also include other models, such as neural machine translation models. I will signal when I do.

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

This example showcases the interaction between grammar and meaning, as *ad sensum* agreement happens with singular head nouns that denote pluralities, such as "group". Aspects of meaning that are conceptual in nature are, indeed, the source of much of language's fuzziness (Wittgenstein, 1953): Word meaning, for instance, is notoriously fuzzy, vague, and slippery. As an example, in contrast to cases like (2-5) 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. Moreover, while most words have many meanings, more often than not they are difficult to delineate (Kilgarriff, 1997). Hence, symbolic formalisms with discrete representations are highly problematic for word meaning (Wittgenstein, 1953; Kilgarriff, 1997; Boleda, 2020).

Construction grammar, a family of theories within cognitive linguistics (Langacker, 1987; Lakoff, 1987; Fillmore et al., 1988; Goldberg, 1995; Croft, 2001), has put the relationship between conceptual meaning and grammar center stage. While these approaches still use discrete representations, they contest the existence of abstract syntactic rules of the sort exemplified in Figure 2 (top), which are advocated by generative linguists. Scholars in construction grammar instead propose the existence of patterns (termed *constructions*) at different levels of abstraction, consisting of pairings of form and meaning.[3] Constructions are often semi-productive and heavily dependent on conceptual aspects of meaning, such that it is again difficult to establish hard and fast rules for their use that can be specified on formal grounds only. For instance, the verb "to sneeze", which is not causative, can sometimes be used felicitously in a causative construction, as in example (7), attributed to Adele Goldberg by Hill (2024).

(7) They sneezed the foam off the cappuccino

It should however be noted that not all aspects of meaning are fuzzy; in particular, reference in language is largely discrete (Frege, 1892). We use language to refer to entities 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.[4] This is in contrast to conceptual aspects of meaning.

To sum up, this overview suggests that language is indeed both discrete and continuous; and that there is no neat discrete/continuous divide, nor any area of language that is completely discrete or completely continuous. At the same time, there *are* clearly areas that are more discrete (such as grammar) and areas that are more continuous (such as word meaning). On the other hand, largely because of methodological limitations, most linguistic formalisms to date continue to be discrete.[5] Given the properties of language just discussed, and the fact that, as discussed in the introduction, neural networks afford the potential for both continuous and near-discrete behavior, we can expect LLMs to exploit this potential. And this is indeed what recent literature on interpretability suggests. In what follows, I will focus on providing evidence of near-discrete behavior, as continuous behaviors are already widely recognized in the field (e.g., in the literature on word embeddings, both static and contextualized). Moreover, I will focus mainly on morphosyntax, an area that has received considerable attention in the interpretability literature.

3 Near-discrete language processing in deep learning models

Figure 2 schematically illustrates the contrast between symbolic formalisms and deep learning architectures regarding syntactic processing: while symbolic formalisms are entirely discrete, neural networks afford both continuous and near-discrete processes. However, what counts as near-discrete behavior in the context of neural networks? In my

[3] "Any linguistic pattern is recognized as a construction as long as some aspect of its form or function is not strictly predictable from its component parts or from other constructions recognized to exist. In addition, patterns are stored as constructions even if they are fully predictable as long as they occur with sufficient frequency." (Goldberg, 2005, p. 5)

[4] As of 2025, there are at least two Elizabeth Blackburns: a Nobel laureate and a judge in Florida.

[5] I should note that there have been several developments in integrating a probabilistic component, especially in semantics and pragmatics (see Erk, 2022, for an overview). Computational linguistics has also participated in the debate; for instance, researchers in the field have explored the combination of symbolic and distributed approaches to semantics, building on their complementary strengths and weaknesses (see Boleda and Herbelot, 2016). However, I think it is fair to say that these efforts have not as yet succeeded in providing a unified linguistic framework that encompasses the phenomena reviewed in this section.

