International Journal of Educational Sciences and Arts (IJESA)
Vol. (4), No. (1)
IJESA
January 2025
المجلة الدولية للعلوم التربوية والآداب
الإصدار (4)، العدد (1)

Understanding Syntax in Large Language Models: Successes and Limitations

Mohammed Mahmoud Alhilal
PhD. In Linguistics, English Language Department, King Faisal University, Hofuf, Kingdom of Saudi Arabia
malhilal@kfu.edu.sa, alhilal8979@gmail.com
ORCID: https://orcid.org/0000-0003-4089-1920

Abstract

This study investigates the ability of Large Language Models (LLMs) to process complex syntactic phenomena, including relative clauses, wh-movement, and center-embedding. By analyzing examples derived from linguistic literature, the study highlights both the strengths and limitations of LLMs in handling syntax. The results reveal that while LLMs exhibit competence in simpler syntactic constructions, they struggle with deeper hierarchical dependencies and abstract syntactic constraints. The study underscores the need for integrating explicit syntactic principles into LLM architectures to bridge the gap between surface-level fluency and generative linguistic competence.

Keywords: Syntax, Large Language Models, Natural Language Processing, Syntactic Competence, Neural Language Models, Computational Linguistics.

Introduction

Syntax, the study of sentence structure and the rules governing the arrangement of words and phrases, is a cornerstone of linguistic theory. As Chomsky (1965) asserted, syntax is central to understanding the generative capacity of human language, describing it as "the set of principles and processes by which sentences are constructed in particular languages" (p. 15). The study of syntax has long been

fundamental to linguistic inquiry, providing insights into how meaning is shaped by structure.

In recent years, Large Language Models (LLMs) such as OpenAI's GPT series and Google's BERT have revolutionized natural language processing (NLP). These models leverage vast amounts of data and advanced machine learning architectures to generate human-like text, respond to queries, and complete sentences with remarkable fluency. However, despite their apparent mastery of language, questions remain regarding their ability to accurately model syntactic structures. Linguists have raised concerns about whether LLMs possess a true understanding of syntax or simply approximate it through statistical pattern recognition (Manning et al., 2020). For instance, the inability of some models to consistently handle syntactic ambiguities or recursive structures suggests limitations in their syntactic competence.

This article examines how well LLMs handle complex syntactic phenomena, including relative clauses, wh-movement, and center-embedding, as documented in linguistic literature. These phenomena have been extensively studied as test cases for linguistic theory due to their structural complexity and implications for cognitive processing (Hawkins, 2004). By analyzing LLM outputs for examples derived from published syntax research, this study aims to evaluate whether these models align with theoretical predictions or diverge in systematic ways.

Understanding how Large Language Models (LLMs) process syntax is essential for both theoretical and practical purposes. Theoretically, it contributes to discussions about syntactic competence and the degree to which computational systems can emulate human linguistic abilities. Practically, the ability to handle syntax impacts the effectiveness of LLMs in tasks such as machine translation, summarization, and automated question answering. As highlighted by Goldberg (2019), accurately

managing syntax is not merely an academic interest but a fundamental requirement for the successful deployment of NLP systems in practical applications.

This study addresses a fundamental question: How accurately can LLMs handle published examples of complex syntactic phenomena, as described in linguistic theories? By grounding the analysis in established linguistic research, this article provides a comprehensive assessment of the syntactic capabilities of LLMs and their implications for both linguistics and NLP.

Literature Review

The role of syntax in Large Language Models (LLMs) has been a focal point of extensive research, with numerous studies exploring their capacity to manage syntactic structures, dependencies, and constraints. Zhou et al. (2023) assessed LLMs' syntactic competence through natural language questions targeting specific syntactic knowledge points, revealing significant performance discrepancies across different syntactic aspects. This aligns with the present study's aim of evaluating LLMs' handling of complex syntactic phenomena, such as relative clauses and wh-movement. Similarly, Kulmizev and Nivre (2022) examined the implicit syntactic knowledge in neural models, concluding that such knowledge remains shallow—an observation further explored in this research through targeted syntactic constructions.

