# 国際日本学研究科

# 修士学位請求論文要旨

〔論文題名〕

**A Comparative Study of Linguistic Complexity in Human and AI-Generated English Essays Using Average Dependency Distance and Propositional Idea Density**

| 国際日本学専攻　博士前期課程 | |
|---|---|
| 英語教育学　　研究領域 | |
| 入学年度 | ２０２４年度 |
| 学生番号 | ４９１１１２４５００２ |
| 氏名 | 貝瀬 結風 |
| 指導教員 | 大矢 政徳 |
| 提出日 | ２０２６年　１月　９日 |

The rapid development of large language models (LLMs) has fundamentally transformed contemporary language use and has raised urgent questions concerning the nature of linguistic complexity, human cognition, and second language (L2) writing. Generative AI such as ChatGPT are now capable of producing texts that appear fluent, coherent, and stylistically sophisticated, often comparing human performance with conventional evaluations of writing quality. However, surface-level proficiency alone provides limited insight into the cognitive principles underlying language production. From a psycholinguistic perspective, language use is deeply constrained by human cognitive architecture, particularly working memory (WM), and linguistic complexity reflects the ways in which speakers and writers allocate limited cognitive resources during real-time processing. This study aims to investigate how linguistic complexity manifests differently in texts written by Japanese learners of English, native English speakers (ENS), and Generative AI, and to clarify whether AI-generated language adheres to or diverges from cognitively motivated constraints that shape human language production.

WM has long been recognized as a central component of language processing, serving as a limited-capacity system responsible for temporarily storing and manipulating linguistic information (Baddeley & Hitch, 1974; Baddeley, 2000). In

second language acquisition (SLA), WM capacity has been shown to play a crucial role in grammatical development, syntactic processing, and writing performance, particularly under cognitively demanding conditions (Miyake & Friedman, 1998; Juffs & Harrington, 2011; Wen, 2016). Cognitive Load Theory (CLT) further refines this perspective by distinguishing among intrinsic, extraneous, and germane cognitive load (CL) and by emphasizing that performance deteriorates when total load exceeds WM capacity (Sweller, 1988, 1994). Applied to L2 writing, CLT predicts systematic trade-offs among syntactic complexity, semantic density, and processing efficiency, especially when learners are required to produce extended texts. Despite the theoretical relevance of WM and CLT, many empirical studies of L2 writing have relied on surface-level indices of complexity, such as sentence length or subordination ratios, which only indirectly reflect cognitive processing demands and may obscure the mechanisms underlying learner performance.

To address this limitation, the present study adopts a dependency-based approach to syntactic complexity and a proposition-based approach to semantic complexity, both of which are closely linked to cognitive processing. Dependency Grammar (DG) conceptualizes sentence structure as a network of asymmetric relations between heads and dependents (Tesnière, 1959; Hudson, 1984), offering a representation that aligns

well with incremental processing models. Within this framework, dependency distance (DD)—the linear distance between syntactically related words—has been proposed as a direct correlation of processing difficulty, as longer dependencies require linguistic information to be maintained in WM for extended periods (Liu, 2008). The Dependency Distance Minimization (DDM) hypothesis posits that natural language (NL) tends to minimize DDs to reduce CL, a tendency that has been observed across typologically diverse languages and genres (Futrell et al., 2015; Temperley, 2007, 2008). Average Dependency Distance (ADD), calculated as the mean distance between heads and dependents in a sentence or text, therefore provides a theoretically grounded and cognitively interpretable index of syntactic complexity.

Complementing syntactic measures, Propositional Idea Density (PID) captures semantic and conceptual complexity by quantifying the number of propositions expressed per unit of text (Kintsch, 1974; Turner & Greene, 1977). Propositions represent basic units of meaning, typically involving predicates and their arguments, and higher PID values indicate denser conceptual packaging. PID has been widely used in cognitive psychology and aging research, where it has been shown to predict comprehension difficulty and long-term cognitive outcomes (Kemper, 1987; Snowdon et al., 1996). More recently, PID has been applied to cross-linguistic and translation

studies, demonstrating its utility as a language-independent measure of semantic complexity (Oya, 2023, 2024). From a WM perspective, producing or processing texts with high PID imposes substantial cognitive demands, as multiple propositions must be planned, integrated, and maintained simultaneously.