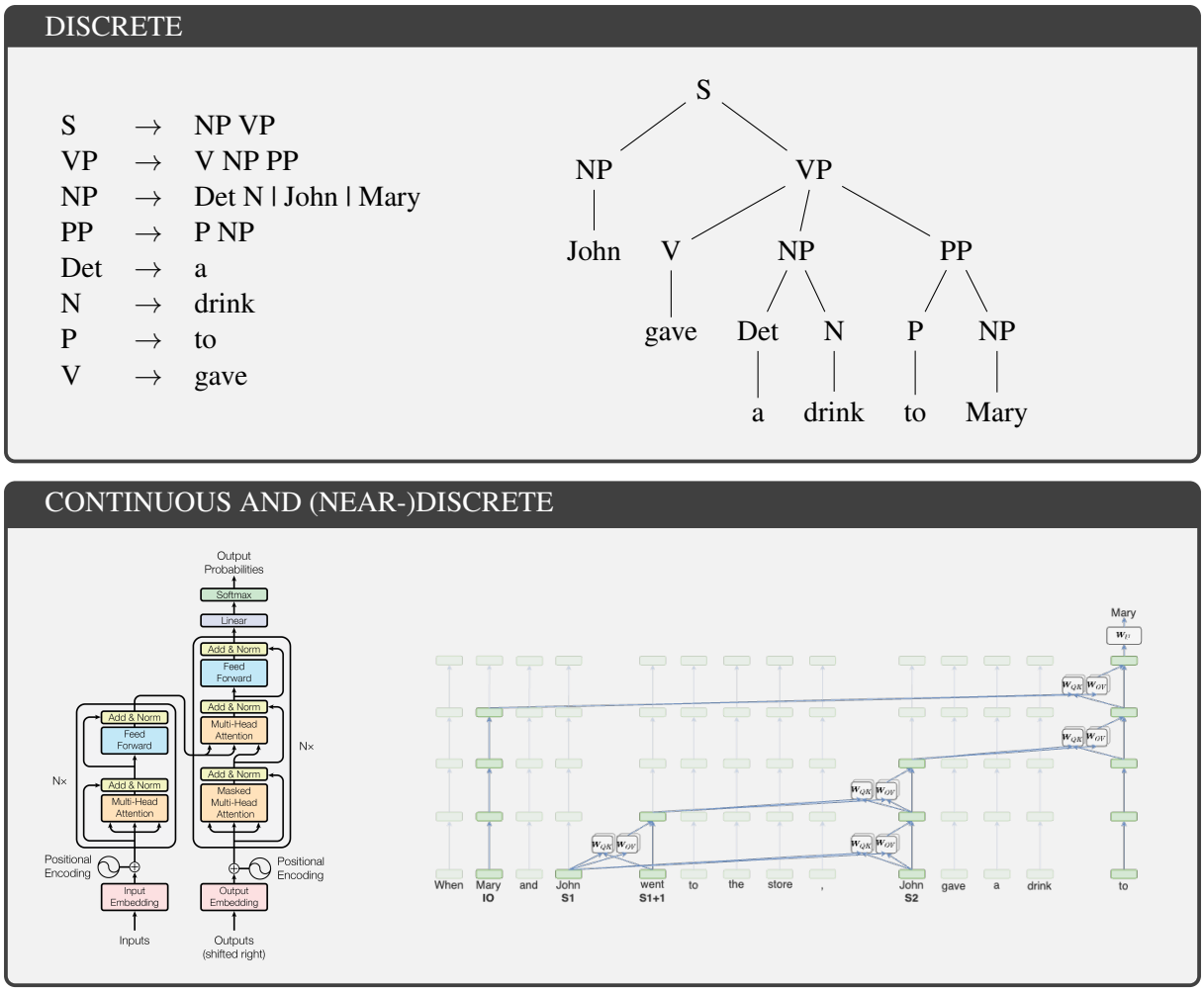

Figure 2: Schematic illustration of the contrast between symbolic formalisms and deep learning. Top: context-free grammar and parse tree for the sentence "John gave a drink to Mary". Bottom: transformer architecture and circuit for the fragment "When Mary and John went to the store, John gave a drink 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).