Building on this, Marvin and Linzen (2018) developed a framework for assessing structure-sensitive phenomena like subject-verb agreement and reflexive anaphora, emphasizing the utility of benchmarks in syntax evaluation. This methodological approach informs the framework of the current study. Van Schijndel, Mueller, and Linzen (2019) analyzed whether increasing training data improves syntactic understanding, concluding that inherent architectural limitations, rather than data quantity, constrain performance—a finding particularly relevant to the observed challenges LLMs face with center-embedding.

To refine syntax evaluation methods, Newman et al. (2021) proposed disentangling broad evaluation objectives into distinct goals, underscoring the need for precision—a principle adopted in this study to assess relative clauses, wh-movement, and center-embedding. In a complementary perspective, Wilcox et al. (2023) employed psycholinguistic stimuli to evaluate neural language models' ability to process syntactic dependencies, offering insights into the models' handling of abstract structural relationships directly relevant to the present research.

Highlighting LLMs' syntactic limitations, Manning et al. (2020) demonstrated their struggles with hierarchical and long-distance dependencies despite their surface fluency, findings further interrogated in this study through systematically selected examples. Goldberg (2019) addressed the practical implications of syntactic errors in NLP applications like machine translation and question answering, underscoring the real-world significance of the syntactic phenomena analyzed here.

Lastly, Shi and Knight (2017) proposed a neural model capable of jointly learning syntax and lexicon, suggesting that structural information could enhance language modeling. Their approach points to potential pathways for addressing LLMs' syntactic shortcomings, which this study discusses in its implications and future directions. Collectively, these contributions establish a foundation for the present investigation into the syntactic competence of LLMs, offering insights into their strengths, limitations, and avenues for improvement.

Theoretical Background

Syntax has long been a central focus of linguistic theory, offering a systematic framework for understanding how words and phrases combine to form grammatically correct sentences. According to Chomsky (1981), syntax encompasses universal principles and language-specific rules that govern sentence structure. Key constructs within syntactic theory include relative clauses, wh-

movement, and center-embedding, all of which are notable for their complexity and their role in human linguistic competence.

- **Complex Syntactic Constructions:**

Relative clauses, which function to modify nouns, are a core topic in syntactic studies. Comrie (1989) emphasized the importance of relative clauses in demonstrating the interplay between syntax, semantics, and discourse. In English, such constructions frequently involve relativizers like who or which, as in "The book that John read was fascinating." These clauses are particularly significant for analyzing syntactic embedding, a defining feature of structural complexity.

Wh-movement, another critical syntactic phenomenon, involves the displacement of constituents to form questions or relative clauses. In sentences such as "What did she say he wanted to buy?", the interrogative pronoun what is moved to the sentence's beginning, leaving a gap in its original position. Chomsky (1977) explained that this process is regulated by constraints like subjacency, which limits the distance over which constituents can move. Wh-movement presents significant challenges for computational models due to its reliance on abstract structural dependencies.

Center-embedding, an extreme form of recursion, features nested clauses and is exemplified by sentences like "The boy the girl liked ran away." Frazier and Fodor (1978) described how such constructions strain cognitive processing, particularly when multiple layers of structure are involved. Gibson (1998) highlighted that working memory constraints exacerbate the difficulty of processing center-embedded sentences, making them challenging benchmarks for both human and machine syntactic competence.

- **Syntax in Computational Linguistics:**

The role of syntax in computational linguistics has evolved considerably over time. Early rule-based systems explicitly incorporated syntactic principles, whereas

modern neural models typically learn implicit representations of syntax from data. Manning et al. (2020) pointed out that while LLMs achieve impressive fluency in surface-level language, their ability to handle complex syntactic relationships often falls short of human competence. This limitation has fueled ongoing efforts to explore how syntactic theory can better inform and evaluate the capabilities of these models.