Although ADD and PID have each been investigated in isolation across various domains, few studies have integrated these measures within a unified analytical framework to compare human and AI-generated writing. Moreover, existing research on AI writing has primarily focused on performance-oriented outcomes such as coherence, accuracy, or human-likeness (Brown et al., 2020; Jiao et al., 2023; Gilardi et al., 2023), often neglecting the cognitive interpretability of structural features. As a result, it remains unclear whether AI-generated texts reflect the same complexity constraints that shape human language production or whether they represent fundamentally different patterns optimized through statistical learning rather than cognitive processing. The present study seeks to fill this gap by examining how syntactic and semantic complexity, as measured by ADD and PID, differ across Japanese learners of English, ENS, and Generative AI, and by interpreting these differences through the lens of WM and CLT.

The study addresses four research questions: how syntactic complexity differs among the three groups as measured by ADD; how semantic complexity differs as measured by PID; how the interaction between syntactic and semantic complexity reflects different patterns of CL allocation; and to what extent AI-generated writing resembles or diverges from human writing in terms of cognitively motivated constraints. To answer these questions, the study draws on three comparable corpora. The learner corpus consists of argumentative essays written by Japanese learners of English from the ICNALE corpus (Ishikawa, 2013), representing intermediate to advanced proficiency levels. The native speaker corpus comprises essays written by ENS under similar task conditions. The AI corpus consists of essays generated by ChatGPT (GPT-4o) in response to the same prompts, ensuring comparability in topic, genre, and communicative purpose. Texts were controlled for length and preprocessing procedures were applied to minimize confounding effects unrelated to syntactic or semantic structure.

All texts were annotated using the Universal Dependencies (UD) framework (de Marneffe et al., 2021; Nivre et al., 2020), which provides a cross-linguistically consistent and cognitively motivated representation of grammatical relations. Dependency parses were generated using state-of-the-art parsers and manually

inspected to identify systematic errors. ADD was calculated by measuring the linear distance between each head and dependent pair and averaging these values across sentences. PID was calculated by identifying propositional units based on predicate–argument relations and functional elements, following established procedures in cognitive and discourse-based analyses. Statistical analyses included descriptive statistics, distributional analyses, and inferential tests to examine group differences and interaction patterns.

The present study set out to examine differences in linguistic complexity across English essays written by Japanese learners of English, ENS, and a ChatGPT, adopting a cognitively grounded perspective. By employing ADD as a measure of syntactic complexity and PID as a measure of semantic complexity, the study aimed to move beyond surface-level indicators of writing quality and to explain the underlying principles that shape language production in humans and AI. The results revealed clear, systematic, and theoretically meaningful differences among the three groups, reflecting the fact that linguistic complexity is governed by fundamentally different mechanisms in NL and AI-generated language.

With respect to syntactic complexity, as indexed by ADD, Japanese learners of English exhibited the lowest values, followed by ChatGPT, while ENS showed the

highest ADD values. This ordering suggests that Japanese learners tend to produce sentences with shorter dependency relations, relying on relatively simple syntactic structures. From a cognitive perspective, this pattern can be interpreted as a strategy for coping with limited WM capacity. In L2 production, grammatical representations are often less automatized than in native language use, which increases the cognitive cost of syntactic processing. As a result, L2 learners may avoid constructions that involve long DDs, thereby reducing the burden on WM during sentence planning and production. Shorter DDs allow learners to keep syntactically related elements active over a shorter span, minimizing the risk of processing breakdown.

In contrast, ENS demonstrated the highest ADD values, indicating an ability to sustain longer dependency relations within sentences. This finding suggests greater flexibility and efficiency in syntactic processing, likely attributable to highly automatized grammatical knowledge. For ENS, maintaining longer dependencies does not necessarily impose excessive CL, as syntactic operations are largely automatized and require fewer WM resources. Importantly, higher ADD values in ENS writing should not be interpreted as inefficiency. Rather, they reflect the fact that ENS can afford to prioritize discourse-level goals, stylistic variation, or semantic clarity over strict minimization of DD. In this sense, the results are compatible with the DDM

hypothesis, which posits that DDs tend to be minimized within cognitively tolerable limits rather than reduced to an absolute minimum.

