ISSN: 2716-9189 E-ISSN: 2676-1750
Journal of Studies in Language, Culture, and Society (JSLCS). (08) 01, 2025 (PP. 295-309 )
Received: 22/02/2025 Accepted: 08/04/2025 Published: 30/04/2025

# SYNTACTIC AND LEXICAL COMPARISON BETWEEN AI-GENERATED READING PASSAGES AND JAPANESE UNIVERSITIES' NATIONAL TEST

**Satoshi Kurokawa** [1] **Maelys Salingre** [2]
[1] **Nagoya University (Japan)**
kurokawa.satoshi.g6@f.mail.nagoya-u.ac.jp
[2] **Shizuoka University (Japan)**
salingre.maelys@shizuoka.ac.jp

**Abstract:** This research aimed to develop a methodology for creating mock English reading tests using AI-generated passages to help decrease the workload of EFL teachers who struggle in creating the high-quality mock English reading test for universities' entrance examinations. To achieve this goal, this paper examined the lexical and syntactical differences between English Subject of the Center Test at Japanese universities' entrance exams (ESCT) and AI-generated passages. Three different generative AI were used: OpenAI ChatGPT version 4, Google Gemini version 1.5 Flash, and DeepSeek-V3. To make the vocabulary coverage between AI-generated passages and ESCTs meaningful, topics based on ESCTs were used to create 11 prompts for AI-generated passages. This paper examined text coverage using CEFR-based wordlist and syntactic complexity using the Python library spaCy. Findings revealed that the proportion of A2 level tokens does not differ greatly between AI-generated passages (ChatGPT: 19.1%; Gemini: 17.4%; DeepSeek: 17.3%) and ESCT (15.6%); ESCT had more complex but shorter sentences comparing to AI-generated passages, and personal pronouns accounted for 1.3% of ESCT tokens, while they accounted for less than 1% of generative AIs tokens (ChatGPT, 0.39% ; Gemini, 0.71% ; DeepSeek 0.45%). Few wh-pronouns and the existential *there* were found in AI-generated passages. This study concluded that when the EFL teachers convert AI-generative text to the mock ESCT, they should (a) add a wider variety of A1 level lemmas, (b) rewrite more complex and shorter sentences, (c) increase personal pronouns and determiners, and (d) reduce adjective modifiers, conjuncts, and coordination.

***Keywords***: Generative AI, natural language processing, text coverage, syntactic complexity, university entrance examination

**How to cite the article :**
Kurokawa, S. & Salingre, M. (2025). Syntactic and Lexical Comparison Between AI-Generated Reading Passages and Japanese Universities' National Test. *Journal of Studies in Language, Culture, and Society (JSLCS), 8*(1), 295-309

[1] **Corresponding author :** Satoshi Kurokawa **ORCID ID:** https://orcid.org/0009-0009-0325-806X

## 1. Introduction

Owing to the rapid technological improvements in AI, the language-teaching environment has the potential to change dramatically. Since the COVID-19 pandemic, the education system has shifted to online education, and the trend of AI-based content generation tools continues to increase (Qazi et al., 2024). Wang et al. (2023) showed that AI applications in education have significant implications and that educational institutions must strive to maximize the benefits and minimize the risks associated with their implementation. As Jiang (2022) noted, AI tools such as neural machine translation tools, intelligent tutoring systems, AI chatbots, and Virtual Reality (VR) tools are widely used in the English as a Foreign Language (EFL) teaching context. Generative AI is being used in the classroom to help students (e.g., Ho, 2023; Young & Shishido, 2023 *inter alia*). Ho (2023) used ChatGPT as a tool to develop ESL learners' paraphrasing skills. Young and Shishido (2023) investigated the potential of OpenAI's ChatGPT for generating chatbot dialogues for EFL.

Generative AI may contribute to decreasing EFL teachers' workload. In Japan, for example, although the Ministry of Education, Culture, Sports, Science, and Technology (MEXT) (2019) announced a guideline on reducing teachers' workload, in 2022, teachers still worked an average of 10 hours per day on weekdays and 2 hours per day on weekends, according to a survey on teachers' workload conducted by MEXT (2023). Along with this regular workload, teachers are expected to help students prepare for university entrance examinations. In Japanese society, as well as in other East Asian countries, the findings suggest that entering a prestigious university provides a considerable advantage in one's career after graduation (Erikawa, 2011). Hence, high school teachers are under considerable pressure to prepare students for the nationwide first-round general test for national and public universities: the Japanese universities' National Test (Center Test), conducted until 2020, and the Common Test for University Admissions (Common Test), which replaced it in 2021. Although preparing students for these tests is a crucial task, many high school EFL teachers are already overworked and may not have enough time to analyze past tests or conduct reliable mock tests. Furthermore, most Japanese high school EFL teachers speak Japanese as a native language; thus, when they prepare reading comprehension tests, those tests must be proofread by native English speakers or they must search for proofread passages. Furthermore, relying on past tests is difficult, as the Center Test and Common Test are conducted only once a year in two sessions. Therefore, using generative AI may contribute to reducing EFL teachers' workloads by helping them perform mock English reading tests for the Center Test and Common Test (mock ESCT).

However, little is known about the linguistic features of AI-generated texts. While many researchers have highlighted using generative AI as a support tool in education (Ho, 2023; Young & Shishido, 2023), few have explained the differences between AI-generated and human-made texts. It is important to understand these differences before conducting mock English reading tests using generative AI.

Vocabulary coverage is calculated as the percentage of sentences occupied by words on a particular list. High vocabulary coverage is essential for the test takers' accurate text comprehension (Hsueh-Chao & Nation, 2000), thus, it has been widely used in studies of the Center Test (Chujo & Hasegawa, 2004; Tani, 2008; Sakurai, 2022). However, vocabulary coverage alone may not be effective in uncovering the differences between AI-generated and human-made passages, because the syntax itself may be different. In contrast, with the advancement in Natural Language Processing (NLP), many tools are now available to easily parse texts into dependency constituents. Salingre and Kurokawa (2023) demonstrated that dependency parsing can be used to quantify the syntactic complexity of reading passages in entrance examinations at Japanese medical universities.