- **Significance of Complex Syntax:**

Relative clauses, wh-movement, and center-embedding serve as critical test cases for linguistic theories and computational systems alike. They provide opportunities to explore universal syntactic principles and to assess the extent to which models trained on extensive datasets can replicate them. Johnson and Goldberg (2013) stressed that syntactic complexity is not only a theoretical concern but also a practical challenge that tests the boundaries of natural language understanding.

Building on these foundational insights, this study evaluates the performance of LLMs on syntactic phenomena that are both theoretically significant and computationally demanding.

Methodology

This section outlines the approach used to assess how Large Language Models (LLMs) process complex syntactic phenomena, relying on examples drawn from well-established linguistic literature. The methodology prioritizes the use of published syntactic constructs to evaluate model performance, eliminating the need for direct human interaction or custom data collection.

The study focuses on three primary syntactic phenomena: relative clauses, wh-movement, and center-embedding. Examples representing these constructions were carefully chosen from authoritative syntax textbooks and scholarly research to maintain theoretical rigor and relevance. For instance, Comrie (1989) provided a

quintessential example of a relative clause with the sentence, “The book that John read was fascinating,” which effectively illustrates the structural embedding typical of such constructions.

For wh-movement, sentences like “What did she say he wanted to buy?” from Chomsky’s (1977) work were utilized to analyze how LLMs manage syntactic dependencies and the displacement of constituents. These examples are particularly suited to testing whether models adhere to constraints governing syntactic movement, such as maintaining the integrity of hierarchical relationships.

Center-embedding examples, such as “The boy the girl liked ran away,” as discussed by Gibson (1998), were selected to evaluate the models’ ability to process recursive structures. Nested clauses like these are well-documented for their cognitive complexity, offering a robust benchmark for examining how effectively LLMs handle multiple levels of syntactic integration.

By grounding the evaluation in these carefully curated examples, the study ensures that the syntactic phenomena under investigation are theoretically significant and align with well-established linguistic principles.

The evaluation framework employs both qualitative and quantitative methods to analyze the syntactic capabilities of Large Language Models (LLMs). This dual approach ensures a comprehensive assessment of their performance across complex syntactic phenomena**.**

The qualitative evaluation involves comparing LLM-generated outputs with theoretical expectations from established linguistic literature. Success is determined by the model's ability to adhere to grammatical rules and accurately represent syntactic dependencies. Errors are analyzed for recurring patterns, including misplacement of syntactic constituents or the inability to resolve ambiguities

effectively. These patterns provide insights into systematic limitations in the models' handling of syntax.

Quantitatively, the success rate for each syntactic phenomenon is calculated based on the proportion of outputs that are accurate. Different error types are categorized to identify recurring weaknesses in syntactic processing, such as failure to manage hierarchical dependencies or recursive structures.

The testing was conducted using publicly available LLM interfaces, including OpenAI's GPT-4 and Google's BERT, which are recognized for their advanced performance in natural language processing. To minimize variability, syntactic examples were input into the models under controlled conditions. Each syntactic phenomenon was tested as follows:

- Relative Clauses: Incomplete sentences were provided to the models, and their completions were evaluated for grammaticality and syntactic integrity.
- Wh-Movement: Questions were posed to assess whether the models generated responses that were both grammatically correct and semantically appropriate.
- Center-Embedding: Nested constructions were used to evaluate the models' ability to maintain coherence and manage complex recursive structures. The performance of the LLMs was evaluated using the following criteria:
    1. Grammaticality: Whether the output sentences conformed to the syntactic rules outlined in linguistic literature.
    2. Coherence: Whether the generated sentences were interpretable, both semantically and syntactically, particularly in cases involving complex constructions like center-embedding.
    3. Theoretical Alignment: Whether the outputs adhered to theoretical syntactic constraints, such as subjacency as proposed by Chomsky (1977) and locality principles as discussed by Gibson (1998).

This study relies on resources such as syntax textbooks (e.g., Chomsky, 1981; Comrie, 1989) and datasets specifically designed for syntactic evaluation, including the Benchmark of Linguistic Minimal Pairs (BLiMP). These resources provide a robust foundation for assessing the syntactic competence of LLMs, ensuring that the findings are both linguistically grounded and computationally relevant.

