Lark Journal of Philosophy, Linguistics, and Social Sciences

E-ISSN 2663-5836
P-ISSN 1999-5601
Original Research

Journal homepage: https://lark.uowasit.edu.iq

Vol. 18 No.3 July 2026, pp.792-802

https://doi.org/10.31185/lark.5400

A Comparative Analysis of Language Generation Mechanisms in Cartesian Universal Grammar and Transformer-Based AI Models

M.M. Raqib Imad Jassim
University of Wasit, College of Arts
Gheahebaltamimi@gmail.com

Received Feb1, 2026 Revised Apr29, 2026 Accepted Apr 29, 2026 Online Jul.1, 2026

ABSTRACT

This paper provides a strict comparative analysis of two fundamentally opposed paradigms towards the perception and creation of human language: the nativist, computational-hierarchical model (Cartesian Universal Grammar) and the empiricist, statistical-associative model (Transformer-based LLMs). The objective of this study is to evaluate the explanatory depth of UG against the predictive power of LLMs to determine if statistical models can truly replicate the discrete infinity of human language. Using a qualitative comparative methodology, the analysis applies a four-part framework: Computational Primitive, Representational Structure, Source of Knowledge, and Explanatory Scope. Results show that while UG offers a principled account of linguistic competence, LLMs excel in performance, suggesting a potential hybrid model for future linguistic theory.

Keywords: UG, Cartesian Universal Grammar, MP, Transformer-Based AI Models

تحليلٌ مقارن لآليات توليد اللغة في النحو الكلّي الديكارتي ونماذج الذكاء الاصطناعي القائمة على بنية المحوّلات

م.م رقيب عماد جاسم
جامعة واسط / كلية الآداب
Gheahebaltamimi@gmail.com

الملخص

تقدم هذه الورقة البحثية تحليلًا مقارنًا دقيقًا لنموذجين متناقضين جوهريًّا لفهم اللغة البشرية وإنشائها: النموذج الفطري الحسابي الهرمي (القواعد النحوية العالمية الديكارتية) والنموذج التجريبي الإحصائي الترابطي (نماذج اللغة القائمة على المحولات). ويهدف هذا البحث إلى تقييم العمق التفسيري للقواعد النحوية العالمية مقابل القدرة التنبؤية لنماذج اللغة، وذلك لتحديد ما إذا كانت النماذج الإحصائية قادرة على محاكاة التعددية اللانهائية للغة البشرية. باستخدام منهجية مقارنة نوعية، يطبق التحليل إطارًا رباعي الأجزاء: العنصر الحسابي الأساسي، والبنية التمثيلية، ومصدر المعرفة، والنطاق التفسيري. تُظهر النتائج أن القواعد النحوية العالمية، على الرغم من أنها تقدم تفسيرًا مبدئيًّا للكفاءة اللغوية، إلا أن نماذج اللغة القائمة على المحولات تتفوق في الأداء، مما يشير إلى نموذج هجين محتمل لنظرية لغوية مستقبلية.

الكلمات المفتاحية: القواعد النحوية العالمية، القواعد النحوية العالمية الديكارتية، نموذج المحاكاة، نماذج الذكاء الاصطناعي القائمة على المحولات

1. Introduction

The technological development that the world witnessed, after the Fourth Industrial Revolution, opened the door for major companies and media platforms in the world to move towards investing in artificial intelligence technology, and employing it in the field of digital media (Alsaray & Altimimiy, 2023, p.1). Since the ability of human language, its infinite creativity, its order and arrangement, and its easy learning by children- has long been viewed as the mark of our species. The leading theoretical model that has been used to explain this capacity over the past half-century has been Universal Grammar (UG), a rationalist tradition derived out of Cartesian philosophy and formalized by Noam Chomsky. UG assumes that the fundamental principles of language are inbuilt and they constitute a special faculty of the human mind called language (Chomsky, 2002, p. 118).

