English Language Teaching; Vol. 17, No. 9; 2024
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Can ChatGPT Understand Malapropism Correctly? Challenges to Davidson's Passing Theory in Generative AI

Tsukasa Yamanaka[1]

[1] College of Life Sciences, Ritsumekan University, Kusatsu, Japan

Correspondence: Tsukasa Yamanaka, College of Life Sciences, Ritsumeikan University, Kusatsu, Shiga, 525-8577, Japan. E-mail: yaman@fc.ritsumei.ac.jp

Received: August 2, 2024 Accepted: August 20, 2024 Online Published: August 22, 2024

doi: 10.5539/elt.v17n9p48 URL: https://doi.org/10.5539/elt.v17n9p48

Abstract

This study examines ChatGPT's performance in understanding Japanese malapropisms, aiming to explore its capacity for linguistic inference compared to humans. Despite its remarkable fluency in conversation, ChatGPT shows significant limitations in comprehending malapropisms, particularly in handling phonetic, lexical, and contextual errors. Using a specialized dataset, the research highlights these gaps, suggesting that while ChatGPT excels in fluency, its understanding of nuanced language phenomena remains distinct from human comprehension. The findings contribute to the discourse on the potential and limitations of generative AI, advocating for a reevaluation of linguistic-philosophical theories in light of AI advancements.

Keywords: ChatGPT, malapropism, Davidson, passing theory, probability

1. Introduction

Language is a complex and dynamic system, as evidenced by phenomena like malapropisms, where words are mistakenly substituted for similar-sounding ones. Understanding such nuances is a testament to the sophistication of human language comprehension. This study examines the capabilities of OpenAI's ChatGPT, a prominent generative AI model, in understanding and interpreting malapropisms, and how this challenges existing linguistic-philosophical theories, particularly Donald Davidson's passing theory.

1.1 ChatGPT

ChatGPT is a generative AI model developed by OpenAI, designed to understand and generate human-like text using a deep neural network architecture known as the Transformer. Trained on vast amounts of textual data, ChatGPT can engage in a variety of conversational tasks, from casual dialogue to more intricate language-based interactions. Its relevance to this research stems from its ability to simulate human-like language understanding, making it an ideal subject for investigating how well AI can grasp nuanced language phenomena such as malapropisms. The rise of generative AI models like ChatGPT has sparked debate over whether machines can genuinely understand the complexities of human language. As noted by Wu et al. (2023), ChatGPT demonstrates impressive fluency in handling complex language tasks, but this raises important questions about its ability to comprehend deeper linguistic phenomena.

1.2 Passing Theory

Donald Davidson's passing theory offers a philosophical framework for understanding language communication, emphasizing that successful language communication relies not just on shared conventions but on the continuous reinterpretation of meaning. This process, Davidson argues, is uniquely human, as it requires the ability to adapt and construct new interpretations in response to each communication (Davidson, 1986). Passing theory is particularly relevant when examining malapropisms, where words or phrases are mistakenly used in place of similar-sounding ones. These instances challenge the conventional understanding of language and require a deeper level of interpretation to grasp the intended meaning. The ability (or inability) of AI, such as ChatGPT, to navigate these linguistic challenges serves as a critical test of its communicative understanding and raises questions about the extent to which AI can replicate human-like language communication.

In recent years, the debate surrounding AI's capacity to understand language has intensified, particularly with the development of models like ChatGPT. Critics such as Titus (2024) argue that the apparent comprehension exhibited by generative AI is merely a result of probabilistic modeling, lacking the true semantic understanding, as exemplified by Davidson's passing theory. This study aims to explore these concerns by comparing ChatGPT's performance in interpreting Japanese malapropisms with human understanding, thereby assessing the strengths and limitations of generative AI in replicating human language capabilities. To the best of the author's knowledge, no research has yet been undertaken from this specific standpoint.

This research contributes to the broader discourse on AI's role in language processing by challenging the assumptions underlying current language theories. As Piattelli-Palmarini (1980) discusses in the debate between Piaget and Chomsky, language is often viewed as a uniquely human trait, deeply rooted in cognitive processes that are difficult to replicate artificially. Our findings suggest that while ChatGPT excels in fluency, its understanding of malapropisms—a task that requires deep semantic processing—remains limited. This supports the view that true linguistic comprehension involves more than probabilistic modeling, as Davidson and others have argued.