By anchoring the methodology in well-documented examples and theoretical principles, this study provides a reliable framework for evaluating the syntactic performance of LLMs.

Results

This section details the findings of the study, focusing on the performance of Large Language Models (LLMs) in processing complex syntactic phenomena. The results are organized by the three selected phenomena—relative clauses, wh-movement, and center-embedding—and categorized into successes, failures, and recurring error patterns.

Relative clauses, which embed a clause within a noun phrase, were a relatively manageable aspect of syntax for LLMs. For simpler constructions, such as "The book that John read was fascinating," both GPT-4 and BERT generated grammatically accurate outputs. When tasked with producing or completing sentences involving restrictive and non-restrictive relative clauses, the models generally adhered to syntactic rules. For instance, when prompted to generate a sentence with a non-restrictive clause, GPT-4 successfully produced: "The author, who won the Nobel Prize, is giving a lecture."

However, the models struggled with more complex sentences involving multiple embeddings or long-distance dependencies, such as "The car that the mechanic who lives in the city fixed is red." Common errors included truncation of sentences or

failure to preserve correct syntactic relationships, indicating that the models encounter difficulty with increased structural complexity.

Wh-movement posed a moderate challenge for LLMs. In straightforward cases, such as "What did she say he wanted to buy?" the models displayed high accuracy, resolving the moved element appropriately and maintaining grammatical correctness. For example, GPT-4 correctly responded: "She said he wanted to buy a book. What did she say he wanted to buy?"

Nevertheless, in more complex instances involving multiple wh-phrases or syntactic constraints (such as island constraints), the models exhibited inconsistencies. For example, in sentences like "What did the manager claim that the assistant forgot to mention?", the models occasionally failed to resolve the syntactic dependencies accurately, resulting in incomplete or ungrammatical outputs. These challenges align with Manning et al.'s (2020) observation that LLMs often face difficulties when processing hierarchical structures and long-distance dependencies.

Center-embedding, a recursive syntactic structure exemplified by sentences like "The boy the girl liked ran away," posed the most significant challenge for the models. While they were capable of handling single-level embeddings, their performance deteriorated as the depth of embedding increased. For example:

- In simpler center-embedding cases, such as "The man the woman admired left the room," GPT-4 produced correct and coherent responses.
- However, for sentences with more than two levels of embedding, like "The dog the cat the boy saw chased ran away," the models frequently generated ungrammatical or incomplete sentences. This performance decline reflects the cognitive and computational difficulty associated with recursive structures.

These findings are consistent with Gibson's (1998) theory of working memory constraints, which posits that increased embedding imposes a significant cognitive

load, making such constructions challenging even for humans. The limitations observed in the models underscore the need for more explicit training to address recursion and hierarchical structure.

Quantitative Summary

A summary of the success rates across the three phenomena is presented in **Table 1** below:

Phenomenon	Success Rate	Common Errors
Relative Clauses	85%	Mismanagement of deeply embedded structures
Wh-Movement	75%	Island constraint violations, unresolved gaps
Center-Embedding	50%	Unfinished sentences, ungrammatical outputs

Analysis of errors revealed systematic patterns:

1. Truncation: Sentences involving deeply embedded structures were often truncated before completion.
2. Unresolved Dependencies: In wh-movement, gaps were sometimes left unresolved, leading to ungrammatical outputs.
3. Semantic Incoherence: Center-embedding errors often involved semantically incoherent sentences, indicating a failure to integrate syntactic and semantic information.

Discussion

The findings of this study offer valuable insights into both the strengths and limitations of Large Language Models (LLMs) in processing complex syntactic phenomena. By evaluating LLM performance on relative clauses, wh-movement, and center-embedding constructions, the discussion explores key theoretical implications for linguistics as well as practical considerations for natural language processing (NLP). The results show that LLMs demonstrate a significant degree of competence in managing simpler syntactic structures, such as single-layer relative

clauses and basic instances of wh-movement. This suggests that these models are capable of internalizing many surface-level syntactic patterns that are common in their training data. For example, the study found that models correctly generated relative clause constructions, including examples like "The book that John read was fascinating," reflecting their ability to follow established grammatical rules. Similarly, in handling wh-movement, the models accurately resolved syntactic dependencies in straightforward sentences such as "What did she say he wanted to buy?"