This faculty has a generative engine which is a basic recursive process that assembles elaborate structures of language out of simple components. Conversely, the last ten years have seen the inception of Transformer-based Large Language Models (LLMs), with GPT and BERT, that have established more than ever before the skill to generate language confidently and logically. These models are virtually numerical and data-driven, exercise representations and relations on vast groups of human text. Their success puts into question the principles of UG, which posits that complex language behavior may be an emergent phenomenon due to massive scale and highly specialized statistical learning, as opposed to a consequence of a specialized, coded module.

This paper is attempting to compare these two radically different language generation methods. The point is not only that they worked out and contrasted the mechanisms on which the two paradigms basing their statements rest, but that they critically analyzed those mechanisms. The paper seeks to go beyond the top-level discussion about the nativism and empiricism to question the computational primitives each system is based on. Particularly, it makes a comparison between the Merge of UG, which is the only structure-building mechanism in the Minimalist Programme, and the Self-Attention mechanism of the Transformer the essence of creating meaning that allows it to understand and generate context.

The research questions that will guide this analysis will be:

1. What are the essential differences between the language-generation processes of Cartesian Universal Grammar (i. e. Merge) and Transformer-Based AI Models (i. e. Self-Attention)?
2. What do these differences tell us about the nature of language, and especially about the discrete infinity and structure-dependence of human language?

Moreover, in this study 'Merge' is defined as the binary operation that combines two syntactic objects into a single set. 'Self-Attention' is the mechanism that calculates the relevance of each token

in a sequence relative to others using Query, Key, and Value vectors. 'Competence' refers to internalized linguistic knowledge, while 'Performance' refers to the actual use of language in context.

2. Literature Review

The intellectual climate of language generation is currently divided into two powerful traditions which appear to be mutually exclusive. To make a meaningful comparison, there is no way one can do without a careful understanding of their historical underpinnings and details of their computations.

2.1Cartesian/Nativist Tradition and UG

Universal Grammar builds upon the Cartesian tradition in linguistics, which gives priority to the rationalist notion that the human mind is initially endowed with innate ideas and forms that exist before experience (Chomsky, 2002, p. 206). This position was formalized by Chomsky in the first volume. Besides, syntactic structures, argue that the speed and uniformity with which children learn language, despite the poverty of the stimulus (weakness and lack of richness of the input data) requires an innate species-specific endowment. This gift is what makes up the UG (Chomsky, 1957, p. 118). The development of UG reached the Minimalist Program (MP), which aims at bringing the complexity of the language faculty down to the simplest possible computational system (Chomsky, 2014, p. 420). The key principle of the MP is that the human language faculty (HLF) is an ideal solution to the condition of interface with sensorimotor manufacturing system to externalization, and the conceptual-intentional system to thought. The exclusive structure-building operation in the MP, which is called merge, is a binary set-formation operation that inputs two syntactic objects, X and Y and forms a new unordered set {X, Y).

More importantly, Merge is recursive: the output of one application may be used as an input of another application. This recursive property is the formal apparatus that produces the discrete infinity of the human language, the ability to form an infinite number of well-formed sentences using a finite number of elements. An example is the word the man, which can be defined as a Merge (the, man); the word the man who saw the dog, which comes out as a result of a sequence of Merge operations in which the relative clause the man who saw the dog is embedded, hence demonstrating the unlimited embedding and hierarchical structuring ability.

Accordingly, the generative ability of UG is essentially combinatorial and hierarchical, an expression of an entrenched belief in the idea that language is a rule system, which creates structure, but not a set of acquired associations. The UG paradigm is mainly focused on competence the idealized, internalized knowledge of language, and not on performance, which deals with the actual application of language in real life contexts (Chomsky& et. al. , 2023, p. 192).