This study provides a comparison of vocabulary coverage and syntactic complexity between AI-generated reading passages and the Center Test reading passages. It is expected that the results of this comparison will contribute to the development of high-quality mock tests with generative AI, and eventually, help decrease EFL teachers' workload. The methods presented in this paper are also applicable to the comparison of AI-generated text with other English tests, such as the TOEFL, TOEIC, IELTS, etc. The research question (RQ) of this study is as follows:

What linguistic characteristics must be accounted for when modifying AI-generated passages to create mock ESCT?

## 2. Materials

### *2.1 Center Test*

The Center Test was designed by the National Center for University Entrance Examination in Japan (Watanabe, 2013). There were only two sessions per year: a main session and a remedial session for students who could not attend the first session because of health issues or technical problems. The Center Test aimed to measure the extent to which high school students achieved basic learning goals at the high school level and target of the test was 3rd year students who wished to enter university by taking a general entrance exam (Kuramoto, 2017). The Center Test featured a wide variety of subjects, including mathematics, science, physics, biology, Japanese language, foreign language (English, French, German, Chinese, or Korean), history, civics, and geography, depending on the chosen university's admission criteria. Watanabe (2013) explained that the Center Test was initially developed as a first-round exam for national and local public universities, although some private universities began administering the test to use the score for their own entrance exams. Based on their scores, students could apply for second-round university tests specific to each national or local public university (MEXT, 2020). Successful candidates were selected based on their scores on both the Center Test and the second-round university test.

Many prestigious national universities, such as the University of Tokyo, require students to take a foreign language test; for most students, this was the English Subject of the Center Test (ESCT). The ESCT was divided into two parts, reading and listening, and consisted of only multiple-choice questions. The reading section comprised approximately 30% of questions about basic English knowledge, pronunciation, accent, grammar, and idioms, and about 70% of reading comprehension—conversations and long passages. Among the long reading passages, at least one passage was approximately 400 words and focused on academic English.

In this study, to compare ESCT reading passages with AI-generated passages, long academic passages from each year were extracted from the Center Test past exams from 2010-2020. This type of passage was chosen because it provides the most text, is impersonal, and thus, easy to generate, unlike essays and fiction. The last 11 years of the Center Test were selected because they were conducted under the same Course of Study.

After 2020, The Center Test was replaced with the Common Test for University Admissions. The Common Test is currently the latest first-round national test. However, there were only four past tests available at the time of conducting this study, and the style of the reading tests was reformed so dramatically that few have analyzed this test. Moreover, there is still some debate regarding the significance of the Common Test over the Center Test (Kuramoto, 2017). Therefore, this study focused on the well-documented Center Test, and a comparison with the Common Test is left for future research.

*2.2 AI-generated Passages*

Three different generative AI were used: OpenAI ChatGPT version 4, Google's Gemini version 1.5 Flash, and DeepSeek's DeepSeek-V3. All passages were generated between December 2024 and January 2025. To compare the vocabulary coverage between AI-generated passages and ESCTs, their topics and styles must be similar. For each selected ESCT, the authors read the passages and extracted the main topics (Table 1). These topics were later used to create 11 prompts for the AI-generated passages. Each prompt followed the same template: "Explain [main topic] in an academic tone, without bullet points." "Without bullet points" was added after a few trials because bullet points were often generated, which would be a problem when comparing syntactic complexity.

**Table 1.**
*Topics Used for Creating the Prompts*

| ESCT year | Main topic |
|---|---|
| 2010 | popular sightseeing spots in Japan for overseas tourists |
| 2011 | shared values in communication |
| 2012 | wood used in house construction and how it must be stable in size |
| 2013 | availability and distribution of human health resources around the world |
| 2014 | state-to-state migration in the US |
| 2015 | dangers of Social Network Services |
| 2016 | fresh fruit imports to the US |
| 2017 | relation between the type of schoolyard and the physical activity of children |
| 2018 | how the color of a product can influence shoppers |
| 2019 | how art can portray clothing and social settings |
| 2020 | how training programs can enhance sports performance |

## 3. Methodology

All reading passages were tokenized, lemmatized, and parsed for dependencies using the Python library spaCy (Honnibal & Montani, 2017) and its large web-trained English model.

Vocabulary coverage and passage level were examined based on the Common European Framework of References for Languages (CEFR) using the CEFR-J wordlist version 1.6 (Tono, 2020). The CEFR-J is a version of the CEFR specifically adapted for English education in Japan, because Japanese learners tend to skew towards lower levels (Uchida & Negishi, 2021). The CEFR-J wordlist consists of four sublists, one for each level between A1 and B2, with a total 7,801 words (Tono, 2017). This wordlist was created based on English textbooks in Japan and neighboring Asian countries such as China and Korea. Furthermore, it was checked against the English Vocabulary Profile (EVP), and B1-B2 levels were adjusted to be closer to the EVP (Tono, 2019). Although the CEFR-J wordlist was made with Japanese learners in mind, it is still relevant for learners from various backgrounds. Therefore, CEFR-J was the most suitable wordlist for this study. When calculating text coverage, proper nouns were ignored; therefore, the total number of tokens and lemmas considered was slightly lower than their actual numbers.

For syntactic complexity, two metrics were calculated: sentence and passage levels. The four sentence-level metrics were based on those adopted by Salingre and Kurokawa (2023). The first metric is the sentence length, which is the number of tokens in a sentence. According to a literature review by Szmrecsányi (2004), this is one of the most common measures of syntactic complexity and correlates strongly with other measures such as the number of nodes in the constituency tree. The other three metrics are the height of the dependency tree,

maximum distance (i.e., number of tokens) between a dependent and its head, and average distance between a dependent and its head. They capture the cognitive effort required by readers to store the necessary information in their working memory, which is representative of syntactic complexity, according to von Glaserfeld (1970).