These successes point to the ability of LLMs to approximate syntactic rules through pattern recognition. This proficiency is likely attributed to their exposure to extensive linguistic data during training, enabling them to replicate and generalize common syntactic structures with notable fluency and accuracy.

- **Limitations of LLMs in Complex Syntax (Paraphrased):**

Despite their notable successes with simpler syntactic structures, Large Language Models (LLMs) struggled as structural complexity increased. Center-embedding constructions, in particular, presented significant challenges, with accuracy rates declining to approximately 50% for deeply embedded structures. This finding aligns with Gibson's (1998) theory, which attributes the difficulty of processing center-embedding to working memory constraints. Common errors in these cases, such as truncation and incoherence, highlight the limitations of LLMs in representing hierarchical syntactic dependencies.

Similarly, LLMs showed weaknesses in resolving syntactic constraints related to wh-movement, such as violations of subjacency. In some instances, the models failed to adhere to rules governing long-distance dependencies, resulting in outputs that were either ungrammatical or incomplete. These challenges underscore the difficulty LLMs face in processing abstract syntactic principles that demand a nuanced understanding of structural relationships.

- **Theoretical Implications:**

The findings contribute to ongoing debates about the nature of syntactic competence in LLMs. While these models exhibit a notable ability to produce grammatically plausible sentences, their struggles with complex syntactic phenomena suggest that their understanding of syntax is largely superficial. Manning et al. (2020) argue that LLMs, while adept at approximating linguistic patterns, often fail to grasp the deeper generative principles that underpin language. This raises important questions about the extent to which current neural architectures can replicate human linguistic competence and whether incorporating explicit syntactic theory into these models is necessary.

The results also challenge assumptions about the sufficiency of large datasets for training LLMs. Although extensive corpora allow these models to generalize syntactic patterns effectively, their struggles with recursion and long-distance dependencies indicate that training data alone may not fully capture the complexities of syntax.

- **Practical Implications:**

The limitations observed in this study have significant implications for the application of LLMs in real-world NLP tasks. Functions such as machine translation, question answering, and text summarization rely heavily on accurate syntactic processing. Errors in handling complex constructions, including center-embedding and wh-movement, can lead to misinterpretations or reduce the reliability of these systems. Goldberg (2019) emphasized that achieving syntactic accuracy is a critical prerequisite for deploying NLP systems effectively in practical scenarios.

Addressing these challenges may require hybrid approaches that combine data-driven methodologies with explicit syntactic modeling. For example, integrating linguistic parsers or rule-based systems into LLM architectures could enhance their

ability to manage complex syntactic structures. Such approaches could bridge the gap between the surface-level fluency of LLMs and the deeper syntactic competence required for robust linguistic performance.

- **Future Directions:**

The findings of this study highlight several promising avenues for future research. One direction involves incorporating explicit syntax into LLM architectures by integrating syntactic parsers to enhance their ability to manage hierarchical dependencies. Another avenue is conducting cross-linguistic comparisons to evaluate how LLMs handle syntactic phenomena across languages with varying typological characteristics. Finally, refining training methods by developing targeted strategies that expose models to a broader range of complex syntactic structures offers potential for improving their syntactic competence. These directions provide pathways for addressing the limitations observed in current LLMs and advancing their syntactic capabilities. By addressing these areas, future research can advance both the theoretical understanding and practical capabilities of LLMs in handling syntax.