2. 2 The Tradition Empiricist/Connectionist and Transformer Models

Unlike the paradigm of the UG, the empiricist tradition understands the concept of language as something that is constructed during the interaction between a person and the environment, and the generalization between statistics is made due to the presence of data. This point of view found its modern implementation in statistical Natural Language Processing (NLP), which used models like N -grams and Hidden Markov Models to give forecasts about word sequences in terms of frequency distributions (Manning & Schutz, 1999, p. 680). However, these models failed to work once long-range dependencies and subtle contextual insights were taken into consideration. Transformer architecture: the beginning of the Transformer architecture in the 2017 seminal object Consideration "Is All You Need"(Manning & Schutz, 1999, p. 680). This transformed the background of language modeling forever. The Transformer replaced sequential execution of Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) with the Self-Attention mechanism which attends to all the input tokens simultaneously. This computation primitive enables one model to include the weight of each other token in the input sequence in determining a given token. The Self-Attention mechanism works through the allocation of every token with three learned vectors, namely Query (Q) and Key (K) and Value (V). The similarity between two tokens is calculated as a dot product between the Query of a given token and the Key of the other one, which is after that subjected to a SoftMax to produce a probability distribution. This score dictates how a token is centered on another during processing, and the ultimate representation of the token comes out as a weighted total of the related Value vectors, weighted by the attention scores (Vaswani et al. , 2017, Pp. 5998-6008).

In addition, generation of language using large language models (LLMs) generally follows an autoregressive-style, i. e. , the model predicts the next token based on the past tokens and the acquired correlations of its training corpus. The special presentation of LLMs is also clarified by the Scaling Hypothesis, which shapes that the presentation increases predictably with the development of the model size, data size, and computing properties. As a result, the main characteristic of the performance-oriented systems is that LLMs are capable of showing remarkable fluency and coherence in terms of the natural language usage (Kaplan & et. al. , 2020, Pp. 1-23).

2.3 UG vs. LLM Debate: New Critiques and Reactions

The unequivocal triumph of large language models (LLMs) has given birth to a raging modern-day discussion that can be viewed as the archetypal rationalism vs. empiricism dichotomy. Critics like Chomsky and others have been criticize, arguing that, they are 'stochastic parrots', and

that they are good at pattern matching, but not the essential cognitive properties on which human language and intelligence are founded (Chomsky, Watumull, and Roberts, 2023, p. 194). At the heart of their criticism is the fact that the LLMs have not managed to manifest discrete infinity and depend on structure in a principled manner. Chomsky claims that since the models are statistical, this does not mean that they can reliably differentiate a grammatical sentence and a statistically probable but grammatically unsound sentence, or that they can support the unrestrained recursion inherent in the Merge operation (Marcus and Davis, 2020, p. 336). For instance, an LLM may produce syntactically multifaceted sentences, but its capability to abstract a rule to a new and subjectively deep recursive construction may be unsure, representing a failure to code the fundamental Merge instruction. To explain LLMs, their groups, particularly Piantadosi and others, believe that such models really exhibit developing syntactic information, and that since the amount of the preparation data is large, they can indirectly produce the hierarchical constructions that UG assumes as essential (Piantadosi, 2023, Pp. 1-16). They uphold that long-range dependencies of the models can be tracked by self-attention, and they are a functional analogy of the structure-dependence principle. Besides, they claim that so-called poverty of stimulus argument is exaggerated because the corpora that is used to train LLMs is much richer and vast compared to the linguistic input a normal child has.

The discussion has therefore moved away to the question as to whether or not the LLMs can simply simulate human language to whether the underlying processes can be used as a viable, non-nativist explanation of linguistic phenomena. Such a contrast is therefore indispensable in defining whether the human language ability is a special and rule-based computational system or a comprehensive statistical engine modified to the linguistic contribution (Hupkes & De Raedt, 2024, Pp. 105-116).

3.Methodology

The current study uses a conceptual and theoretical comparative study, which is supported by the latest publications in the sphere of theoretical linguistics, cognitive science, and computational linguistics. Since the two systems that are subject to scrutiny namely UG, formal abstraction of the human mind and LLM, large-scale computational paradigm, cannot be compared directly and empirically in one experimental paradigm, the current study focuses on the formal properties and explanatory commitments that comprise the core generative mechanisms of each system.