view, it is the existence of a small sub-unit of the network that is causally involved in encoding or processing a single piece of linguistic information in an interpretable fashion.[6]

An illustrative example is Bau et al. (2019), who identified individual neurons associated to specific morphosyntactic properties in a neural Machine Translation model from the pre-transformer era. Altering the values of these neurons changes the morphosyntactic properties of the translations. For example, in (8) modifying the activation of a single neuron in the representation of the token “supported” changes the tense of the French translation from past (“a appuyé”) to present (“appuie”). Similarly, in (9), altering the activation of a single neuron changes the translation into Spanish from feminine to masculine.[7] Larger sub-units can also manifest near-discreteness, such as attention heads and what has been called “circuits” (subgraphs within neural networks; Cammarata et al., 2020).

(8) The committee supported the efforts of the authorities
Original: Le Comité a appuyeé les efforts des autorités
Modified: Le Comité appuie les efforts des

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

[7]Remarkably, both are potentially correct translations, but the latter has a narrower meaning in which “party” must refer to a political party.

autorités

(9) The interested parties
Original: Las partes interesadas
Modified: Los partidos interesados

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 mechanistic 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.[8] This literature provides robust evidence for near-symbolic representation and processing of both morphosyntactic properties (e.g. part of speech, number, gender, and tense) and syntactic relations (dependencies and agreement).

As for individual 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), as showcased in examples (8-9) above. As another example, Durrani et al. (2023) find neurons sensitive to part of speech in three multi-lingual LLMs (BERT, RoBERTa, and XLNet); for instance, neuron 624 in layer 9 of RoBERTa responds 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.

If the work reviewed up to here focuses on neurons that detect input properties, other studies look at the effects of specific neurons on the output. Geva et al. (2022) 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"). Ferrando et al. (2023) identify 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.

Attention heads specializing in specific syntactic relations have also been amply shown to be present 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 head 7 in layer 6 in BERT for the sentence "many employees are working at its giant Renton, Walsh, 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. Not all heads are near-discrete; Figure 3(b) depicts a head with a broad attention pattern.

As for circuits,[9] which have only recently gained attention, a particularly relevant example in the context of our paper is Wang et al. (2023). This study describes in detail a circuit in GPT2-small that governs the prediction of the indirect object of a sentence. Figure 2 (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 that have specialized functions: 1) Duplicate Token Heads perform duplicate token detection by attending to the duplicate token and writing its position into another head; 2) S-Inhibition Heads remove the duplicate from Name Mover Heads by inhibiting the attention of these heads to the duplicate token; and 3) Name Mover Heads output the re-

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

[9] Definition of "circuit" in Olah et al. (2020): "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 [...]".

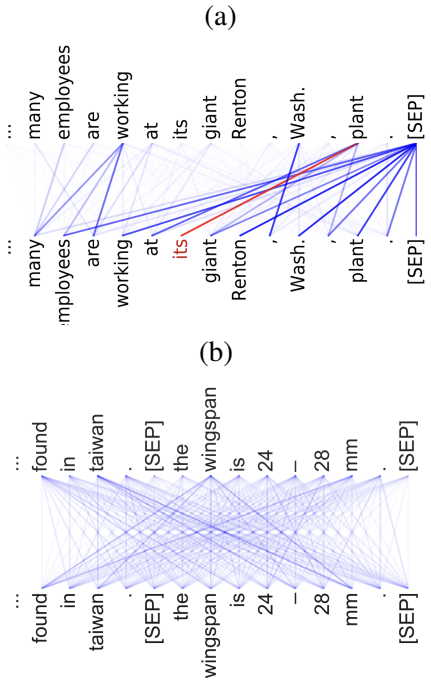

Figure 3: Near-discrete and continuous attention heads in BERT (adapted from Clark et al. (2019); line thickness is proportional to amount of attention). (a) Head 7 in layer 6 tracks dependencies between possessive determiners and their head nouns dependency in a near-discrete fashion: the determiner "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.

maining name by attending to previous names in the sentence and copying the name they attend to (since S-Inhibition Heads inhibit attention to the duplicate token "John", this name will be "Mary" in the example).