This study critically examines the performance of ChatGPT in understanding malapropisms, drawing connections to existing linguistic theories and raising important questions about the future role of AI in language understanding. The findings challenge the notion that AI can fully replicate human linguistic abilities, suggesting a need for further research and theoretical development in this rapidly evolving field.

2. Methodology

2.1 Research Design

This study employs a mixed-methods approach to evaluate ChatGPT's ability to understand and infer the correct meanings of Japanese malapropisms. The research design is structured to address the central research question: How effectively can ChatGPT comprehend malapropisms compared to human understanding?

2.2 Data Collection

The primary data source for this study is a corpus of 1,500 Japanese malapropisms, derived from Tamaoka & Terao (2004). This corpus includes examples of spontaneous speech malapropisms, categorized by phonology, morphology, lexicon, and context. Each example is annotated with the correct expression and the type of malapropism it represents.

2.3 Procedure

2.3.1 ChatGPT Interaction

ChatGPT (version 3.5) was presented with each malapropism example from the corpus. The AI was given the malapropism and its context and asked to infer the correct expression. For instance, when given the phrase "magirawa magirawashikute" where "magirawa" is a malapropism, ChatGPT was prompted (Figure 1)

"Magirawa magirawashikute" to iu hatsuwa nanodesuga "magirawa" toiunoga iimachigai nanodesuga hontouha nante iitakattanoka wakarimasuka?

(Regarding the utterance "magirawa magirawashikute," the "magirawa" part is a malapropism, but do you know what it really meant?)

ChatGPT responded with:

"Magirawa magirawashikute" to iu hatsuwaha osoraku "magirawashii magirawashikute" no kotowo iitakatta nodehanaikato omowaremasu.

(The utterance "magirawa magirawashikute" was probably intended to refer to "magirawashii magirawashikute".)

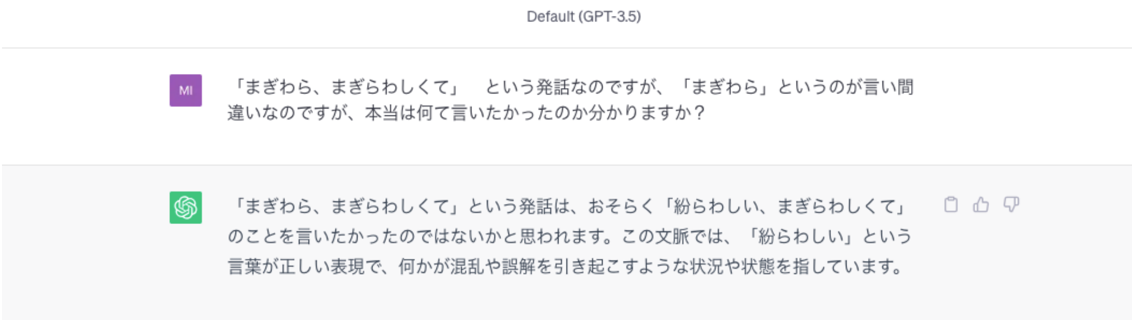

Figure 1. Examples of prompts and interactions with ChatGPT

2.3.2 Human Evaluation

Four human evaluators, all proficient in Japanese and with some prior experience using ChatGPT, were tasked with scoring ChatGPT's responses. They were provided with the malapropism, the correct expression, and ChatGPT's inferred response. Evaluators rated the accuracy of ChatGPT's understanding on a scale from 0 to 100, where 0 indicates no understanding and 100 indicates perfect understanding.

2.4 Data Analysis

The analysis focused on comparing the average scores given by human evaluators across different categories of malapropisms (e.g., phonological vs. lexical errors). Statistical methods, including frequency distribution and mean score analysis, were employed to identify patterns in ChatGPT's performance. For example, in cases where phonological errors were prominent, the average score was analyzed to determine if ChatGPT consistently struggled more with these types of errors compared to lexical errors.

2.5 Reliability and Validity

To ensure the reliability of the findings, the evaluation process was conducted in multiple sessions, with evaluators scoring responses independently. The validity of the research design was supported by using a well-established corpus and involving multiple human evaluators to mitigate individual biases.

2.6 Research Questions

This research design is specifically tailored to address the central research question by systematically evaluating ChatGPT's performance across a wide range of malapropisms. By incorporating detailed examples and rigorous analysis, the methodology ensures that the study directly assesses the AI's ability to understand and process linguistic anomalies, providing a robust comparison with human language comprehension.