The research summaries a strong and fourfold-criterion outline that obviously questions the differences and similarities between Cartesian Universal Grammar, explicitly the Minimalist Program with Merge, and transformer-based artificial intelligence schemes, and in particular the

Self-Attention mechanism. This structure provides a framework of the Data Analysis section to follow:

1. Computational Primitive: The earlier principle examines the simplest, irreducible procedure that gives rise to linguistic output in each model (Merge in UG or Self-Attention in LLM).
2. Representative Construction: The second principle likens the internal representative structures that organize linguistic knowledge and obviously hierarchy trees in UG and indirectly distributed vectors in LLM.
3. Knowledge: The third principle is the ontological sources of linguistic aptitude, associating the distinctive and domain-specific donations to the numerical optimization of large-scale domain-general studies.
4. Explanatory Scope: The fourth criterion evaluates the hypothesized explanatory area of every theory: idealized competence in UG and observable performance in LLM.

This analysis has a practical source on influential works in generative grammar (Chomsky, 2014, p. 420), influential works on the transformer architecture, and an edited sample of peer-reviewed research, book chapters, and pre-prints issued in 2020-2025 that obviously address the UGLLM debate. Through the application of the above analytic framework, the study aims to provide a rigorous, sophisticated comparison at the level going beyond surface based descriptions of output fluency.

4.Comparative Analysis

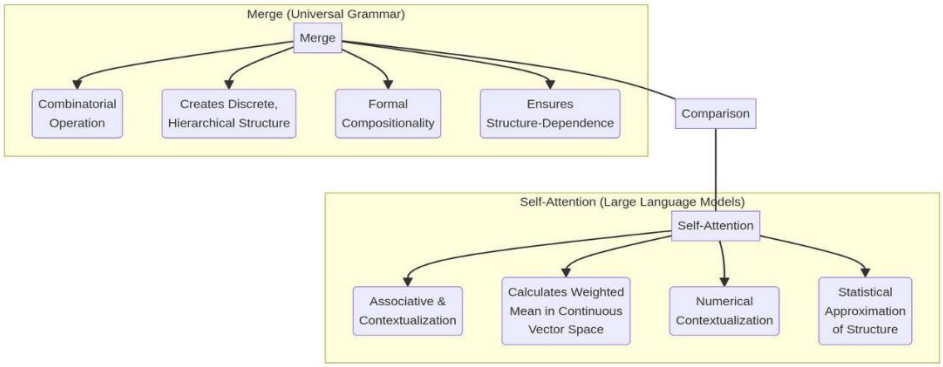
The comparative framework that was created in the methodology is now applied to the basic generative processes of UG and LLM. This discussion shows that there are deep, structural differences signifying the philosophical gap between the two paradigms.

4.1. Computational Primitives Analysis:

The most radical difference is on nature of the core computational primitive. Merge is a combinatorial operation; it is the operation of taking two different elements and gluing them into another new discrete labeled object. This process is recursive and structural in nature. The result of Merge is a hierarchical structure (A phrase or a sentence) that has a qualitative difference with its inputs. Merge is powerful because it is simple and produces infinite collection of expressions of a language using finite lexicon (Chomsky & et. al., 2023, p.220). The process of its generation is motivated by the necessity to meet the formal requirements of the interfaces (CI and SM). Self-Attention on the other hand is an associative and contextualization. It does not make a new and discrete hierarchical object. Rather, it calculates a weighted mean of all the input elements (the Value vectors) with weights based on the relevance (Attention scores) of each element to the current token

(Query) (Vaswani & et.al., 2017, p.5998-6234). It is not a construction of an edifice of a formal structure but the enrichment of a representation by statistical association. While hierarchical relations can indirectly be implicit by multiple layers of consideration, the multiplication is a fixed-depth computational task in a continuous vector interplanetary space. It is a generative process based on the probability of the next token based on the context that occurred. This difference is crucial: Merge is a structure of official compositionality, and Self-Attention is a structure of numerical contextualization. Merge ensures the formal property of structure-dependence, whereas Self-Attention only offers a very useful, albeit statistically determined, approximation to the same.