In addition to the above four sentence-level metrics, two passage-level metrics were calculated to measure the syntactic complexity of the passages. The first passage-level metric was the proportion of pronouns in a passage. This metric is specifically aimed at measuring the complexity of Japanese learners. Pronouns, especially personal pronouns, are rarely used in Japanese; thus, understanding what pronouns refer to is one of the greatest challenges faced by Japanese EFL learners (Shirahata, 2019). Consequently, if a passage contained a high proportion of pronouns, it would be more difficult for Japanese EFL learners to understand. Six different types of pronouns were investigated: personal pronouns, possessive pronouns, wh-pronouns, wh-determiners, determiners, and the existential *there*. The wh-determiners and determiners examined were pronouns and were not confused with articles. Examples of wh-determiners and determiners as pronouns are presented below:

(1) Wh-determiner

a. Wood is often painted to prevent sharp changes in moisture content, which cause expansion and shrinkage. (ESCT: 2012)

b. Then they calculated the percentage of the paintings from these countries that included each food. (ESCT: 2019)

(2) Determiner

a. The researchers offered some explanations for this. (ESCT: 2019)

b. Florida is a good example of a state that ranks high on both. (ESCT: 2014)

The last passage-level metric is the proportion of each type of dependency (e.g., adjectival modifier, noun subject, and direct object). This metric focuses on the types of dependencies because a higher command of English is required to correctly recognize a wider variety of dependencies.

## 4. Results

### *4.1 Descriptive Statistics*

Table 2 indicates descriptive statistics for each passage. The length differs between ESCT and AI-generated passages but also varies greatly among Generative AIs. Gemini generated the shortest passages, with an average of 283.2 tokens per passage; ChatGPT generated the longest ones, with an average of 432.3 tokens per passage. When focusing on length, passages generated by DeepSeek are quite similar to ESCT passages: the average number of tokens is 379.7 for DeepSeek and 383.1 for ESCT. However, when looking at the number of lemmas, passages generated by DeepSeek contained on average more lemmas (203.9) than ESCT (173.1). This means that there were more repetitions of tokens in ESCT. Only passages generated by Gemini contained fewer lemmas (151.0 on average) than ESCT; although they previously contained on average 100 fewer tokens than ESCT, there was proportionally less repetition of tokens than ESCT.

**Table 2.**

*Descriptive Statistics*

| | ESCT | | | ChatGPT | | | Gemini | | | DeepSeek | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| ID | *s* | *t* | *l* | *s* | *t* | *l* | *s* | *t* | *l* | *s* | *t* | *l* |
| 2010 | 10 | 246 | 140 | 11 | 317 | 192 | 18 | 332 | 192 | 19 | 474 | 262 |
| 2011 | 14 | 239 | 126 | 22 | 452 | 207 | 12 | 235 | 127 | 15 | 349 | 183 |
| 2012 | 20 | 374 | 164 | 25 | 451 | 212 | 15 | 293 | 140 | 18 | 409 | 199 |
| 2013 | 19 | 381 | 200 | 23 | 488 | 221 | 14 | 305 | 149 | 13 | 295 | 181 |
| 2014 | 25 | 403 | 176 | 22 | 445 | 212 | 13 | 288 | 148 | 25 | 501 | 235 |
| 2015 | 17 | 422 | 164 | 22 | 466 | 253 | 14 | 292 | 173 | 17 | 322 | 214 |
| 2016 | 21 | 473 | 195 | 19 | 489 | 242 | 12 | 269 | 145 | 15 | 428 | 225 |
| 2017 | 25 | 502 | 218 | 21 | 440 | 192 | 14 | 312 | 142 | 10 | 297 | 150 |
| 2018 | 21 | 379 | 180 | 21 | 408 | 199 | 13 | 265 | 148 | 13 | 333 | 180 |
| 2019 | 27 | 390 | 174 | 19 | 437 | 207 | 12 | 261 | 149 | 15 | 405 | 221 |
| 2020 | 24 | 405 | 167 | 19 | 362 | 186 | 16 | 263 | 148 | 16 | 364 | 193 |
| Total | 223 | 4,214 | 1,904 | 224 | 4,755 | 2,323 | 153 | 3,115 | 1,661 | 176 | 4,177 | 2,243 |
| Average | 20.3 | 383.1 | 173.1 | 20.4 | 432.3 | 211.2 | 13.9 | 283.2 | 151.0 | 16.0 | 379.7 | 203.9 |

*Note*. *s* denotes the number of sentences, *t* is the number of tokens, and *l* is the number of lemmas.

*4.2 Vocabulary Coverage*

Figure 1 demonstrates the vocabulary levels of the tokens for all passages. The proportion of each CEFR-J level was similar among all three Generative AIs but differed from that of ESCT.

**Figure 1**:

*CEFR-J Levels of Tokens*

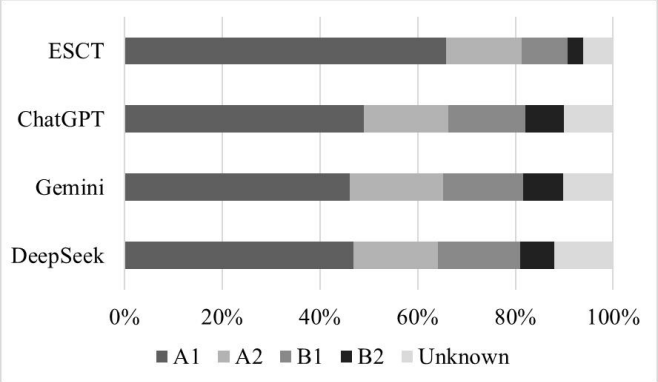
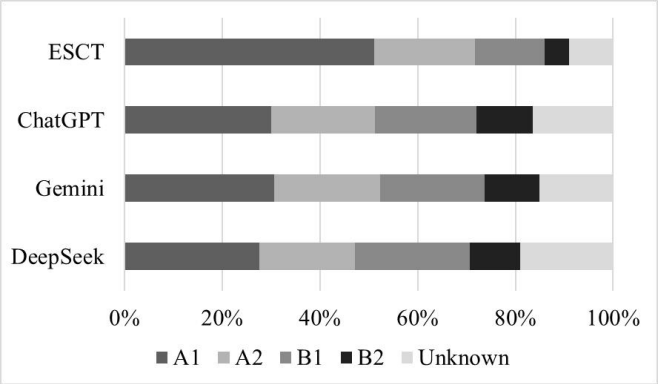



Figure 2 displays the vocabulary level of lemmas instead of tokens—tokens that are repeated several times count as only one lemma.

