Feasibility of NLP Metrics in Automatic Paraphrase Evaluation for EFL Learners

Minkyung Kim[a]
Department of English Language and Literature, Seoul National University, Gwanak-Ro, Seoul, Korea

Keywords: Writing Assessment, Automatic Essay Scoring, Paraphrase Evaluation, Evaluation Metrics, Natural Language Processing.

Abstract: This study investigates the feasibility of evaluation metrics for English as Foreign Language (EFL) learners' paraphrases. Paraphrasing can effectively measure learners' writing skills, yet much attention has not been oriented to automated systems in this area. While considerable efforts have been made to reduce burdens for teachers by developing automatic essay scoring system, there is little research on bridging the automatic assessment and paraphrasing in terms of language testing. Thus, this study explores three evaluation metrics in natural language processing (NLP) – dependency distance, cosine similarity, and Jaccard distance – mainly designed for machine translation to assess syntactic and word change as well as semantic congruence. A total of 1,000 paraphrases from Korean EFL undergraduate and graduate students were evaluated via target metrics with the results compared to human rating. Pearson correlation coefficient turned out to be moderate and high in semantic equivalency and lexical diversity, but in syntactic change, there were few significant correlations. Finding appropriate alternative metrics for syntactic complexity and developing automatic evaluation could be crucial steps for future research.

1 INTRODUCTION

In academic writing, to build rationale, it is inevitable to use other scholars' works on our own. In this process, directly restating some sentences in others work can be a serious problem, which is plagiarism. It is necessary for students to indirectly cite the sources to support theirs by changing sentence structure and altering some words in the original works. This is called paraphrasing – syntactically and lexically different but semantically congruent.

Paraphrasing is a writing strategy that helps writers cite indirectly, avoid plagiarism and make their materials more reliable when especially for academic writing (Chen, Huang, Chang, & Liou, 2015; Hawes, 2003). However, as Eberle (2013) emphasizes, improper paraphrasing—where ideas are rephrased without learning proper methods or citation—can still lead to plagiarism.

Moreover, the increasing reliance of learners—particularly EFL (English as a Foreign Language) students—on large language models (LLMs) and other online tools for writing assignments exacerbates the issue of improper paraphrasing (Pecorari, 2023). While these tools can provide support, they often encourage unconscious copy-and-paste behaviors, raising concerns about the academic honesty (Warschauer et al., 2023).

As such, the need for effective teaching and scalable assessment of paraphrasing skills is more pressing than ever, especially given the growing influence of AI-based writing tools. This underscores the need for students to develop appropriate paraphrasing skills through targeted instruction and practice.

Testing whether students have learned to paraphrase properly is as important as teaching the skill itself. However, assessing paraphrasing ability presents significant challenges for educators, as it is resource-intensive and often relies on manual scoring (Page, 1966). This process can be particularly burdensome when evaluating large numbers of students, requiring considerable time and effort to ensure accurate and fair assessment.

For efficient evaluation process for paraphrasing practices, developing the automatic scoring system

[a] https://orcid.org/0009-0004-4588-2980

Kim, M.
Feasibility of NLP Metrics in Automatic Paraphrase Evaluation for EFL Learners.
DOI: 10.5220/0013297400003932

In *Proceedings of the 17th International Conference on Computer Supported Education (CSEDU 2025) - Volume 2*, pages 760-767
ISBN: 978-989-758-746-7; ISSN: 2184-5026

for paraphrasing is also getting important. Previous research on paraphrase quality evaluation has primarily focused on machine learning and NLP, including building models for paraphrase generation (Andrews & Witteveen, 2019; Hegde & Patil, 2020) and developing paraphrase corpora (Jayawardena & Yapa, 2024; Pehlivanoğlu et al., 2024).