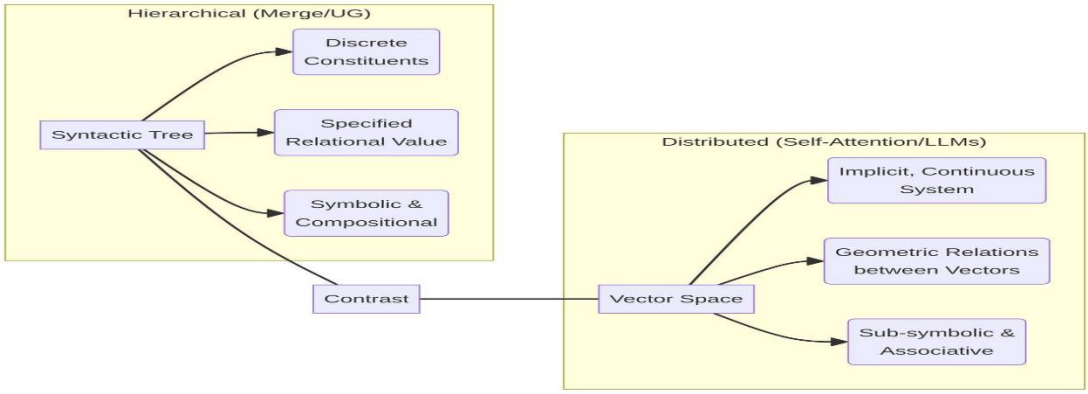
Figure 4.1: Merge and Self Attention



4.2. Representational Structure: Structure Hierarchical vs. Distributed

The computational primitive is directly manifested in the internal representation of language. With the help of the Merge operation, UG assumes a highly hierarchical and discrete representational structure, typically represented as a syntactic tree (Chomsky, 2014, p.460). Every constituent in this structure has a specified relational value- head, complement, or specified to all other constituents and these are inviolable syntactic principles. The manner of representation is also symbolic and compositional that is, the meaning of a complex expression is a provisional of the meanings of that expression, and also how the parts are syntactically assembled. In contrast, self-attention-based large-scale language models use an implicit distributed and continuous system of representation in high-dimensional vector space. Knowledge of language is coded into the parameters of the network and the embedded vectors. There are no clear lexical labels or structural relatives, but these possessions are implicitly signified in the geometric relations between vectors (Goldberg, 2019, p.10). Experiential studies have established that detailed attention heads in a Transformer can learn syntactic dependences that are meaningful of hierarchical construction (Clark & et. al., 2019, p.120), though are emergent features of statistical exercise as opposed to geometrically distinct structural components. The representation is therefore sub-symbolic and associative and the meaning of a lexical item is described by distributional profile of the lexical item.