**Figure 2**:

*CEFR-J Levels of Lemmas*




*4.3 Sentence Length and Dependency Tree Height*

Figure 3 shows the sentence lengths for all passages. The average length of sentences in the ESCT passages is 18.9 tokens, which is shorter than ChatGPT (21.2 tokens), Gemini (20.4), and DeepSeek (23.7). However, ESCT passages include a few extremely long sentences. While the maximum sentence length is 36 tokens for Gemini, 40 tokens for DeepSeek, and 45 tokens for ChatGPT, it is 50 tokens for ESCT. ESCT passages contained four sentences with more than 40 tokens.

Figure 4 illustrates the proportion of each tree's height dependence. The proportion of trees with a height of 2 is similar for all types of passages: 69.1% for ESCT, 75% for ChatGPT, 77.1% for Gemini, and 76.7% for DeepSeek. The proportion of trees with a height of 3 is much higher for ESCT (22.4%) than for generative AIs (ChatGPT, 8.5%; Gemini, 12.4%; DeepSeek, 13.1%).

**Figure 3**: *Sentence Length*

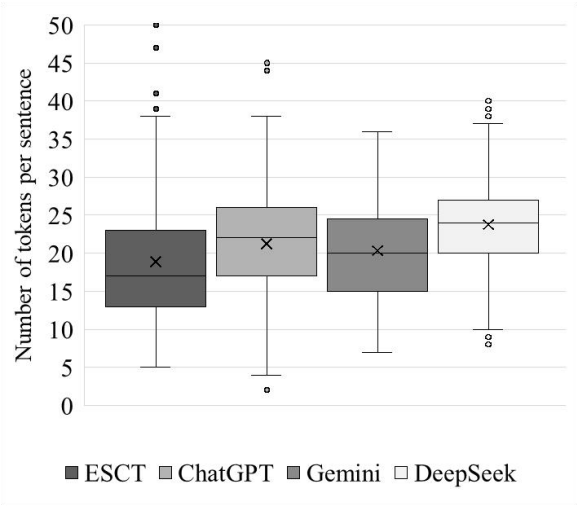
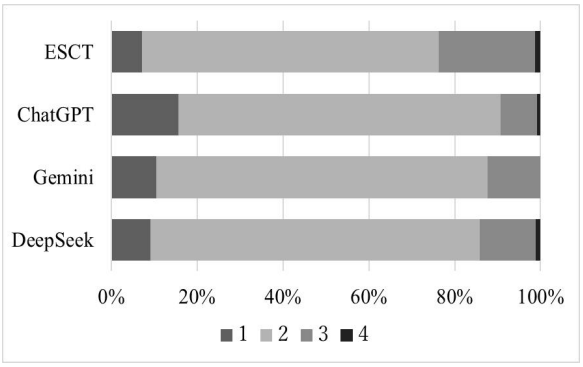



**Figure 4**: *Dependency Tree Height*




*4.4 Maximum and Average Distance between Dependents and Heads*

Figures 5 and 6 display the distance between dependents and heads. In Figure 5, the average maximum distance between dependents and heads in a sentence is slightly shorter for ESCT (8.0 tokens) than for generative AIs (ChatGPT: 8.3 tokens; Gemini: 8.4 tokens; DeepSeek: 9.7 tokens). The maximum distance between dependents and heads of ESCT was significantly shorter than DeepSeek ($t(397) = -3.72$, $p < .01$).

**Figure 5**: *Maximum Distance between Dependents and Heads*

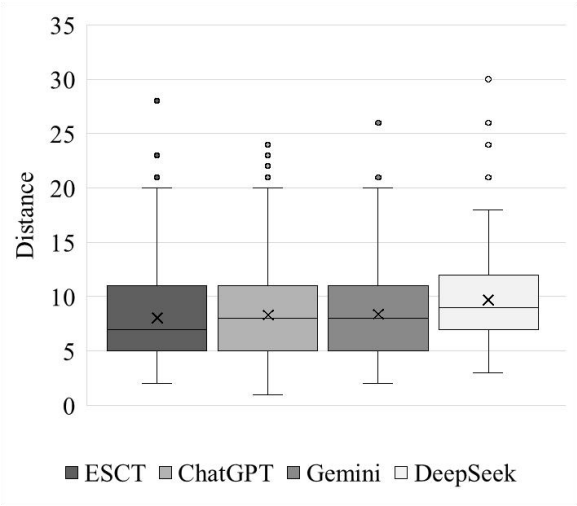
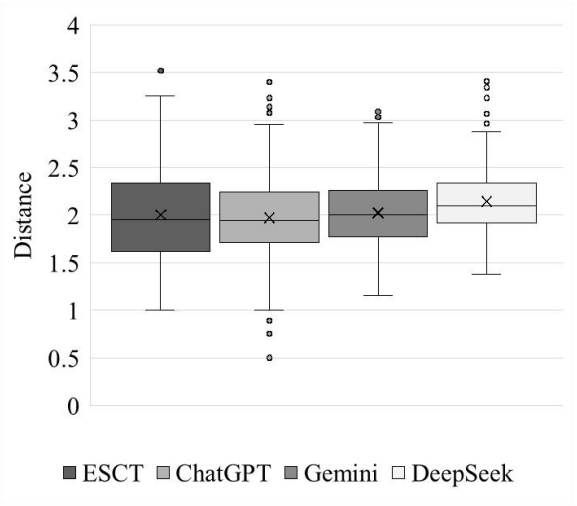



**Figure 6**: *Average Distance between Dependents and Heads*




The average distance between the dependents and heads in a sentence (see Figure 6) was similar for all passages; the average point was between 1.9 and 2.1. The maximum average distance in a sentence was slightly lower for Gemini (3.1) and ESCT (3.5) than for the other passage types (ESCT: 3.5; ChatGPT: 3.4; DeepSeek: 3.4). The minimum average distance for ChatGPT was 0.5, which corresponds to section titles. Regarding the maximum distances, DeepSeek showed the lowest standard deviation (0.36) for average distances, followed by Gemini (0.42), ChatGPT (0.45), and ESCT (0.48).

