International Conference on Innovation, Reflection, and Creativity in Arts, Literature, and Language Education (I-CIRCALLED)
Fakultas Bahasa dan Seni, Universitas Negeri Jakarta
14 Oktober 2025

A Reassessment of Chomsky's View on the Use of Corpus Databases in Linguistic Research: Between Theoretical Challenges and Empirical Opportunities

Ahmad Syafiq Amir Abdullah ZAWAWI[1*], Fazal Mohamed Mohamed SULTAN[2]
[1] Malay Linguistics Programme, Academy of Malay Studies, Universiti Malaya
[2] Center for Research in Language and Linguistics, Universiti Kebangsaan Malaysia

Email Adress
syaf.zawawi@um.edu.my
***Corresponding author**

Keywords: corpus database; linguistics; artificial intelligence; syntax; digital data

Abstract
In the generative linguistics tradition, Noam Chomsky has consistently rejected the use of empirical corpus data to study language structure, especially in syntax research. He believes that native speaker intuition is more important in language studies and argues that corpus data is not reliable because it can be affected by variation and does not show true linguistic competence. However, with the fast growth of artificial intelligence and language technologies, the availability of large corpus databases, and the increasing need for wider empirical analysis, this view has been debated again in today's linguistic research. This paper aims to re-examine Chomsky's arguments against corpus use by applying a corpus-based method in syntax studies. This can help us understand universal syntactic structures more clearly. Some challenges of using corpora include their limits in showing native speaker competence, the lack of negative data, their inability to reflect how the mind works, and the possibility of biased or limited data. However, there are also new opportunities in corpus-based research, such as having access to billions of words from many sources and types of texts, using advanced technology to find morphosyntactic patterns, and using big data to test hypotheses and theories. In conclusion, combining corpus-based research with theory is very important today. Corpus data is not an enemy of theory, it is a valuable tool that supports and strengthens modern linguistic analysis.

Introduction

The rapid development in the field of contemporary linguistics has witnessed a paradigm shift in research approaches—from purely introspective and theoretical methods to more data-driven and empirical ones (Baese et al. 2020; Taguchi & Ishihara 2018; Georgiou & Theodorou 2022; Andringa & Godfroid 2020). One approach that has increasingly gained attention is the corpus-based approach, which offers advantages in terms of access to authentic, diverse, and large-scale linguistic data (Arellano 2020; Bryfonski & Sanz 2018). Linguistic corpora allow researchers to analyze language phenomena based on actual evidence from language use in real-world contexts, thereby providing a more comprehensive view of linguistic variation and patterns (Chapman & Routledge, 2009; Carnie, 2013; Radford, 2009).

However, within the generative linguistic tradition, Chomsky (1965, 2002) consistently rejected the use of corpus data as the primary source in language analysis, particularly in syntactic studies. Chomsky argued that data derived from language performance is unstable and subject to various external interferences, such as speech errors, social variation, and contextual factors. Therefore, he emphasized that linguistic research should focus on linguistic competence that is, the speaker's mental knowledge of language structure which, according to him, can only be accessed through native speaker intuition and not through corpus data.

Chomsky's (1965, 2002) perspective has shaped the epistemological foundation of generative linguistics for several decades. However, with advancements in linguistic technology and artificial intelligence that allow large-scale processing and analysis of language data, the corpus-based approach now presents a competitive alternative in contributing to the development of linguistic theory. This raises new questions about the relevance of Chomsky's rejection of corpora in the context of modern linguistic research, which is increasingly corpus-driven.

Accordingly, this paper aims to reassess Chomsky's (1965, 2002) views on the use of corpus data in linguistic research, particularly in the field of syntax. This study revisits the key arguments put forth by Chomsky and compares the generative approach with the advantages offered by corpus methodology. In addition, this paper discusses the main challenges of using corpus data in linguistic studies, as well as the new opportunities arising from the development of large-scale big data corpora to support corpus-driven syntactic analysis. It is hoped that this discussion will contribute to a reevaluation of current linguistic paradigms and open the door to a more integrative approach between theory and empirical data in language analysis.