Figure 4.2: Structure Hierarchical vs. Distributed

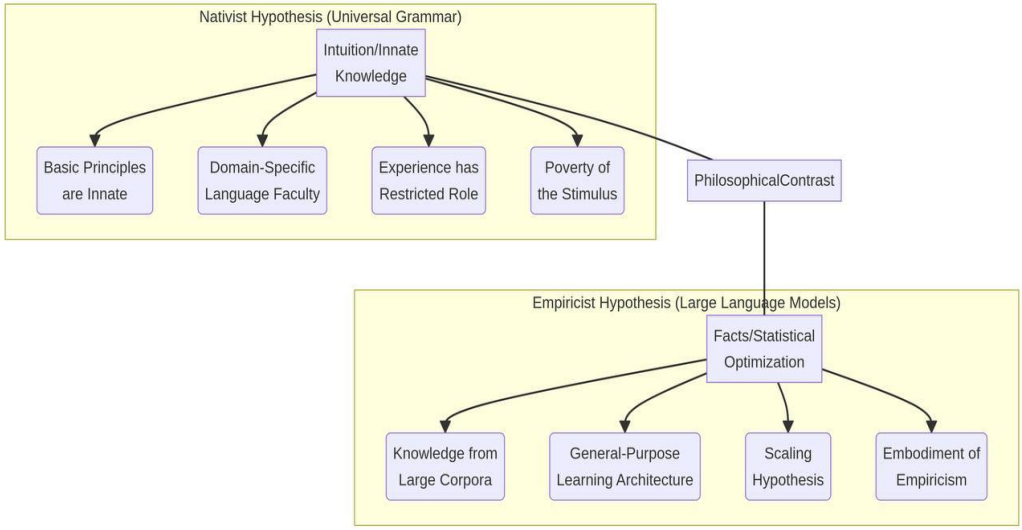


4.3 Comparison of Origin of Knowledge: Intuition vs. Facts

The philosophical point that is grounded is the provenance of the knowledge of language. The universal grammar is based on the nativist hypothesis, which holds that the basic principles of language the original state of the language faculty is innate and domain specific (Chomsky, 2002, p.209). In this conception, experience has a very restricted role of establishing the parameters of universal grammar and overcoming the so-called poverty of the stimulus. The capacity to use language has thus been viewed as an in-born aspect of biology and not a learned art. On the contrary, huge language models, are the embodiment of the empiricist hypothesis. They are all information-based and obtained in the context of statistical optimization on large, heterogeneous corpora (Kaplan & et. al., 2020, p.20).

Moreover, the general-purpose learning architecture, the Transformer as it is, and the learning algorithm, back propagation is the only innate components. The fact that large language models are successful is frequently used as a supporting fact to the scaling hypothesis, meaning that the complexity of language can be represented with general-purpose learning processes, and thus there is no need to have domain-specific, innate universal grammar (Piantadosi, 2023, p.14)

Figure 4.3 : Intuition vs. Facts



4.4 Exploratory Scope Analysis: Competence vs. Performance

The last criterion is the one that takes care of the field which both theoretical frameworks aim to shed light on. Cartesian Universal Grammar is thought of as a competence theory, an idealized, internalized theory, which lets a speaker produce and understand an infinite amount of sentences. It is obsessed with the formal, necessary qualities of language, consciously disengaged with cognitive limitations, memory capacities, and sociocultural background. It has its main objective: explanatory adequacy: to explain the very fact that language has the attributes it has. Besides, the Huge Language Models are models of performance, the obvious, physical practice of language, they are efficient at creating fluent, contextually-appropriate, and stylistically-different text, which is a characteristic of the effective performance. Their main goal is to predict accuracy: this is to predict the next token in a sequence with high likelihood. Even though they are extraordinarily effective in as far as the workings of the linguistic faculty are concerned, their inner workings do not, in a direct way, provide a clear explanation of the formal properties of the human language. The main controversy, then, is whether an extremely successful model of performance can be, by definition, a model of competence or whether the cognitive reality beneath requires an explanation that is more constricted, formal like UG.

5. Results

The comparative analysis gives a concise list of results as to the strengths and limitations of each paradigm as summarized in the table below (See Table 1):

The analysis has confirmed Cartesian UG and Transformer-Based AI Models symbolize two fundamentally incommensurable language approaches. UG proposes a rigorous, formal and parsimonious theory of the combinatorics of language, and manages to explain the property of

discrete infinity as a recursive operation, Merge. Its power lies on its explanatory profundity on the principles of human language that are formal. On the other hand, the unparalleled strength of statistical association and contextualization can be seen in LLMs. Their accomplished results in the production of very fluent and contextually competent text highlight the degree to which the application of language is directed by statistical regularities and distributional qualities. They are better than their competitors in predictive power and ability to model language performance at large scale.