*4.5 Percentage and proportion of each type of pronouns*

Table 3 illustrates the percentage of each type of pronoun in each passage type. ESCT had the highest proportion of pronouns (3.73%), followed by ChatGPT (2.73%), Gemini (2.44%), and DeepSeek (2.23%).

**Table 3.**

*Percentage of Type of Pronouns (all passages)*

| | ESCT | ChatGPT | Gemini | DeepSeek |
|---|---|---|---|---|
| Personal pronouns | 1.38% | 0.39% | 0.71% | 0.45% |
| Possessive pronouns | 0.81% | 0.96% | 0.71% | 0.67% |
| Wh-pronouns | 0.21% | 0.06% | - | 0.05% |
| Wh-determiner | 0.55% | 0.98% | 0.80% | 0.84% |
| Determiner | 0.57% | 0.28% | 0.22% | 0.22% |
| Existential *there* | 0.21% | 0.06% | - | - |
| Total | 3.73% | 2.73% | 2.44% | 2.23% |

Finally, 40 different types of dependency relationships were found in ESCT passages, 36 in ChatGPT, 33 in Gemini, and 35 in DeepSeek passages. No type was only found in AI-generated passages, and parataxis and quantifier phrase modifiers were only found in ESCT passages. Examples of parataxis and quantifier phrase modifiers are given below (the dependent is underlined and the head is in italics).

(3) Parataxis

> […] the volume of US orange imports has *grown* steadily since the 1990s, with occasional sudden increases when the US crop experienced freezing weather (see Figure 1). (ESCT: 2016)

(4) Quantifier phrase modifier

> No more than *two* consecutive throws were allowed from the same location for this group. (ESCT: 2020)

Other relation types absent from some AI-generated passages include dative and object predicates (not found in Gemini), negation modifiers (not found in DeepSeek), and predeterminers (not found in either ChatGPT or Gemini). The negation adverb "not" was found in DeepSeek; however, it was solely in the collocation "not only," where its relation to "only" is a preconjunction, and not a negation modifier. Finally, as seen in Table 3, because the existential *there* is not present in the Gemini and DeepSeek passages, there are also no expletive relations in these two types of passages. Illustrations of these dependency relations are provided below.

(5) Dative

> Thus, knowing how schoolyards are used by students may *give* us some helpful ideas to promote their physical activity. (ESCT: 2017)

(6) Object predicate

> The other is to put it in a special oven *called* a kiln. (ESCT: 2012)

(7) Negation modifier

> However, the results here *were* not clear. (ESCT: 2015)

(8) Predeterminer

> Another important factor is a country's health care spending, shown in Table 1 as a percentage of its gross domestic product (GDP), or the total value of all its *goods* and services. (ESCT: 2013)

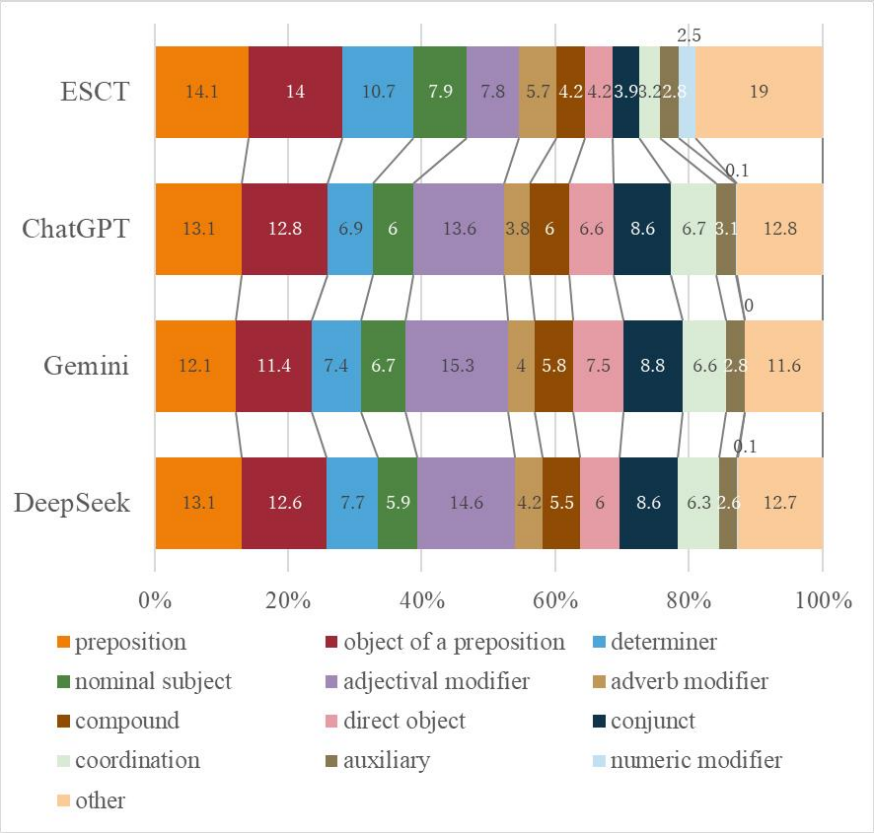
(9) Expletive

> There *were* also states that were only magnet or only sticky. (ESCT: 2014)

Figure 7 shows the proportions of each type of dependency. Dependencies that accounted for less than 2% in every type of passage are counted in "others."