Methods

This study adopts a qualitative approach in the form of scholarly discourse analysis, focusing on two primary sources: theoretical texts in generative linguistics representing Chomsky's views (1965, 2002), and data and findings from corpus-based studies related to contemporary syntactic analysis. The aim of this approach is to critically evaluate Chomsky's main arguments against the use of corpus data and to compare the relevance of these arguments with the evidence and potential offered by modern corpus methodologies. Theoretical data is obtained through close reading and analysis of Chomsky's key works, including major texts that reflect the development of his thought. These perspectives are then contextually reviewed and linked to methodological issues in linguistic research, with a focus on topics such as the rejection of empirical data, the prioritization of native speaker intuition, and the conceptual framework of linguistic competence. A corpus contains large-scale linguistic data spanning various genres and domains, and has been employed in numerous modern syntactic studies. This research assesses how such data can contribute to the understanding of syntactic structures and the extent to which it can be used to support or challenge generative views. Through this approach, the study not only analyzes the tension between theory and data, but also explores the potential for integrating theoretical and empirical approaches to enrich linguistic research, particularly in the field of syntax.

Literature Review

One of the earliest and most influential criticisms by Chomsky (1965, 2002) regarding the use of corpus data is his emphasis on the distinction between linguistic competence and performance. In his work Aspects of the Theory of Syntax, Chomsky (1965) asserted that linguistic inquiry should focus on linguistic competence—namely, the speaker's mental knowledge of the language system—rather than actual language use, which is often influenced by cognitive, social, and contextual disturbances. He argued that performance data, such as that found in corpora, is inherently unstable and does not accurately reflect the internal structure of language. However, this approach overlooks the possibility that variation in performance may also offer valuable insights into the real language system and how it functions within a speech community.

In addition, Chomsky (1965, 2002) placed great emphasis on native speaker intuition as the primary source for evaluating the grammaticality of syntactic structures. Within the generative tradition, intuition-based tests—such as the acceptance or rejection of sentences by native speakers—are considered more valid than observations derived from corpus data. He viewed intuition as a direct reflection of a speaker's mental competence, whereas corpus data was seen as merely indicative of usage habits, not structural validity. However, this approach has been questioned by contemporary researchers, who argue that speaker intuition is often influenced by context, educational background, and social factors. As such, exclusive reliance on intuition also

carries limitations in producing findings that accurately represent the broader language community..

Another key argument put forward by Chomsky (1965, 2002) concerns the absence of negative evidence in corpora. He argued that corpus data only contains uttered or written sentences, but does not provide information about structures that are ungrammatical or disallowed within the language system. In syntactic theory-building, negative evidence is essential for delineating the boundaries between acceptable and unacceptable structures. However, corpus-based approaches have begun to develop alternative strategies, such as low-frequency analysis and the identification of anomalies, as indirect indicators of potentially ungrammatical constructions even though these do not entirely replace explicit negative evidence.

In addition, Chomsky (1965, 2002) also criticized the corpus approach for relying on actual usage data, which is heavily influenced by genre, situational, and social variation. He argued that data closely tied to real-world contexts is difficult to use in formulating universal syntactic principles that are free from external interference. However, in the context of current research, this view is increasingly challenged, as various modern corpora including structured corpora such as the DBP Corpus now provide more balanced and diverse datasets. In fact, variation in the data is seen as a strength rather than a limitation, as it helps researchers understand how linguistic structures are realized across different contexts and user communities.

Lastly, Chomsky (1965, 2002) consistently emphasized that linguistics is a branch of cognitive science, and therefore the primary focus should be on constructing idealized mental models of language. According to him, reliance on external data such as corpora cannot adequately represent the internal, abstract, and mentalistic nature of linguistic knowledge. Nevertheless, advances in linguistic technology and big data analysis now make it possible to generate syntactic hypotheses based on extensive and detailed empirical data. This development suggests that corpora can not only contribute to understanding surface-level patterns, but also play a role in supporting and testing theoretical models including generative syntax itself when used critically and integratively.