ChatGPT occupied an intermediate position in terms of ADD, producing DDs longer than those of Japanese learners but shorter than those of ENS. This pattern is particularly noteworthy because ChatGPT does not operate under human cognitive constraints such as WM limitations. Nevertheless, the AI system consistently generated syntactically compact structures with relatively short DDs. This suggests that the observed ADD values in AI-generated texts are not the result of online processing constraints but rather reflect statistical regularities learned from large-scale training data. In other words, ChatGPT appears to favor syntactic configurations that are frequent, stable, and broadly acceptable across contexts, resulting in DDs that are neither minimal nor extreme. This intermediate ADD profile highlights the importance of distinguishing between cognitively motivated constraints and distributional optimization when interpreting complexity measures in AI-generated language.

Turning to semantic complexity, as measured by PID, a different pattern emerged. Japanese learners consistently exhibited lower PID values than ChatGPT, indicating a more limited capacity to package multiple propositions within a given textual unit. PID captures the density of meaning expressed in a text by quantifying the number of

propositional units, such as predicates and their arguments. Lower PID values in learner writing suggest that learners tend to express ideas in a more linear and less condensed manner, often distributing meaning across multiple clauses or sentences. From a WM perspective, producing or integrating multiple propositions within a single sentence requires substantial cognitive resources, as the writer must simultaneously plan, encode, and maintain several semantic units. For L2 learners, this level of semantic integration may exceed available WM capacity, leading them to reduce propositional density as a compensatory strategy.

ChatGPT, by contrast, produced texts with higher PID values, demonstrating an ability to generate semantically dense language. In many cases, and depending on the topic, the PID values of AI-generated texts were comparable to those observed in ENS writing. On the surface, this similarity might suggest that AI-generated texts achieve a level of semantic sophistication comparable to that of human expert writers. However, it is crucial to recognize that this similarity is limited to the outcome rather than the process. While native speakers achieve high PID through cognitively managed planning and integration of meaning, ChatGPT does so without any underlying CL, relying instead on probabilistic associations between linguistic forms and meanings learned from data.

ENS writing exhibited relatively high PID values alongside higher ADD values, indicating an ability to balance syntactic and semantic complexity effectively. This balance suggests that ENS can maintain dense conceptual expression while allowing for longer syntactic dependencies, without exceeding WM capacity. Such a pattern reflects an optimized allocation of cognitive resources, where syntactic efficiency and semantic richness are jointly managed. The coexistence of moderate-to-high ADD and high PID in ENS texts underscores the importance of considering multiple dimensions of complexity simultaneously. Evaluating syntactic or semantic complexity in isolation would obscure the nuanced trade-offs that characterize proficient language use.

When syntactic and semantic complexity are considered together, the three groups display qualitatively distinct profiles of complexity management. Japanese learners appear to simplify both syntactic structure and semantic density, producing texts characterized by low ADD and low PID. This pattern reflects the cognitive constraints faced by L2 writers, who must divide limited WM resources among grammatical encoding, lexical retrieval, semantic planning, and discourse organization. The resulting trade-offs lead learners to prioritize stability and accuracy over complexity, yielding texts that are structurally simpler and conceptually less dense.

ENS, in contrast, demonstrate a capacity to optimize the relationship between syntax and semantics. Their writing exhibits higher ADD and high PID, suggesting that they can sustain complex dependency structures while simultaneously packaging rich semantic content. This optimized balance aligns with cognitively motivated accounts of language production, which propose that skilled language users allocate cognitive resources efficiently across multiple levels of representation. The ENS profile thus serves as an empirical benchmark for cognitively sustainable complexity in human writing.

AI-generated texts represent a third, distinct pattern. ChatGPT produces syntactically compact sentences with relatively high semantic density, resulting in a combination of moderate ADD and high PID. However, a key characteristic of AI-generated language is the markedly reduced variability in both ADD and PID distributions compared to human writing. Human texts, especially those produced by learners, exhibit a wide range of values, reflecting individual differences, situational demands, and moment-to-moment fluctuations in CL. In contrast, AI-generated texts cluster within a narrow band of values, rarely exhibiting extreme simplicity or complexity. This lack of variability suggests that ChatGPT operates within a constrained

space of statistically optimal configurations, shaped by the central tendencies of its training data rather than by adaptive responses to cognitive constraints.