3. Results

The primary focus of this study was to evaluate how effectively ChatGPT can understand and infer the correct meanings of Japanese malapropisms. The results are presented below, categorized by the types of errors and their associated performance metrics.

3.1 Overall Performance

Table 1 shows the distribution of average scores assigned by four evaluators across 1,500 cases.

Table 1. Distribution of Average Scores for 1,500 Cases Scored by Four Evaluators

Score Range	Number of Cases	Percentage
0-20	658	44%
21-40	201	13%
41-60	126	8%
61-80	135	9%
81-100	377	25%

The average score across the 1,500 cases was 42.1, indicating that ChatGPT is not consistently accurate in identifying malapropisms. Only 25% of the cases were rated between 81-100, while 44% were rated between 0-20, suggesting that, despite ChatGPT's fluency in conversation, it often fails to correctly assess malapropisms.

The limitations of the database and the small number of evaluators (four) should be noted, and definitive conclusions should be approached with caution. Nevertheless, it appears that ChatGPT currently lacks the ability to reliably identify malapropisms.

3.2 Error Level Analysis

Given that each case in the database contains between one and five tags, and there are 32 different tags in total, we performed an analysis focusing on the frequency of these tags. Table 2 summarizes the occurrence of each tag based on the "level" of the malapropism.

Table 2. Frequency of Occurrence of Each Tag Information According to the Level of Malapropism

Category	Number of Cases	Percentage
Phonology	961	64%
Morphology	47	3%
Lexicon	121	8%
Phrase	26	2%
Sentence	0	0%
Root	12	1%
Conjugation	16	1%
Grammar (Function Words)	0	0%
Grammar (Root)	0	0%
Grammar (Particle)	0	0%

Phonological errors are the most prominent, accounting for 64% of all cases, followed by lexical errors (8%) and morphological errors (3%). This distribution suggests that the database is biased towards phonological malapropisms, which are often related to sound.

3.3 Error Type Analysis

We further analyzed the data based on the "type" of malapropism. Table 3 presents the frequency of each error type.

Table 3. Frequency of Occurrence of Each Tag Information According to the Type of Malapropism

Category	Number of Cases	Percentage
Substitution	1161	77%
Replacement	52	3%
Addition	39	7%
Omission	127	8%
Movement	7	0%
Blend (Mixed)	0	0%

Substitution is the most common type of malapropism, representing 77% of all cases, followed by omission (8%) and replacement (3%). This indicates that substitution errors are by far the most frequent in everyday speech.

3.4 Phonological Error Analysis

Phonological errors were further categorized to identify specific types of errors. The results are summarized in Table 4.

Table 4. Frequency of Occurrence of Each Tag Information According to the Malapropism Error in Phonology

Category	Number of Cases	Percentage
Vowel	162	11%
Consonant	200	13%
Special (Exceptional)	49	3%
Mora	396	26%
Syllable	0	0%
Distinctive Feature	0	0%
Mora / Vowel	34	2%
Mora / Consonant	73	5%

The majority of phonological errors are related to mora (26%), followed by consonant errors (13%) and vowel errors (11%).

3.5 Lexical Error Analysis

Lexical errors were categorized based on the type of error. The results are presented in Table 5.

Table 5. Frequency of Occurrence of Each Tag Information According to the Malapropism Error in Lexicon

Category	Number of Cases	Percentage
Semantics	37	2%
Pragmatics	7	0%
Syntax	4	0%
Phonetics / Phonology	8	1%

Lexical errors account for a smaller portion of the total dataset, with semantic errors being the most frequent (2%).

3.6 Contextual Error Analysis

Contextual errors were also tagged, and the frequency of occurrence is summarized in Table 6.

Table 6. Frequency of Occurrence of Each Tag Information According to the Context-related Malapropism

Category	Number of Cases	Percentage
Context	797	55%
Semantic Context	19	1%
Perceptual Context	0	0%

Just over half of the malapropisms in this study are context-related. These errors often co-occur with phonology and substitution errors, indicating the complexity of understanding malapropisms without context.

3.7 Analysis of High-Frequency Tags

To better understand how ChatGPT handles malapropisms, we analyzed the top five high-frequency tags. Table 7 lists the top eight tagging frequencies.