Conclusion

This study examined the performance of Large Language Models (LLMs) on three complex syntactic phenomena: relative clauses, wh-movement, and center-embedding. By analyzing examples from established linguistic literature, the study highlighted both the strengths and limitations of these models in handling syntax. The findings reveal that LLMs exhibit considerable competence in processing simple and moderately complex syntactic constructions, such as straightforward relative clauses and basic wh-movement. These successes underscore the models' ability to generalize syntactic patterns from large training corpora. However, significant limitations emerged in more complex scenarios, particularly in handling deeply embedded structures and long-distance dependencies. The difficulties encountered

with center-embedding and syntactic constraints, such as subjacency, indicate that LLMs lack a robust internal representation of hierarchical syntactic principles.

From a theoretical standpoint, the findings add to the ongoing discussion about syntactic competence in Large Language Models (LLMs). Although these models display remarkable fluency, their difficulties with abstract and recursive structures indicate that their understanding of syntax relies more on pattern recognition than on generative principles. As Manning et al. (2020) pointed out, LLMs are adept at replicating linguistic patterns but often struggle to grasp the deeper generative rules that underlie natural language.

Practically, these findings have implications for the deployment of LLMs in natural language processing tasks. Applications such as machine translation, summarization, and question answering require precise handling of syntactic structures to ensure accuracy and coherence. Addressing the observed limitations may involve integrating explicit syntactic knowledge into neural architectures or enhancing training methods to expose models to a wider variety of complex syntactic phenomena.

In conclusion, while LLMs represent a significant advancement in natural language processing, their handling of complex syntactic phenomena remains an area requiring further exploration. Future research should focus on hybrid approaches that combine data-driven methods with linguistic theory to create models capable of both surface-level fluency and deep syntactic competence. By bridging this gap, we can advance not only the performance of LLMs but also our understanding of the interface between computational models and human linguistic knowledge.

Summary

This study investigates the ability of Large Language Models (LLMs) to process complex syntactic phenomena, focusing on relative clauses, wh-movement, and

center-embedding. Using examples from established linguistic literature, the study highlights both the strengths and limitations of LLMs in handling syntax. The findings reveal that LLMs demonstrate significant competence with simple and moderately complex syntactic constructions, such as basic relative clauses and straightforward wh-movement. These successes reflect their capacity to generalize syntactic patterns from large training corpora. However, their performance declines with increased structural complexity, particularly in deeply embedded or recursive structures like center-embedding and in adhering to syntactic constraints such as subjacency. These limitations suggest that LLMs rely on surface-level pattern recognition rather than a deep, generative understanding of syntax. Theoretically, the results contribute to discussions about syntactic competence in computational systems, emphasizing the need for models to incorporate explicit linguistic principles to address their shortcomings. Practically, the study highlights the importance of accurate syntactic processing for NLP tasks such as translation and summarization, where errors in complex syntax can reduce system reliability. The research concludes by advocating for hybrid approaches that integrate data-driven methods with linguistic theory to enhance LLMs' syntactic capabilities and bridge the gap between computational models and human linguistic knowledge.

Further Research

The findings of this study open several avenues for future research into the syntactic capabilities of Large Language Models (LLMs). One promising direction involves integrating explicit syntactic theory into neural architectures. Investigating how linguistic frameworks, such as the Minimalist Program (Chomsky, 1995) or dependency grammar, can enhance the models' handling of hierarchical and recursive structures could address some of the observed limitations. Another area for exploration is the cross-linguistic performance of LLMs. While this study focused on English syntax, extending the analysis to other languages with varying syntactic

properties—such as free word order languages or those with extensive agreement systems—could provide insights into the universality and adaptability of LLMs' syntactic competence. Additionally, future research could involve designing more linguistically focused benchmarks. Current evaluation datasets, while useful, may not sufficiently test deeper syntactic principles. Developing benchmarks based on theoretical constructs, such as syntactic islands or scope ambiguity, could provide more robust evaluations. Lastly, understanding the relationship between syntactic capabilities and downstream task performance warrants further investigation. Examining how specific syntactic weaknesses affect tasks like machine translation, summarization, or dialogue systems can inform practical improvements and identify critical areas for enhancement.