Table 1: Comparative Findings

No.	Feature	Cartesian Universal Grammar (UG)	Transformer-Based AI Models (LLM)
1.	Computational Primitive	Merge (Combinatorial, Recursive)	Self-Attention (Associative, Weighted)
2.	Representational Structure	Explicitly Hierarchical (Syntactic Trees)	Implicitly Distributed (Vector Space)
3.	Source of Knowledge	Innate (UG, Domain-Specific)	Data-Driven (Massive Corpus, General-Purpose)
4.	Explanatory Scope	Competence (Formal Constraints, Explanatory Adequacy)	Performance (Fluency, Predictive Accuracy)
5.	Key Strength	Clarifying the nature of language (discrete infinity, structure-dependence)	Demonstrating the use of language (fluency, coherence, context)

6. Conclusion

Such a comparative analysis has made a systematic comparison between the language generation mechanisms of Cartesian Universal grammar and Transformer-Based AI Models, and the underlying differences between the combinational, recursive action of Merge and the associative, weighted action of Self-Attention. It has been stated that there is a deep theoretical gulf between the two paradigms: UG is a theory of formal structure and innate competence and LLMs are models of statistical association and data-driven performance. The success of LLMs does not amount to a refutation of UG, but is a dire threat to its extent. The structure of language in general and the existence of apparent hierarchical dependencies, in particular, is implicitly acquired by statistical optimization on large volumes of data, which has been demonstrated by LLMs.

6.1 Constructions for Future Research

Synthesis and not mutual exclusion is probably the most fruitful way to go on. Forthcoming studies ought to reflect the prospective of hybrid models which could apply the formal restraints of Merge in the architecture of a Modifier. As an example, syntactic tree construction models that explicitly use a Merge-like operation to inform the attention mechanism might be able to leverage the explanatory capability of UG, in addition to the predictive capability of the LLM. Also, empirical studies should proceed to test the LLMs based on the specific, non-statistical properties of language

which UG predicts, e.g. the capability to support unbounded recursion and structure-specific movement under novel, out-of-distribution conditions. Only when we challenge the boundaries of statistical learning with the expected predictions of the nativist theory with rigor can we have the hope of fully comprehending what exactly the nature of the human language faculty. It could be that ultimately the long-lasting enigma of language production is to be solved by a theory which acknowledges both the beautiful, natural simplicity of the Cartesian computational core and the mighty, statistical capability of the information-driven mind.

References

Alsaray, A. A. D., & Altimimiy, K. K. (2023). Employing artificial intelligence techniques among mobile journalism practitioners when covering daily events: A field study in Wasit Governorate. *Lark Journal of Philosophy, Linguistics and Social Sciences*, 15(3/Pt2), 537–577. https://doi.org/10.31185/lark.Vol2.Iss50.3165

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

Chomsky, N. (1966). *Cartesian linguistics: A chapter in the history of rationalist thought*. Harper & Row.

Chomsky, N. (2002). *On nature and language.* Cambridge University Press.

Chomsky, N. (2014). *The minimalist program*. MIT Press.

Chomsky, N., Seely, T. D., Berwick, R. C., & Fong, S. (2023). *Merge and the strong minimalist thesis*. Cambridge University Press.

Chomsky, N., Watumull, J., & Roberts, I. (2023, March 8). Chomsky: The fallacy of Chat GPT and large language models. *The New York Times*, A21.

Clark, K., Khandelwal, U., Levy, O., & Manning, C. D. (2019). What does BERT look at? An analysis of BERT's attention. Proceedings of the 2019 ACL Workshop Black boxNLP: *Analyzing and Interpreting Neural Networks for NLP*, 276-286.

Goldberg, Y. (2019). Assessing the ability of LSTMs to learn syntax-sensitive dependencies. *Language and Linguistics Compass*, 13(1), e12319.

Hupkes, D., & De Raedt, L. (2024). The role of recursion in large language models. *Trends in Cognitive Sciences*, 28(2), 105-116.

Kaplan, J., McCandlish, S., Henighan, T., Brown, T. B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., & Amodei, D. (2020*). Scaling laws for neural language models.* arXiv preprint arXiv:2001.08361.

Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing*. MIT Press.

Marcus, G., & Davis, E. (2020). *Rebooting AI: Building artificial intelligence we can trust*. Pantheon.

Piantadosi, S. T. (2023). Modern language models refute Chomsky's approach to language. *Cognitive Science*, 47(10), e13361.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30, 5998-6008.