However, there are limited studies examining the suitable metrics to form an automated process of paraphrase assessment with learners' data. For instance, Han et al. (2022) explored NLP-based metrics to evaluate the quality of paraphrasing and translation produced by L2 learners (people learning a second language, typically distinct from their native language). However, their metrics were designed to consider both paraphrasing and translation skills, which limits their applicability for assessing paraphrasing in isolation. Consequently, identifying appropriate and credible metrics for paraphrasing is a crucial first step toward developing an automated scoring model specifically tailored to this task.

Paraphrasing, as defined in this study, requires maintaining the meaning of the source text while rephrasing its structure and vocabulary. To evaluate a paraphrase effectively, three scoring dimensions are essential: (a) syntactic complexity, (b) lexical complexity, and (c) semantic similarity. In the current study, three evaluation metrics were selected to correspond to these dimensions. The scores generated through these metrics were then compared with ratings provided by human experts using an analytic scoring rubric to conduct a reliability analysis.

This study contributes to the ongoing investigation of reliable metrics for paraphrase evaluation, laying a foundation for establishing a quick and efficient process to assess the writing ability of EFL learners, who are non-native English speakers learning English as their second or additional language. By providing convenience for teachers and opportunities for learners to engage in self-training, these metrics bridge the gap between automated evaluation and practical pedagogy. The reliability measured for each metric offers valuable insights for the development of an automated paraphrase scoring system within an objective and systematic evaluation framework.

Furthermore, incorporating NLP metrics into educational tools holds significant potential to address key challenges in teaching and evaluating paraphrasing skills. For educators, these metrics can automate the evaluation of paraphrase quality, significantly reducing the burden of large-scale assessments while ensuring consistency in feedback. For learners, NLP-based systems can provide instant, actionable feedback, enabling them to refine their paraphrasing techniques and avoid errors like semantic distortion or copying.

Ultimately, this study highlights the potential of NLP-based metrics as a valuable tool for alleviating the challenges faced by educators relying on traditional evaluation methods. These metrics not only streamline the assessment process but also enhance the quality of paraphrasing instruction, promote academic integrity, and support the development of critical writing skills in EFL contexts.

2 RELATED WORK

Recent research with the topic of paraphrase usually falls into two research fields of natural language processing (NLP) and language assessment. The current study bridges these two areas together to explore suitable metrics of paraphrase assessment and examine the feasibility of them.

2.1 Paraphrase Assessment in NLP

Paraphrase quality assessment is a significant area of focus within natural language processing, given the importance of paraphrasing in large language models (LLMs) and summarization tasks. As one of the fundamental aspects of NLP, evaluating the quality of paraphrases has gained considerable attention, relying on a range of metrics commonly employed in computational linguistics. Notably, many NLP metrics originally developed for machine translation have been adapted to automatically assess the quality of paraphrases generated by LLMs.

Investigating LLM-generated paraphrases with NLP metrics has started with using relative early models such as OpenAI's GPT-2 and Google AI's T5 (Andrews & Witteveen, 2019; Hegde & Patil, 2020; Palivela, 2021). More recent studies have constructed paraphrase data from state-of-the-art (SOTA) models such as GPT-3.5 turbo, GPT-4 as well as ChatGPT (Jayawardena & Yapa, 2024; Kim & Kim, 2024; Pehlivanoğlu et al., 2024).

For example, Andrews & Witteveen (2019) assessed the quality of paraphrases produced from OpenAI's GPT-2 with metrics such as Universal Sentence Encoder (USE), ROUGE-L, and BLEU. In the similar vein, Jayawardena & Yapa (2024) tested their paraphrase corpus developed via GPT-3.5 turbo model with a total of 19 NLP metrics.

However, the metrics used in previous studies to evaluate rating criteria of syntactic diversity, semantic similarity, and lexical diversity were mainly designed to investigate the quality of machine translation (MT), not paraphrase (Lee et al., 2023).

Capturing this research gap, Kim & Kim (2024) conducted a comparative analysis of ChatGPT- and human-generated paraphrases with other NLP metrics which were considered suitable for assessing paraphrase. They selected dependency distance for assessing syntactic diversity, cosine similarity for semantic similarity and Jaccard distance for lexical diversity. Their preliminary results indicated that ChatGPT has superior ability in syntactic change and semantic retention but only comparable skill in lexical diversity.