Table 7. Top 8 Most Frequently Appearing Tagging

Ranking	Tagging Type	Number of Cases
1	Substitution	1164
2	Phonology	961
3	Context	797
4	Mora (Under Phonological Error)	396
5	Consonants (Under Phonological Error)	200
6	Vowels (Under Phonological Error)	162
7	Omission	127
8	Lexicon	121

Among these, mora, vowel, and consonant errors are types within the phonological error category. For further analysis, we focused on substitution, phonology, context, omission, and lexicon.

3.8 Analysis of Combined Tags

To explore ChatGPT's strengths and weaknesses, I analyzed cases where malapropisms involved combinations of the five main tags. Table 8 summarizes the results.

Table 8. Summary of Assessments for Each of the Five Main Elements of Malapropism

	Phonology	Substitution	Context	Lexicon	Omission
Number of Cases	961	1162	795	121	127
Average Total	38045.0	47986.3	33369.2	4133.3	6340.0
Average Mark	39.6	41.3	42.0	34.2	49.9
Number of Subjects Who Gave 0	1514.0	1732.0	1132.0	196.0	147.0
Number of 0 Per Case	1.6	1.5	1.4	1.6	1.2
Number of Subjects Who Gave 100	927.0	1184.0	817.0	82.0	166.0
Number of 100 Per Case	1.0	1.0	1.0	0.7	1.3

It is important to note that the data may be biased due to the unequal distribution of cases across the elements. Nevertheless, the average scores suggest that ChatGPT has a lower ability to infer malapropisms related to phonology, with an average score of 39.6, which is below the overall average of 42.1 as discussed in Table 1. This indicates that ChatGPT is relatively "weak" at reasoning about phonological errors.

Similarly, the average score for substitution was 41.3, close to the overall average, while the score for context was 42.0, also near the average. However, the score for lexicon dropped significantly to 34.2, suggesting that ChatGPT struggles even more with understanding malapropisms related to context than those related to phonology. This is further supported by the higher number of 0 scores per case (1.6) and the lower number of 100 scores per case (0.7), indicating that many evaluators rated ChatGPT as not understanding the malapropism at all.

On the other hand, for omission, the results were the opposite of those for lexicon. With an average score of 49.9, well above the overall average, and the lowest number of 0 scores per case (1.2), omission appears to be an area where ChatGPT performs relatively well. The higher number of 100 scores per case (1.3) further supports this observation, suggesting that omission is an area of strength for ChatGPT.

To further explore ChatGPT's "weak" areas related to lexicon and its "strong" areas related to omission, we examined cases where errors included these two elements. Table 9 summarizes the results of the evaluation of malapropisms containing two of the five elements discussed in Tables 7 and 8.

Table 9. Summary of Assessments for Malapropisms Involving Combinations of Two of the Five Main Elements

	Phonology + Substitution	Phonology + Context	Phonology + Lexicon	Phonology + Omission	Substitution + Context	Substitution + Lexicon	Substitution + Omission	Context + Lexicon	Context + Omission	Lexicon + Omission
Number of Cases	823	672	3	50	744	95	7	25	1	13
Average Total	32670	27527.1	187.5	2112.5	30730.4	3352.1	496.3	760	0	365
Average Mark	39.7	41.0	62.5	42.3	41.3	35.3	70.9	30.4	0	28.1
Number of Subjects Who Gave 0	1285	1003	2	74	1081	157	3	41	4	22
Number of 0 Per Case	1.6	1.5	0.7	1.5	1.5	1.7	0.4	1.6	4	1.7
Number of Subjects Who Gave 100	793	677	4	60	754	72	13	18	0	3
Number of 100 Per Case	1	1	1.3	1.2	1	0.8	1.9	0.7	0	0.2

By comparing the data in Table 9 with that in Table 8, we can infer that some of the high scores for omission observed in Table 8 decrease when combined with other factors. For example, the combination of phonology and omission yields an average score of 42.3, which is close to the overall average of 42.1 and not significantly higher than the 49.9 seen for omission alone in Table 8. This suggests that the small number of omission cases may be contributing to the higher average score, rather than the combination with other tags of higher frequency.

In contrast, lexicon shows a tendency to lower the average score when combined with other elements, except in cases where the number of instances is significantly low.

Based on the above analysis, we decided to further investigate cases involving tags not selected in the top five, as well as combinations of omissions and lexicon, to better understand their impact on inference accuracy. Table 10 summarizes the results.