The reduced variability observed in AI-generated texts has important theoretical implications. Human language production is inherently variable, influenced by cognitive capacity, task demands, emotional state, and communicative goals. Variability is not a flaw but a defining feature of human language use. The relative uniformity of AI-generated complexity, by contrast, points to a fundamentally different mode of language generation. Rather than dynamically adjusting complexity in response to internal constraints, AI systems generate language by selecting forms that are most probable given prior patterns. As a result, AI-generated texts may appear consistently fluent and well-structured yet lack the adaptive complexity modulation characteristic of human writing.

From a cognitive standpoint, the contrast between human and AI writing underscores the central role of WM constraints in shaping linguistic complexity. For human writers, especially L2 learners, complexity emerges from the interaction between linguistic knowledge and cognitive limitations. ADD and PID reflect not only linguistic proficiency but also the ability to manage CL during production. AI-generated complexity, in contrast, is non-cognitive in nature. Although AI systems can

approximate human-like patterns on surface measures, the principles governing their output differ fundamentally from those underlying human language use. This distinction lends empirical support to critiques of LLMs as systems that generate language through probabilistic pattern matching rather than genuine understanding.

The pedagogical implications of these findings are substantial. In L2 writing instruction, AI-generated texts are increasingly used as reference models or learning aids. However, the high semantic density and compact syntactic structures characteristic of AI-generated language may impose unrealistic cognitive demands on learners if presented as ideal targets. Learners attempting to emulate AI-generated texts may struggle to manage the simultaneous demands of syntax and meaning, potentially leading to cognitive overload and reduced learning effectiveness. Instruction informed by CLT should therefore scaffold complexity in a manner that aligns with learners' WM capacities, gradually increasing syntactic and semantic demands rather than promoting direct imitation of AI output.

At the same time, the findings highlight the pedagogical value of cognitively interpretable complexity measures such as ADD and PID. Unlike traditional surface-level indices, these measures provide insight into the processing demands associated with different types of writing. They can serve as diagnostic tools for identifying learner

difficulties and for designing instructional materials that balance challenges and feasibility. In the context of assessing learners' performance, ADD and PID may also contribute to more nuanced evaluations of writing proficiency by capturing dimensions of complexity that are closely linked to cognitive processing.

Several limitations of the present study should be acknowledged. The most important limitation in the present study is that it focuses on only two metrics: ADD and PID, and future research should focus on other metrics as well. While these metrics provide theoretically grounded insights into syntactic and semantic complexity, they do not capture the full range of dimensions involved in linguistic complexity. Adopting multi-dimensional approach to complexity would allow for a more comprehensive and reliable evaluation of differences between human and AI-generated writing. In addition to that, the AI corpus represents a specific model and version of ChatGPT, and different models or prompting strategies may yield different complexity profiles. Additionally, while dependency parsing and propositional analysis were conducted using established frameworks and carefully checked, automatic annotation is inherently subject to error. Future research should extend this approach to other learner populations, genres, and modalities, including spoken language, as well as longitudinal designs that track developmental changes in complexity over time. Investigating how learners' ADD and

PID profiles evolve with increasing proficiency would provide further insight into the cognitive mechanisms underlying L2 development.

Despite these limitations, the present study demonstrates the value of integrating dependency-based syntactic analysis, propositional measures of semantic complexity, and CLT in the study of language production. By comparing Japanese learners of English, ENS, and AI-generated texts within a unified analytical framework, the study reveals fundamental differences in how linguistic complexity is structured and constrained across agents. While generative AI systems can produce linguistically sophisticated texts that rival or even surpass human output on surface measures, the principles governing their complexity are fundamentally distinct from those that shape human language production.

In conclusion, this study provides a cognitively grounded account of linguistic complexity in L2 learner writing, native speaker writing, and AI-generated texts. The findings demonstrate that human linguistic complexity is deeply shaped by WM constraints and CL management, whereas AI-generated complexity reflects statistical optimization unconstrained by human cognition. Recognizing this distinction is essential for advancing theoretical understanding of language complexity, refining

evaluation practices, and making informed pedagogical decisions in an era of increasingly pervasive AI-generated language.

**Reference**

Baddeley, A. D., & Hitch, G. (1974). Working memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 8, pp. 47–89). Academic Press.