However, the reliability of three metrics used in Kim & Kim (2024) was not investigated by comparing with the scores rated through human evaluation. It is necessary to examine the potential of dependency distance, cosine similarity and Jaccard distance in paraphrase assessment.

The metrics employed in prior studies to assess paraphrase quality in LLMs also hold significant potential for application in EFL writing assessment research, particularly in evaluating learner-generated paraphrases. Leveraging these established metrics can help develop automatic scoring models. Such an approach would reduce the burden of manual scoring for educators. Identifying suitable metrics for this purpose represents a critical step toward enhancing the efficiency and reliability of evaluating learner-generated paraphrases in educational contexts.

2.2 EFL Paraphrase Evaluation

Paraphrasing ability serves as a crucial writing strategy especially in academic writing and has been regarded as one of efficient language learning methods in second language acquisition (SLA) (Han et al., 2022). Several studies focused on developing paraphrasing test and evaluation system (Chen et al., 2015; Kim, 2018; Ji, 2012) but they used traditional methods in the field of language assessment, indicating that there has been little research in the aim of automatic rating for paraphrases as well as using NLP metrics.

Han et al. (2022) introduced NLP evaluation framework for assessing translation and paraphrasing from L2 learners to develop standardized and objective criteria. However, the metrics used in the study were chosen for rating both of translation and paraphrasing. There were nine metrics for investigating syntactic and lexical complexity of L2 translation and paraphrases but they were not mainly chosen with the priority of paraphrase in mind. They measured semantic similarity for paraphrasing tasks and coherence of content in translation tasks with the same metric.

However, putting two different scoring components together and measuring them simultaneously can cause an overlap for other studies with priority in paraphrasing. Thus, considering the special features of paraphrase and applying suitable metrics for evaluation process is a necessary step.

With the rationale in mind, the research questions of the current study are as follows:

1) Which NLP metrics are suitable for assessing EFL paraphrases?
2) What is the inter-reliability between the scores rated via NLP metrics and those by human raters?

3 METHODOLOGY

The following illuminates some details of paraphrase material, scores, target metrics, and statistical analysis of current study which aims to figure out the feasibility of suitable metrics for paraphrase assessment. As background information for readers, the paraphrase data analyzed in this study were collected four years ago in a larger project investigating the reliability and validity of a paraphrasing test for Korean EFL learners (Kim, 2020). The present analysis took advantage of that data with new methods to identify effective metrics for assessing paraphrasing.

3.1 Paraphrasing Data

There were 1,000 paraphrases from 100 Korean EFL learners who were undergraduate and graduate students in South Korea. They took five paraphrasing tasks consisting two original English sentences each as shown in Table 1. The original source texts were adapted from Chen et al. (2015) and went under modification. Participants paraphrased English source sentences into English to evaluate their ability to rephrase while retaining the meaning of the original text.

3.2 Paraphrasing Scores

For the paraphrasing data written by Korean EFL learners, there were two human raters recruited to evaluate them following the analytic and holistic scoring rubrics of paraphrasing in Kim (2020). They were graduate students majoring in applied linguistics and have expertise in the field of SLA.

The analytic rubric they used is composed of four scoring dimensions such as syntactic change, word change, semantic equivalency and mechanical errors. On the other hand, the holistic scoring rubric encompasses all the rating criteria in one explanation not with details. The inter-rater reliability between them was on average of .86 in Pearson Correlation Coefficient across all paraphrases which indicates a *very strong* value of correlation.

For the purposes of the current study, the analytic scores were chosen as the target scores to align with the NLP metrics being examined. Specifically, the dimensions of syntactic change, word change, and semantic equivalency were analyzed, excluding mechanical errors to maintain focus on the core features of paraphrasing. These three dimensions were used in the inter-rater reliability analysis in conjunction with the selected NLP metrics.