Table 10. Assessment of Cases Containing at Least One Element Other Than the Five Major Elements, Grouped by Omission and Lexicon

	Omission + Function	Omission + Morphology	Omission + Particle	Omission + Phrase	Lexicon + Semantic Context	Lexicon + Meaning
Number of Cases	29	15	10	10	17	22
Average Total	1672.5	1271.3	443.8	413.8	542.5	723.3
Average Mark	57.7	84.8	44.4	41.4	31.9	32.9
Number of Subjects Who Gave 0	23	2	14	15	27	37
Number of 0 Per Case	0.8	0.1	1.4	1.5	1.6	1.7
Number of Subjects Who Gave 100	43	37	12	9	9	13
Number of 100 Per Case	1.5	2.5	1.2	0.9	0.5	0.6

The analysis confirms our predictions, showing significantly higher average scores for combinations of omission with function and morphology. This suggests that ChatGPT is generally good at inferring malapropisms related to omissions, especially when these are combined with functional or morphological elements. Conversely, lexicon consistently shows lower ratings, even when combined with various tags, indicating that ChatGPT struggles with making inferences related to lexicon in general.

The data presented in this section reveal key insights into the strengths and limitations of ChatGPT in understanding malapropisms. While the AI shows some ability to handle contextual errors, its performance in phonological and lexical areas is notably weaker. These findings suggest that while ChatGPT's fluency in conversation may be impressive, it lacks the deeper semantic understanding necessary to accurately interpret malapropisms.

4. Discussion

The results of this study suggest that ChatGPT has not yet reached a level where it can fully understand human malapropisms. While ChatGPT exhibits a fluency that may give the impression of effective communication, it struggles with irregular linguistic phenomena such as malapropisms, indicating that it cannot yet be said to possess Davidson's passing theory. The implications of these findings are discussed in the following sections.

4.1 Critical Consideration

ChatGPT operates on a different level of language understanding compared to humans. Although it may seem to process language on par with human abilities, this is likely not the case. The malapropisms analyzed in this study likely deviated too far from normal language patterns for ChatGPT to handle effectively due to limited relevant data, especially in Japanese. This reveals the "false" linguistic behaviors of ChatGPT, particularly when facing non-standard language use. Moreover, philosophical critiques, such as Gubelmann's (2023) discussion on the symbol grounding problem, suggest that without a physical embodiment, ChatGPT cannot fully grasp meaning or experience in the way humans do. However, as McGregor (2023) notes, the significance of corporeality in language understanding remains a topic of debate, with some arguing that behavior, rather than physicality, should be the focus.

4.2 Optimistic Consideration

There is also the possibility that improvements in the environment, such as the methodology used in this study or advancements in technology, could enable future versions of ChatGPT to better understand malapropisms. For example, while this study used ChatGPT 3.5, iterations like version 4o could potentially utilize larger datasets and more parameters to enhance probabilistic calculations. The incorporation of multimodal information (e.g., audio and video) alongside textual data could provide a richer understanding of context, thereby improving output accuracy. Additionally, increasing the number of samples and subjects in future studies and implementing more rigorous scoring criteria could yield more statistically significant results. Furthermore, it is worth questioning whether the assumption that humans should be able to fully understand all malapropisms is valid, given that the current study relied solely on textual data stripped of contextual information. With these methodological enhancements and technological advancements, it is conceivable that ChatGPT could eventually reach a level where it understands malapropisms to some extent.

4.3 Linguistic-Philosophical Exploration

Up to this point, I have discussed ChatGPT's ability to understand malapropisms from both critical and optimistic perspectives. This issue will likely be revisited by various researchers, even if not by me, as the field evolves.

One goal of AI research, including the development of systems like ChatGPT, is to gain insights into human cognition. By studying ChatGPT's behavior, we can deepen our understanding of human linguistic processes. In particular, the study of ChatGPT's handling of malapropisms offers a unique opportunity to explore the "reality of language" from a constructivist perspective, revealing aspects of language use that might otherwise remain hidden.

During this study, we explored whether ChatGPT, as a probabilistic generative model, can identify and correctly interpret malapropisms. It is possible that ChatGPT treats malapropisms as "bad data" or "noise," leading to their exclusion from its learned models. This technical consideration raises a broader question: Could all linguistic phenomena, including malapropisms, ultimately be explained through probabilistic models? If so, Davidson's passing theory, which emphasizes the need for continuous reinterpretation in language, might be less relevant than previously thought.