Although Chomsky (1965, 2002) upheld the dominance of intuition and a deductive approach in linguistics, corpus linguistics researchers such as Bramo (2018), Ameka (2018), and Vulchanova et al. (2022) have completely rejected the principle that only native speaker intuition is a valid basis for linguistic data. According to Bramo (2018), Ameka (2018), and Vulchanova et al. (2022), they argue that intuition is insufficient and often influenced by individual biases or pre-existing theoretical beliefs. They criticize the introspective approach as being prone to speculation without empirical verification and insist that linguistics should be grounded in actual language use evidence, much like other scientific fields that prioritize systematic observation. This view directly challenges the epistemological foundation of generative linguistics established by Chomsky and opens the way for more data-driven research.

Cui & Zhou (2020), Maran et al. (2022), and Schneider (2020) also criticize the traditional approach for overlooking the diversity of language use. They highlight how certain syntactic structures may dominate in specific genres or communication contexts and may not necessarily appear in the intuition of speakers with limited language experience. Cui & Zhou (2020), Maran et al. (2022), and Schneider (2020) emphasize that without corpora, researchers lose the ability to understand the frequency, collocations, and natural patterns of structures in speech and writing. Schneider (2020) views this diversity not as noise but as key to understanding the flexibility and productivity of the language system a perspective that contrasts with Chomsky's desire to filter out 'noise' in performance data.

All these scholars, with their respective approaches, challenge the monopoly of generative approaches and pave the way for a new linguistic paradigm that is more empirical and open to diverse data. Although Chomsky's views remain influential in terms of theoretical value, the corpus researchers discussed have proven that corpus approaches are not only methodologically valid but also necessary in the increasingly complex and big data-oriented contemporary linguistic era. Therefore, this study contends that an integrative approach that acknowledges both the role of theory and the strength of empirical data should be prioritized to address current challenges in the field of syntax. Meanwhile, Lau & Tanaka (2021) emphasize the importance of quantitative and statistical methods in empirical linguistics. In their various works, they introduce approaches based

on collostructional analysis and corpus-based experimental design, enabling the statistical validation of linguistic hypotheses. Lau & Tanaka (2021) reject the view that syntactic structure studies cannot be conducted through corpus data and instead demonstrate how corpus techniques can reveal relationships between structure and meaning in ways that intuition alone cannot achieve. They argue that modern linguistics needs to integrate theoretical approaches with empirical evidence in a complementary, rather than opposing, manner.

Results and Discussion

The Results and Discussion section of this paper is divided into two main subsections. The first subsection will elaborate on the theoretical challenges faced in using corpus data, particularly in syntactic studies, including issues related to the limitations of corpus data in representing linguistic competence and the difficulties of integrating empirical data with Minimalist syntactic theory. Meanwhile, the second subsection will focus on the empirical opportunities offered by the use of corpus data, especially in the era of big data technology and advancements in large-scale corpus databases, which enable more in-depth and evidence-based syntactic analysis.

Challenges of Corpus Data in Linguistic Research

This paper outlines six challenges in using corpus databases in linguistic research, each summarized in Table 1.