**Figure 7**:

*Proportion of each Dependency Relation*




## 5. Discussion

This study aimed to compare vocabulary coverage and syntactic complexity between AI-generated reading passages and actual Center Test reading passages. To answer the RQ "What are the linguistic characteristics that must be accounted for when modifying AI-generated passages to create mock ESCT?" this study concludes that if EFL teachers want to adjust AI-generative texts to be similar to ESCT, they should do as follows:

(a) Add a wider variety of A1 level lemmas

(b) Rewrite more complex and shorter sentences

(c) Increase the number of personal pronouns and determiners

(d) Reduce the number of adjective modifiers, conjuncts and coordination

Claim (a) shows the results of the vocabulary coverage analysis using the CEFR-J wordlist (Figures 1 and 2). For example, A1 level tokens represent 48.9% of tokens for passages generated by ChatGPT, 46.2% for those generated by Gemini, and 46.9% for those generated by DeepSeek. In comparison, A1 level tokens comprise 65.8% of ESCT passages. Conversely, the proportions of B1 and B2 level tokens were higher for AI-generated passages (ChatGPT: 15.8% for B1 and 7.8% for B2; Gemini: 16.3% for B1 and 8.1% for B2; DeepSeek: 16.7% for B1 and 7.1% for B2) than for ESCT (9.4% for B1 and 3.1% for B2). Only the proportion of A2 level tokens did not differ significantly between AI-generated passages (ChatGPT: 19.1%; Gemini: 17.4%; DeepSeek: 17.3%) and ESCT (15.6%). In the case of lemmas, a few A1 level lemmas were used repeatedly in AI-generated passages, whereas a wide variety of A1 level lemmas were used in the ESCT passages. The text coverage of 93.8%, 89.9%, 89.7%, and 88.0% for ESCT, ChatGPT, Gemini, and DeepSeek,

respectively, indicate that the proportion of unknown tokens, which are likely to be C1 or C2 level, is higher in AI-generated passages. The trend demonstrated in Figure 1 is clearer in Figure 2; the proportion of A1 level lemmas was higher for ESCT (51.1%) than for AI-generated passages (ChatGPT, 30.0%; Gemini, 30.7%; DeepSeek, 27.2%). The proportion of A1 level lemmas was much lower than the proportion of A1 level tokens for AI-generated passages indicating that a few A1 level lemmas were used repeatedly, whereas in ESCT passages, a wider variety of A1 level lemmas were used.

Therefore, EFL teachers do not need to change A2 level tokens, but it is necessary to add a wider variety of A1 level lemmas. This tendency for the ESCT to contain more A1 level tokens and lemmas may be because it is tailored for Japanese high school EFL learners.

The claim in (b) is based on the analysis of sentence length, dependency tree height, and the distance between dependents and heads, which are representative of syntactic complexity (see Figures 3－6). ESCT and ChatGPT both have higher standard deviation (8.36 for ESCT and 7.28 for ChatGPT) than Gemini (6.51) and DeepSeek (6.35), which are similar. Although ESCT passages tend to skew toward shorter sentences, they contain many long sentences. Gemini and DeepSeek contain longer sentences and show less variation in sentence length. ChatGPT tends to contain longer sentences, but also many short sentences, which are mostly section titles. A t-test was used to analyze the data. It revealed that the text length of ESCT was significantly shorter than ChatGPT ($t(445) = -3.14$, $p < .01$), and DeepSeek ($t(397) = -6.35$, $p < .01$). However, there was no significant difference between the ESCT and Gemini groups ($t(374) = -1.81$, $p > .05$). ChatGPT had a higher proportion of trees with a height of 1 (15.6%) compared to other passages. These sentences correspond to the section titles. Lastly, it can be noted that Gemini passages do not contain any sentences with a dependency tree height of 4. Overall, ESCT contained the tallest trees. The average maximum distance between dependents and heads in a sentence was slightly shorter for ESCT (8.0 tokens) than for generative AIs (ChatGPT: 8.3 tokens; Gemini: 8.4 tokens; DeepSeek: 9.7 tokens). Among generative AIs, DeepSeek had the longest sentences; therefore, the maximum distance between dependents and heads was longer. However, the ESCT had taller dependency trees. The proportion of trees with a height of 3 is much higher for ESCT (22.4%) than for all generative AIs (ChatGPT, 8.5%; Gemini, 12.4%; DeepSeek, 13.1%). A taller dependency tree results from a more complex sentence structure. In summary, the ESCT has more complex but shorter sentences. DeepSeek had the longest average maximum distance, which correlated with its having the longest average sentence length. The standard deviation was approximately the same for all passage types (ESCT: 4.44; ChatGPT: 4.38; Gemini: 4.44), with only DeepSeek being slightly lower (4.22).

This implies that ESCT aims to distinguish whether test takers can understand complex English sentences by measuring their grammatical knowledge. Hence, EFL teachers should rewrite AI-generated text so that sentences are shorter but dependencies are deeper.

Claim (c) draws on passage-level metrics investigating the proportion of pronouns (Table 3). ESCT passages were found to have a higher proportion of pronouns, particularly in personal pronouns. This accounted for 1.38 % of the tokens, whereas personal pronouns accounted for less than 1% of AI-generated passages (ChatGPT: 0.39%; Gemini: 0.71%; DeepSeek: 0.45%). Furthermore, very few wh-pronouns and existential *there* were found in the AI-generated passages. The proportion of determiners was also higher in ESCT passages (0.52%) than in AI-generated passages (ChatGPT, 0.28%; Gemini, 0.22%; DeepSeek, 0.22%). In summary, it is important to increase the proportion of personal pronouns and determiners when adjusting AI-generated passages.

Claim (d) refers to passage-level metrics that investigate the proportion of each dependency in passages (see Figure 7). In ESCT passages, the prepositions and objects of a preposition are the most common dependencies. The proportions of prepositions and objects of a preposition were similar across all types of passages (14.1% and 14% for ESCT, 13.1% and 12.8% for ChatGPT, 12.1% and 11.4% for Gemini, 13.1% and 12.6% for DeepSeek). For every generative AI, the most common dependency was adjectival modifier (13.6% for ChatGPT, 15.3% for Gemini, 14.6% for DeepSeek), even though it was only the 5th most common dependency in ESCT passages. A significant increase in the proportion of conjuncts and coordination can be found for generative AIs (8.6% and 6.7% for ChatGPT, 8.8% and 6.6% for Gemini, 8.6% and 6.3% for DeepSeek) in comparison to ESCT (3.9% and 3.2%). Conversely, a slight decrease was found for determiners and nominal subjects (6.9% and 6.0% for ChatGPT, 7.4% and 6.7% for Gemini, 7.7% and 5.9% for DeepSeek) in comparison to ESCT (10.7% and 7.9%). Furthermore, almost no numeric modifiers were found in AI-generated passages (between 0.0% and 0.1%), even though they represented 2.5% of ESCT passages.

Accordingly, EFL teachers should reduce the proportion of adjective modifiers, conjuncts, and coordination in AI-generated passages.