Table 1: Original Sentences.

No.	Source Text
Task 1	On the whole, fuel prices have risen in recent years. Similarly, the cost of food has increased quite considerably.
Task 2	The film drew attention to the fact that due to circumstances beyond our control, CO2 emissions could be more than double by 2020. As a result, the problem will lead to worse natural disasters.
Task 3	Laughter protects us from the damaging effects of stress. In other words, laughter is the best medicine.
Task 4	Due to the weeklong rainfall beyond our control, the flood destroyed many houses. As a result, many people became homeless.
Task 5	There is no doubt that burning PVC plastic can severely damage human health. Therefore, some European countries have taken a lead in reducing the use of plastics.

3.3 Target Evaluation Metrics

This study builds upon a future research direction proposed by Kim and Kim (2024), specifically focusing on an in-depth examination of the reliability of the three metrics employed in their work as shown in Table 2.

These metrics served as the foundational standards for rating paraphrasing dimensions of syntactic and lexical complexity as well as semantic equivalency providing a basis for further investigation in the present research.

Table 2: Evaluation Metrics for Paraphrasing.

Scoring Dimension	Target Metric	Python Library
Syntactic Complexity	Syntactic Dependency Distance	spaCy
Lexical Complexity	Jaccard Distance	NLTK
Semantic Equivalency	Cosine Similarity	sklearn

To introduce each feature with details, Dependency Distance (DD) is the linear distance between two linguistic units in terms of syntactic structure (Hudson, 1995). It is a frequently used method in NLP which shows direct relations between words. Previous research argued it can measure syntactic complexity (Jiang & Liu, 2015; Liu, 2008).

DD in the current study is used to measure syntactic complexity of a paraphrased sentence compared to its original source text. This approach involves counting ***M***, the number of matching dependencies, and ***T***, the total number of dependencies across both sentences. A dependency match occurs when a syntactic relationship between a head and a dependent word in the original sentence is identically preserved in the paraphrase. Here, ***T*** is the sum of dependencies in the original and paraphrased sentences. The formula for syntactic similarity is as follows:

$$S = \frac{2 \cdot M}{T} \qquad (1)$$

To get the value of syntactic diversity D, 1-S is required. The formula for syntactic diversity is as follows:

$$D = 1 - S \qquad (2)$$

Another metric for reliability analysis is cosine similarity (Salton et al., 1975) which evaluates semantic similarity between the sentences in a paraphrase pair. There have been other previous works adopted cosine similarity to calculate the semantic similarity in paraphrases (Awajan & Alian, 2020; Salman et al., 2023; Vrbanec & Mestrovic, 2021). The equation for cosine similarity is as follows:

$$\text{Cosine Similarity } (A, B) = \frac{A \cdot B}{||A||||B||} \qquad (3)$$

To evaluate the last component of paraphrase, lexical complexity, Jaccard distance (Besta et al., 2020), which is a modified form of Jaccard similarity (Jaccard, 1980), was chosen in this study. Jaccard similarity, denoted as *J(A, B)* is the ration of the intersection to the union of the two paraphrases as in (3).

$$J(A,\ B) = \frac{|A \cap B|}{|A \cup B|} \tag{3}$$

To get the value which shows the degree of difference, Jaccard Distance, denoted as d_J, is calculated through the formula below in (4). This gives an insight into the lexical diversity of a paraphrase; a higher distance value indicates, the greater the word complexity. A richer vocabulary can be an index related to EFL learners' English proficiency.

$$d_J = 1 - \mathrm{J}(A,B) \tag{4}$$

3.4 Statistical Analysis

To investigate the feasibility of target NLP metrics as suitable ones for paraphrasing evaluation, inter-rater reliability analysis was conducted compared to the scores obtained by human rating process.