Table 1: New Challenges in Using Corpus Data in Linguistic Research

Challenges	Explaination
Data Quality and Cleanliness Issues	Corpus data often contains errors, non-standard language, or language interference that can affect the accuracy of syntactic analysis. The data cleaning process is time-consuming and risks removing important native language variations. However, with advances in Natural Language Processing (NLP) techniques and data cleaning algorithms, many of these disturbances can be automatically reduced without losing meaning. Moreover, such variations actually reflect richer real language use and provide an important empirical dimension.
Limitations in Representing Complex Syntactic Structures	Corpora typically store data in a linear form without sufficiently detailed annotations, making it difficult to analyze complex hierarchical or syntactic dependency relationships. However, many corpora today come with syntactic annotations (parsed corpora) that enable more detailed studies of tree structures and dependencies. Tools like treebanks also facilitate the integration of corpus data with syntactic theory models.
Challenges of Manual Scale Analysis in Syntax	Manual or semi-automatic syntactic annotation requires significant time and resources, limiting the use of very large-scale corpora. However, machine learning and natural language processing technologies now enable increasingly accurate and efficient automatic annotation, reducing reliance on manual annotation and opening opportunities for analyzing larger and more complex data.
Difficulties in Controlling Socio-Contextual Factors	Language variation in corpora is influenced by various social factors that are difficult to control or isolate, complicating the discovery of universal syntactic patterns.

	Nonetheless, the presence of metadata and social annotations in modern corpora allows researchers to systematically control or study the influence of socio-contextual factors, providing a more comprehensive understanding
Limitations in Capturing Prosodic and Intonational Aspects	Written data or text transcriptions in corpora fail to capture important prosodic and intonational aspects in spoken syntax. However, the development of annotated speech corpora containing audio and prosodic data (such as CORPUS or CHILDES) enables more comprehensive studies that integrate both syntactic and prosodic aspects..
Challenges in Adapting Syntactic Models to Various Languages and Dialects	Corpora typically focus on standard languages, posing challenges for studying minority languages or dialects with limited data and annotations. However, initiatives to develop corpora for minority languages and dialects are increasing worldwide, leveraging digital technology and local community collaboration, thereby opening opportunities for more inclusive and diverse syntactic research.

New Opportunities in Corpus-Based Research

Discussion on new opportunities for exploration in corpus-based research is presented in a table organized by themes, as shown in Table 2.

Table 2: New Opportunities in Corpus-Based Research

Opportunities	Explaination
Corpus-based linguistic research	Students and researchers can now access large-scale corpus databases covering various domains, genres, and language variations. This enables them to conduct syntactic and morphosyntactic analyses in a more empirical, detailed, and large-scale manner. For example, databases like the DBP Corpus allow systematic searches of sentence structures, phrase patterns, and usage variations, supporting hypothesis development and data-driven testing of syntactic theories. This makes research more credible and evidence-based, aligning with current scientific standards.
Development of linguistic applications	Access to corpora allows linguistic researchers to understand the actual usage of language forms in real contexts, rather than relying solely on prescriptive norms. Moreover, with user-friendly interfaces and data visualization tools, even non-experts can easily explore language patterns and their variations. In the future, the integration of corpus data in language education, digital dictionary development, and AI-assisted learning tools will continue to grow, making language not only an academic subject but also a living resource that can be utilized by diverse segments of society..
Collaborative approaches in linguistic research	The use of corpora also opens up opportunities for collaborative approaches in linguistic research. With the existence of open corpora and publicly accessible data sources, researchers from various institutions and fields can

	work together to analyze the same data simultaneously, encouraging the exchange of ideas, diverse perspectives, and study replication. This contributes to methodological transparency and the reliability of research findings, which are crucial aspects in the development of modern linguistics. Such an approach also has the potential to support the development of new theories that are more flexible and based on data from multiple languages and their usage variations.
Construction of adaptive models	Modern technology's ability to process large-scale data enables the construction of adaptive, data-driven syntactic models. This allows for the generation of linguistic generalizations that are not necessarily based on existing linguistic theories but are built from empirical observations of real language patterns. This development has the potential to challenge traditional deductive theoretical models and opens the way for a hybrid approach combining theory and data.
Development of authentic teaching materials	In the context of language teaching and learning, corpus data can be utilized to develop more authentic teaching materials based on real contemporary language use. Language teachers can use corpora to find genuine sentence examples, identify common errors among students, and design learning activities based on patterns derived from the data. This enriches teaching methods and enhances the effectiveness of data-driven learning.
Empowerment of local dialects	The use of corpora also has the potential to empower local languages and dialects through the documentation and preservation of linguistic data that may be at risk. Digital corpora can serve as language archives that not only store linguistic information but also act as research and educational tools for future generations. With the support of communities and language institutions, these initiatives can contribute to the sustainability of the nation's linguistic heritage.