It must be noted that recommendations (a) to (d) are not completely independent from each other. For example, sentences could be shortened by reducing the number of adjectival modifiers. Increasing personal pronouns not only shortens sentences but also contributes to increasing the proportion of A1 tokens. Below is an example of such a revision.

Original text generated by DeepSeek;

Rainforests are indispensable to human survival due to their multifaceted roles in maintaining ecological balance, supporting biodiversity, regulating global climate systems, and providing essential resources.

Revised version to be more similar to ESCT;

Rainforests are indispensable to human survival due to their various(a) roles. For example, they(c) help maintaining(b) ecological balance(b, d). They(c) also support biodiversity and regulate global climate systems. Additionally, they(c) provide ~~essential~~(d) resources that are essential to us(b, c).

## 6. Conclusion

By analyzing vocabulary coverage and syntactic complexity, this paper answered the RQ "What are the linguistic characteristics that must be accounted for when modifying AI-generated passages to create mock ESCT?" It was found that (a) adding a wider variety of A1 level lemmas**,** (b) rewriting more complex and shorter sentences**,** (c) increasing the number of personal pronouns and determiners**,** (d) reducing the number of adjective modifiers, conjuncts, and coordination can move AI-generated passages closer to ESCT passages.

In actual classrooms, EFL teachers are required to adapt materials while being aware of the reading skills that they want students to acquire. By using generative AI, the requirement for non-native EFL teachers to write materials themselves or ask native English speakers to proofread is unnecessary. Hence, they can reduce the time spent searching and selecting teaching materials and newspaper articles. Therefore, the results of this study are expected to lead to a reduction in teachers' workload. Furthermore, the time saved leads to improved test quality. Thus, the findings offer EFL teachers a rational way to create mock

ESCT reading passages. MEXT should aim to reduce teachers' workloads; however, implementing new policies requires time and effort. Therefore, EFL teachers can refer to recommendations (a)−(d) to create mock tests using AI-generated passages and partly reduce their workload.

By analyzing AI-generated passages, this study also reveals some of their linguistic features, along with some educational implications. Many previous studies on AI in education have focused on chatbot dialogues (Shishido, 2023), VR tools (Jiang, 2022) and how AI is used to facilitate classroom activities (Ho, 2023). The development of AI-based content generation tools is expected to increase (Qazi et al., 2024). However, the linguistic features of AI-generated passages remain under-researched. Through the present analysis, it became clear that even if they were developed by different companies, all the AI-generative tools generated passages with similar features. For example, the CEFR-J levels of tokens and lemmas were similar across all generated AIs, whereas the ESCT passages had a significantly different makeup. In addition, a significant increase in the proportion of conjuncts and coordination was found for generative AIs (8.6% and 6.7% for ChatGPT, 8.8% and 6.6% for Gemini, 8.6% and 6.3% for DeepSeek) compared with ESCT (3.9% and 3.2%). Moreover, the proportion of each dependency was different between ESCT and generative AIs; however, almost identical proportions were observed across generative AIs. There were only slight differences among AI-generative passages, such as datives and object predicates that were not found in Gemini, negation modifiers not found in DeepSeek, and predeterminers not found in either ChatGPT or Gemini. AI generative tools continue to evolve and improve (Qazi et al., 2024). Therefore, the above features may not remain relevant in the future. However, the findings can help guide EFL teachers when implementing generative AI in the classroom. They will also be useful as a snapshot of the characteristics of generative AI by early 2025.

Additionally, this study contributes to the development of new methods for analyzing reading passages. Previously, vocabulary coverage was widely used in the study of the Center Test (Chujo & Hasegawa, 2004; Tani, 2008; Sakurai, 2022). However, few studies have focused on syntactic features. Following Salingre and Kurokawa (2023), the present study showed that NLP tools can help elucidate the syntactic complexity of a reading passage and proposed the first analysis of dependencies in ESCT. By analyzing both syntactic complexity and vocabulary coverage, it is possible to understand the Center Test and AI-generated passages in greater depth than when relying only on vocabulary coverage.

Nonetheless, further research is required to develop a complete recipe for creating mock English reading tests using AI-generated passages. First, among ESCT reading passages, the present analysis focused solely on long academic reading passages. It is necessary to analyze several types of reading passages to obtain a complete picture of the differences between ESCT- and AI-generated passages. Second, the Center Test targeted Japanese high school students; other English tests, such as the TOEFL or IELTS may have different characteristics. Therefore, a comparison of other high-validity and high-reliability tests with AI-generated passages is recommended. Finally, the fact-checking problem remained unsolved. Even if the goal of reading comprehension is to test students' reading skills from an educational perspective, presenting students with misinformation or false information must be avoided. By discerning and countering misinformation, Saeidnia *et al.* (2025) emphasized the integration of human oversight and continual algorithm refinement emerges as pivotal in augmenting AI's effectiveness. Thus, even if the linguistic features of AI-generated passages can be closer to human-made tests, teachers must check the accuracy of the contents; for example, by using text or tools such as Google's Fact Check Explore (Hartley, 2024) which aimed to facilitate the work of fact checkers, journalists, and researchers.

In conclusion, this paper developed a methodology (a-d) for creating mock English reading tests using AI-generated passages and suggested the development of new methods for analyzing reading passages using NLP. In future studies, the methods presented in this paper can be applied to compare AI-generated text with other English tests, such as the TOEFL, TOEIC, and so on.

## References


Chujo, K., & Hasegawa, S. (2004). Goi no kabā ritsu to rīdabiritī kara mita daigaku eigo nyūshi mondai no nan'ido [The difficulty of English university entrance examination problems from the perspective of vocabulary coverage and readability]. *Bulletin of College of Industrial Technology, Nihon University, B, 37*, 45-55.

Erikawa, H. (2011). *Juken Eigo to Nihonjin: Nyūshi Mondai to Sankōsho kara Miru Eigo Gakushūshi* [*Entrance exam English and the Japanese: English learning seen through entrance exam questions and reference books*]. Kenkyusha.