Baddeley, A. D. (2000). The episodic buffer: A new component of working memory? *Trends in Cognitive Sciences*, *4*(11), 417–423. https://doi.org/10.1016/S1364-6613(00)01538-2

Bender, E. M., Gebru, T., McMillan-Major, A., & Shmitchell, S. (2021). On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)* (pp. 610–623). ACM. https://doi.org/10.1145/3442188.3445922

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., … Amodei, D. (2020). Language models are few-shot learners. In *Advances in Neural Information Processing Systems (NeurIPS)* (Vol. 33, pp. 1877–1901). https://arxiv.org/abs/2005.14165

Futrell, R., Mahowald, K., & Gibson, E. (2015). Large-scale evidence of dependency length minimization in 37 languages. *Proceedings of the National Academy of Sciences, 112*(33), 10336–10341. https://doi.org/10.1073/pnas.1502134112

Gilardi, F., Alizadeh, M., & Kubli, M. (2023). ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences, 120*(6), e2214347120. https://doi.org/10.1073/pnas.2214347120

Hudson, R. A. (1984). *Word Grammar*. Oxford, UK: Blackwell.

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world* (Vol. 1, pp. 91–118). Kobe University.

Jiao, W., Wang, W., Huang, J., Wang, X., & Tu, Z. (2023). Is ChatGPT a good writer? A case study of essay writing. *arXiv preprint* arXiv:2301.07069.

Juffs, A., & Harrington, M. (2011). Aspects of working memory in L2 learning. *Language Teaching, 44*(2), 137–166. https://doi.org/10.1017/S0261444810000506

Kemper, S. (1987). Life-span changes in syntactic complexity. *Journal of Gerontology, 42*(3), 323–328.

Kintsch, W. (1974). *The representation of meaning in memory.* Hillsdale, NJ: Lawrence Erlbaum Associates.

Liu, H. (2008). Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science, 9*(2), 159–191.

de Marneffe, M.-C., Manning, C., Nivre, J., & Zeman, D. (2021). Universal Dependencies. *Computational Linguistics, 47*(2), 255–308.

Miyake, A., & Friedman, N. P. (1998). Individual differences in second language proficiency: Working memory as language aptitude. In A. Healy & L. Bourne (Eds.), *Foreign language learning: Psycholinguistic studies on training and retention* (pp. 339–364). Lawrence Erlbaum Associates.

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C., Pyysalo, S., Schuster, S., Tyers, F., & Zeman, D. (2020). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the 12th International Conference on Language Resources and Evaluation (LREC 2020)*, (pp.4034–4043). Marseille, France.

Oya, M. (2023). Propositional idea density of a Japanese text and its English translation in a parallel corpus. *Global Japanese Studies Review, 15*(1), 97–105.

Oya, M. (2025). Complexity of English Sentences in English Textbooks for High School Students in Japan. *Global Japanese Studies Review 17*(1) 1-10.

Snowdon, D. A., Kemper, S. J., Mortimer, J. A., Greiner, L. H., Wekstein, D. R., & Markesbery, W. R. (1996). Linguistic ability in early life and cognitive function and Alzheimer's disease in late life: Findings from the Nun Study. *JAMA, 275*(7), 528–532.

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science, 12*(2), 257–285. https://doi.org/10.1207/s15516709cog1202_4

Sweller, J. (1994). Cognitive load theory, learning difficulty, and instructional design. *Learning and Instruction, 4*(4), 295–312. https://doi.org/10.1016/0959-4752(94)90003-5

Temperley, D. (2007). Minimization of dependency length in written English. *Cognition, 105*(2), 300–333. https://doi.org/10.1016/j.cognition.2006.09.011

Temperley, D. (2008). Dependency-length minimization in natural and artificial languages. *Journal of Quantitative Linguistics, 15*(3), 256–282. https://doi.org/10.1080/09296170802159512

Tesnière, L. (1959). *Éléments de syntaxe structurale.* Paris: Klincksieck.

Turner, A., & Greene, E. (1977). *The construction and use of a propositional text base.* Boulder, CO: Institute for the Study of Intellectual Behavior, University of Colorado.

Wen, Z. (2016). *Working memory and second language learning: Towards an integrated approach.* Multilingual Matters.