Conclusions

Overall, the development of corpus data usage in linguistic research should not be seen as a threat to the tradition of linguistic theory, especially within the generative approach, but rather as a complement that strengthens efforts to comprehensively understand language structure. Although there are challenges in terms of methodology and theory integration, the great potential offered by corpora in terms of scale, data diversity, and empirical capability—has opened new dimensions in syntactic and general linguistic research. In the context of contemporary linguistics, which demands a more open, inclusive, and evidence-based approach, corpora have now emerged as a major force supporting the development of a more robust, responsive, and relevant linguistic science for the needs of the times.

References

Ameka, F. K. (2018). From comparative descriptive linguistic fieldwork to documentary linguistic fieldwork in Ghana. *Language Documentation & Conservation*: 224-239.

Andringa, S. & Godfroid, A. (2020). Sampling bias and the problem of generalizability in applied linguistics. *Annual Review of Applied Linguistics* 40: 134-142.

Arellano, R. (2020). Challenges and opportunities of teaching applied linguistics in the context of EFL teacher training. *Memory* 8: 14-54.

Baese-Berk, M. M., McLaughlin, D. J. & McGowan, K. B. (2020). Perception of non-native speech. *Language and Linguistics Compass* 14(7): 135-154.

Bramo, E. (2018). Syntactical Analytical Overview of the New Testament Translation by Vangjel Meksi, After the Editing of Grigor of Gjirokastra, Focusing on the Syntax and the Sentence Types Strata. *European Journal of Multidisciplinary Studies 3*(2): 98-107

Bryfonski, L. & Sanz, C. 2018. Opportunities for corrective feedback during study abroad: A mixed methods approach. *Annual Review of Applied Linguistics* 38: 1-32.

Carnie, A. (2013). *Syantax: A Generative Introduction (Third Edition).* United Kingdom: Willey-Blackwell Publication.

Chapman, S. & Routledge, C. (2009). *Key Ideas In Linguistics and The Philosophy of Language*. Edinburgh: Edinburgh University Press.

Chomsky, N. (1965). *Aspects of the Theory of Syntax* (50th ed.). The MIT Press.

Chomsky, N. (2002). *On Nature and Language*. United Kingdom: Cambridge University Press.

Cui, S. & Zhou, C. (2020). Writing Features Influencing Non-Native English Speakers' Publication in International Journals. *Patchwork* 4: 49-62.

Georgiou, G. P. & Theodorou, E. (2022). Comprehension of complex syntax by non-English-speaking children with developmental language disorder: A scoping review. *Clinical Linguistics & Phonetics*: 1-19.

Lau, E. & Tanaka, N. (2021). The subject advantage in relative clauses: A review. *Glossa: a journal of general linguistics* 6(1): 37-45.

Maran, M., Friederici, A. D. & Zaccarella, E. (2022). Syntax through the looking glass: A review on two-word linguistic processing across behavioral, neuroimaging and neurostimulation studies. *Neuroscience & Biobehavioral Reviews* 10(2): 104-137.

Radford, A. (2009). *Analysing English Sentences: A Minimalist Approach*. United States of America: Cambridge University Press.

Schneider, E. W. (2020). *Developmental patterns of English: Similar or different?. The Routledge handbook of world Englishes.* United Kingdom: Routledge.

Taguchi, N. & Ishihara, N. (2018). The pragmatics of English as a lingua franca: Research and pedagogy in the era of globalization. *Annual Review of Applied Linguistics* 38: 80-101

Vulchanova, M., Vulchanov, V., Sorace, A., Suarez-Gomez, C. & Guijarro-Fuentes, P. (2022). The notion of the native speaker put to the test: recent research advances. *Frontiers in Psychology:* 1432.