Merullo et al. (2024) 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 on 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 evidence to this effect. 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 mechanistic interpretability literature provides evidence for near-discreteness in syntactic processing in different sub-units of LLMs (neurons, attention heads, circuits). However, as discussed in Section 2, discreteness in language goes well beyond syntax, and is present in domains such as compositional semantics and phenomena at the syntax-semantic interface. These domains have received much less attention so far, but the existing evidence tentatively also points towards near-discreteness. 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 already mentioned "universal neurons" in Gurnee et al. (2024) selectively responds to negation.[10]

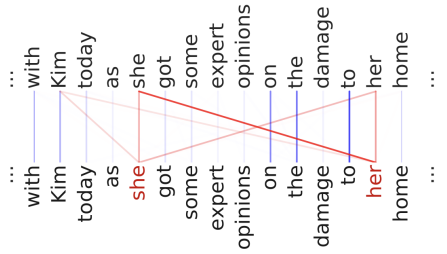

Figure 4: BERT's attention head tracks co-reference dependencies (head 5 in layer 4); adapted from Clark et al. (2019). The anaphoric pronoun "her" sharply attends to antecedent "she".

4 Discussion: LLMs as a synthesis

The previous section has discussed near-discrete encoding and processing of linguistic information in LLMs. However, as mentioned in the introduction, deep learning models can flexibly switch between discrete and distributed modes —and everything in between (see near-discrete vs continuous attention in Figure 3). 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

[10]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 discrete behavior in linguistic representations and processing, but of course its emergence in learning is an exciting topic for further study.

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 there is a high degree of redundancy in neural networks (Durrani et al., 2023). For instance, when Wang et al. (2023) ablated the Name Mover 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 Name Mover 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 many 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. 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 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 and compositional semantics.

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. The advent of deep learning models has added fuel to this debate, with some of its participants continuing to take sides for one or the other with maximalist positions that are, in my view, sterile. Luckily, many scholars are instead increasingly focusing on the huge possibilities that these models bring to the table in terms of advancing scientific knowledge (Manning, 2015; Warstadt and Bowman, 2022; Futrell and Mahowald, 2025). In this article, I have joined this latter camp, putting forth the view that LLMs are a synthesis between the two approaches with respect to how they represent and process language.

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.

Limitations

I am aware that my definition of what counts as near-discreteness 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 the role of quasi-symbolic language processing in LLMs more precisely is an exciting avenue for further work.

Acknowledgments

I used Generative AI to assist with latex formatting in the preparation of this paper.

References

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.

Joan Bresnan. 1982. The mental representation of grammatical relations.

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.

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.

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(1):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.

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.

Mor Geva, Avi Caciularu, Kevin Wang, and Yoav Goldberg. 2022. 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 Goldberg. 2005. *Constructions at Work: The Nature of Generalization in Language*. Oxford University Press.

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

H. P. Grice. 1989. *Studies in the Way of Words*. Harvard University 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.

Randy Allen Harris. 1993. *The Linguistics Wars*. Oxford University Press.

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.

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.

George Lakoff. 1987. *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*. CSLI, Chicago.

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.

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.

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

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*.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. In *Proceedings of ICLR Workshop*.

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*.

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.

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.

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, James L McClelland, PDP Research Group, et al. 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.

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

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.

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.

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.

Stephen Wechsler and Larisa Zlatić. 2003. *The many faces of agreement*.

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.