Research Limitations

This study has several limitations that should be addressed in future research. First, the analysis relied exclusively on examples from English linguistic literature. While these examples are widely studied and provide a solid foundation, they may not capture the full range of syntactic diversity found in natural languages. Extending the study to other languages would provide a more comprehensive understanding of LLMs' syntactic competence. Second, the evaluation focused on selected phenomena—relative clauses, wh-movement, and center-embedding—without exploring the broader spectrum of syntactic constructs. While these phenomena are critical test cases, future studies could expand the scope to include additional structures, such as passive voice, negation, or coordination. Third, the study was limited to publicly available LLMs, such as GPT-4 and BERT, and did not consider the effects of model size, architecture, or training data composition in detail. Variations in these factors may significantly impact performance and should be explored further. Finally, the evaluation relied on qualitative and quantitative analyses of LLM outputs without delving into the underlying representations or

mechanisms used by the models. Employing techniques such as probing tasks or attention visualization could shed light on how LLMs encode syntactic structures internally. Addressing these limitations will help refine our understanding of LLMs and guide efforts to improve their syntactic capabilities, both theoretically and practically.

References

- Chomsky, N. (1965). Aspects of the Theory of Syntax. MIT Press.
- Chomsky, N. (1977). On wh-movement. In P. W. Culicover, T. Wasow, & A. Akmajian (Eds.), Formal Syntax (pp. 71-132). Academic Press.
- Chomsky, N. (1981). Lectures on Government and Binding: The Pisa Lectures. Foris Publications.
- Comrie, B. (1989). Language Universals and Linguistic Typology: Syntax and Morphology. Blackwell.
- Frazier, L., & Fodor, J. D. (1978). The sausage machine: A new two-stage parsing model. Cognition, 6(4), 291-325. https://doi.org/10.1016/0010-0277(78)90002-1
- Gibson, E. (1998). Linguistic complexity: Locality of syntactic dependencies. Cognition, 68(1), 1-76. https://doi.org/10.1016/S0010-0277 (98)00034-1.
- Goldberg, Y. (2019). Neural Network Methods for Natural Language Processing. Morgan & Claypool Publishers.
- Hawkins, J. A. (2004). Efficiency and Complexity in Grammars. Oxford University Press.
- Kulmizev, A., & Nivre, J. (2022). Schrödinger's tree\u2014On syntax and neural language models. Frontiers in Artificial Intelligence, 5, 796788. https://doi.org/10.3389/frai.2022.796788.
- Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U., & Levy, O. (2020). Emergent linguistic structure in artificial neural networks trained by self-supervision. Proceedings of the National Academy of Sciences, 117(48), 30046-30054. https://doi.org/10.1073/pnas.1907367117.
- Marvin, R., & Linzen, T. (2018). Targeted syntactic evaluation of language models. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing (pp. 1192-1202). Association for Computational Linguistics. https://aclanthology.org/D18-1151/.

- Newman, B., Ang, K.-S., Gong, J., & Hewitt, J. (2021). Refining targeted syntactic evaluation of language models. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 3710-3723). Association for Computational Linguistics. https://aclanthology.org/2021.naacl-main.290/.
- Shi, H., & Knight, K. (2017). Neural language modeling by jointly learning syntax and lexicon. arXiv preprint arXiv:1711.02013. Retrieved from https://arxiv.org/abs/1711.02013.
- Van Schijndel, M., Mueller, A., & Linzen, T. (2019). Quantity doesn\u2019t buy quality syntax with neural language models. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (pp. 5831-5837). Association for Computational Linguistics. https://aclanthology.org/D19-1592/.
- Wilcox, E., Levy, R., Morita, T., & Futrell, R. (2023). Neural networks as cognitive models of the processing of syntactic dependencies: An evaluation using controlled psycholinguistic stimuli. Open Mind: Discoveries in Cognitive Science. Advance online publication. https://doi.org/10.1162/opmi_a_00137.
- Zhou, H., Hou, Y., Li, Z., Wang, X., Wang, Z., Duan, X., & Zhang, M. (2023). How well do large language models understand syntax? An evaluation by asking natural language questions. arXiv preprint arXiv:2311.08287. Retrieved from https://arxiv.org/abs/2311.08287.