Davidson's passing theory suggests that understanding malapropisms involves more than just conventional meanings; it requires constructing a new interpretation for each instance. However, if language can be reduced to probabilistic calculations, malapropisms might simply be considered part of habitual language use rather than novel, case-by-case inventions. This perspective challenges the need for theoretical constructs like passing theory and suggests that ChatGPT's relatively high accuracy in detecting certain types of malapropisms might be due to its ability to learn and predict linguistic patterns with high probabilistic accuracy.

This line of reasoning raises further questions about human language. If creativity and linguistic innovation can be understood as a matter of combining elements within a probabilistic framework, then perhaps the distinction between human and AI language processing is not as profound as once believed. From this viewpoint, the language model of generative AI presents a challenge to Davidson's argument, suggesting that some aspects of human language may be less unique than previously thought.

Of course, these conclusions are drawn from theoretical considerations, and future empirical testing will be necessary to verify these ideas. The probabilistic approach may also offer insights into the differences between malapropisms produced by aphasic patients and those by healthy individuals. By studying generative AI like ChatGPT, we gain valuable perspectives on the complexities of human language that are difficult to grasp through traditional methods.

5. Conclusion

This study highlights significant challenges in ChatGPT's ability to fully understand and interpret malapropisms, particularly those involving subtle phonological and lexical variations. While ChatGPT demonstrates impressive fluency in general conversation, its performance in handling linguistic anomalies remains limited, revealing a clear distinction between AI-driven language processing and human linguistic comprehension.

The findings of this study should be interpreted with caution due to several limitations. First, the dataset used was limited to Japanese malapropisms, which may not be representative of other languages or cultural contexts. Additionally, the sample size, both in terms of the dataset and the number of human evaluators, was relatively small. These factors may have influenced the results, and larger-scale studies are needed to validate the findings. Furthermore, the study relied solely on textual data, which lacks the multimodal context (e.g., audio, visual cues) that often aids human understanding. This raises questions about whether a more holistic approach, incorporating multiple forms of data, could improve generative AI's performance in understanding complex language phenomena.

For learners using ChatGPT as a tool for language acquisition, it is essential to recognize both its strengths and limitations. While ChatGPT can serve as a valuable resource for practicing conversational skills and exploring language nuances, it is not infallible, particularly when it comes to understanding irregular language patterns such as malapropisms. Educators and learners should approach ChatGPT as a supplementary tool, complementing traditional language learning methods with AI-driven interaction. Additionally, learners should be encouraged to critically evaluate generative AI's responses and seek clarification when uncertain, fostering a more active and reflective approach to language learning.

As AI technology continues to evolve, the gap between human and machine language comprehension may gradually narrow. The insights gained from this research contribute to ongoing discussions about the role of AI in language learning and processing, and they provide a foundation for future studies aimed at enhancing generative AI's capabilities while also acknowledging the unique aspects of human linguistic experience.

References

Davidson, D. (1986). A Nice Derangement of Epitaphs. In Ernest Lepore (ed.), *Truth and Interpretation: Perspectives on the Philosophy of Donald Davidson*. Blackwell. 433-446.

Gubelmann, R. (2023). A Loosely Wittgensteinian Conception of the Linguistic Understanding of Large Language Models like BERT, GPT-3, and ChatGPT. *Grazer Philosophische Studien, 99*(4), 485-523. https://doi.org/10.1163/18756735-00000182

McGregor, S. (2023). *Is ChatGPT Really Disembodied?* Paper presented at the ALIFE 2023: Ghost in the Machine: Proceedings of the 2023 Artificial Life Conference. https://doi.org/10.1162/isal_a_00597

Piattelli-Palmarini, Massimo (ed.). (1980). *Language and Learning: The Debate Between Jean Piaget and Noam Chomsky*. Harvard University Press.

Rao, V., & Chandra. (2017). A Brief Study of Words Used in Denotation and Connotation. *Journal for Research Scholars and Professionals of English Language Teaching, 1*(1).

Tamaoka, K., & Terao, Y. (2004). Mora or syllable? Which unit do Japanese use in naming visually presented stimuli? *Applied Psycholinguistics, 25*(1), 1-27. https://doi.org/10.1017/S0142716404001018

Titus, L. M. (2024). Does ChatGPT have semantic understanding? A problem with the statistics-of-occurrence strategy. *Cognitive Systems Research, 83*, 101174. https://doi.org/10.1016/j.cogsys.2023.101174

Wu, T. et al. (2023). A Brief Overview of ChatGPT: The History, Status Quo and Potential Future Development. in *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122-1136. https://doi.org/10.1109/JAS.2023.123618