Hartley, R. (2024). Efficacy analysis of online Artificial Intelligence fact-checking tools. *The International Review of Information Ethics*, *33*(1). https://doi.org/10.29173/irie502

Ho, C. C. (2023). ChatGPT as a tool for developing paraphrasing skills among ESL learners. *Journal of Creative Practices in Language Learning and Teaching (CPLT), 11*(2), 85-105. https://doi.org/10.24191/cplt.v11i2.21723

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural Language Understanding with Bloom Embeddings, Convolutional Neural Networks and Incremental Parsing.*

Hsueh-Chao, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language, 13*(1), 403-430.

Jiang, R. (2022). How does artificial intelligence empower EFL teaching and learning nowadays? A review on artificial intelligence in the EFL context. *Frontiers in Psychology*, *13*. https://doi.org/10.3389/fpsyg.2022.1049401

Kuramoto, N. (2017). Daigaku nyūshi seido kaikaku no ronri ni kansuru ichi kōsatsu – Daigaku nyūshi sentā shiken wa naze haishi no kiki ni itatta no ka [An inquiry about the logic of the reform of university entrance examinations: why was the Center Test close to being abandonned?]. *Daigaku Nyūshi Kenkyū Jānaru* [*Journal of Entrance Examination Research*], *27*, 29-35. https://doi.org/10.57513/dncjournal.27.0_29

MEXT. (2019, January 25). *Kōritsu Gakkō no Kyōshi no Kinmu Jikan no Jōgen ni Kansuru G aidorain* [*Guideline on the work time of teachers in public schools*]. Ministry of Educa tion, Culture, Sports, Science and Technology of Japan. https://www.mext.go.jp/comp onent/a_menu/education/detail/__icsFiles/afieldfile/2019/01/25/1413004_1.pdf

MEXT. (2020, October 28). *Wagakuni no Nyūshi Seido no Gaiyō* [*Outline of our country's un iversity entrance exam system*]. Ministry of Education, Culture, Sports, Science and Te chnology of Japan. https://www.mext.go.jp/content/20201028-mxt_daigakuc02-00001 0703_12.pdf

MEXT. (2023, April 28). *Kyōin Kinmu Jittai Chōsa (Reiwa 4 nendo) Shūkei – Kinmu Jikan n o Jikeiretsu Henka –* [*Results of the survey on teachers' workload (2022) – changes in workload with time*]. Ministry of Education, Culture, Sports, Science and Technology of Japan. https://www.mext.go.jp/content/20230428-mxt_zaimu01-000029160_1.pdf

Qazi, S., Kadri, M. B., Naveed, M., Khawaja, B. A., Khan, S. Z., Alam, M. M., & Su'ud, M. M. (2024). AI-driven learning management systems: Modern developments, challenges and future trends during the age of ChatGPT. *Computers, Materials & Continua*, *80*(2), 3289-3314. https://doi.org/10.32604/cmc.2024.048893

Saeidnia, H. R., Hosseini, E., Lund, B., Tehrani, M. A., Zaker, S., & Molaei, S. (2025). Artificial intelligence in the battle against disinformation and misinformation: a systematic review of challenges and approaches. *Knowledge and Information Systems*, 1-20. https://doi.org/10.1007/s10115-024-02337-7

Sakurai, T. (2022). Gakushū shidō yōryō kaitei ni tomonau eigo kyōiku no henkaku – kōpasu wo katsuyō shita korekara no jidai no kyōzai kenkyū [Transformation of English education with the revision of the Course of Study: research on teaching materials for the future using corpora]. *Studies in Arts & Letters: Literature, History, Geography, 56*, 1-15.

Salingre, M., & Kurokawa, S. (2023). Igakubu eigo nyūshi no goi oyobi tōgoteki bunseki – kakariuke kaiseki to bunsan hyōgen wo chūshin ni [Lexical and syntactic analysis of medical English university entrance exams: focus on dependency tree analysis and word embeddings]. *KLA Journal, 7*, 16-32.

Shirahata, T. (2019). Applying research findings from theoretical linguistics to teaching of English as a foreign language: a case of teaching personal pronouns. *The Chubu English Language Education Society, 48*, 243-248.

Szmrecsányi, B. (2004). On operationalizing syntactic complexity. In G. Purnelle, C. Fairon & A. Dister (eds.). *Le Poids des Mots - Proceedings of the 7th International Conference on Textual Data Statistical Analysis,* 1032-1039.

Tani, K. (2008). Daigaku nyūshi sentā shiken goi to kōkō eigo kyōkasho no goi hikaku bunseki: kabā ristu no kanten kara [Comparison between the vocabulary of the Center Test and highschool English textbooks: from the perspective of vocabulary coverage]. *Practical English studies, 14*, 47-55.

Tono, Y. (2017). The CEFR-J and its Impact on English Language Teaching in Japan. *JACET International Convention Selected Papers, 4*, 31-52.

Tono, Y. (2019). Coming Full Circle – From CEFR to CEFR-J and back. *CEFR Journal – Research and Practice, 1*, 5-17.

Tono, Y. (2020). *CEFR-J Wordlist ver 1.6*. https://www.cefr-j.org/download.html

Uchida, S., & Negishi, M. (2021). Eigo dokkai kyōzai no CEFR reberu no suitei: CVLA no datōsei hyōka [Estimation of the CEFR level of English reading materials: evaluation of the validity of CVLA]. *Journal of Corpus-based Lexicology Studies, 3*, 1-14.

von Glasersfeld, E. (1970). The problem of syntactic complexity in reading and readability. *Journal of Reading Behavior, 3*(2), 1-14.

Wang, T., Lund, B. D., Marengo, A., Pagano, A., Mannuru, N. R., Teel, Z. A., & Pange, J. (2023). Exploring the potential impact of artificial intelligence (AI) on international students in higher education: Generative AI, chatbots, analytics, and international student success. *Applied Sciences*, *13*(11), 6716. https://doi.org/10.3390/app13116716

Watanabe, Y. (2013). The national center test for university admissions. *Language Testing*, *30*(4), 565-573. https://doi.org/10.1177/02655322134830

Young, J. C., & Shishido, M. (2023). Investigating OpenAI's ChatGPT potentials in generating Chatbot's dialogue for English as a foreign language learning. *International journal of advanced computer science and applications, 14*(6). https://doi.org/10.14569/IJACSA.2023.0140607