The measurement for inter-rater reliability in this study was Pearson Correlation Coefficient (PCC). It is a typical measure in automatic evaluation of English language skills (Attali, 2013; Hong & Nam, 2021; Pack et al., 2024). The PCC values range from -1 to + 1, indicating +1 shows a perfect agreement, 0 meaning no relationship, and -1 for a perfect disagreement.

Table 3 shows interpretation of PCC with a figure above +0.8 as a very strong correlation, +0.6 for a strong correlation, and +0.4 for a moderate inter-rater reliability (Wahyuni & Purwanto, 2020).

All correlation-related analyses and statistical testing were conducted via SPSS 28. The automatic calculation using NLP metrics was executed via Python ver. 3.1.1 (Python Software Foundation, 2023).

Table 3: Interpreting Pearson Correlation Coefficient.

Range	Interpretation
0.80 – 1.0	Very Strong
0.60 – 0.79	Strong
0.40 – 0.59	Moderate
0.20 – 0.39	Weak
0 – 0.19	Very Weak

4 RESULT

The current study shed light on figuring out appropriate natural language processing (NLP) metrics in terms of paraphrase assessment. A paraphrase should be syntactically and lexically different from the original sentence while remaining the semantic congruence.

Keeping this in mind, there were three target metrics thoroughly chosen to investigate the inter reliability with human raters in seeking a useful and convenient way for teachers and students to assess their paraphrase in a more objective view. The human scores used in the study were rated by two human raters and calculated as the average figure of those two.

4.1 Reliability in Syntactic Complexity

To calculate the syntactic complexity of Korean EFL paraphrases, syntactic dependency distance has been measured by comparing the original source text and the paraphrase respectively. Table 4 shows the inter-rater reliability through Pearson Correlation Coefficient. The inter-rater reliability between the paraphrasing scores assessed dependency distance (DD) and those by human raters has overall weak correlation around the five tasks reaching the highest correlation figure of .28. This suggests that dependency distance, while providing an objective measure of syntactic complexity, may not align closely with human evaluators' perceptions of syntactic change in paraphrases. This may oppose to the previous research of Liu (2008) and Jiang & Liu (2015) arguing that DD has its potential in measuring syntactic complexity.

The weak correlation could imply that human raters consider additional syntactic elements or nuanced linguistic features that are not captured by the dependency distance metric. These findings highlight the limitations of dependency distance as a potential metric and suggest that its use in paraphrase assessment might require supplementation with other measures or adjustments to better reflect human judgment.

Table 4: Inter-rater Reliability in SC.

Task 1	Task 2	Task 3	Task 4	Task 5
0.28**	0.07	0.20*	0.13	0.27

Note: **p* < .05, ***p* <.0.1

4.2 Reliability in Lexical Complexity

In terms of lexical complexity (LC), the scores obtained from Jaccard Distance (JD) have shown a moderate and strong correlation with the human-rated scores as shown in Table 5. The highest PCC value was discovered in Task 4 and the lowest in 0.57. All values were significant at .01 level.

These results suggest that Jaccard Distance is a relatively reliable metric for capturing lexical complexity in paraphrases, as it aligns well with human judgments in most cases. However, the variation in correlation across tasks implies that task-specific factors, such as the complexity of source texts or paraphrasing strategies employed by learners, may influence the metric's performance, warranting further investigation.

Table 5: Inter-rater Reliability in LC.

Task 1	Task 2	Task 3	Task 4	Task 5
0.62	0.69	0.74	0.77	0.57

Note: All significant at .01 level (2-tailed)

4.3 Reliability in Semantic Equivalency

Scores in the rating dimension of semantic retention were measured with cosine similarity (CS) and compared to those from human raters. Table 6 shows the inter-rater reliability values between human-NLP metric correlations. All scores showed a positive correlation and reached moderate correlation compared to human scores. The highest inter-rater reliability value was observed in Task 1 while the lowest was in Task 3. The results were statistically significant in general.

The results were statistically significant across tasks, highlighting the potential of cosine similarity as an effective metric for assessing semantic retention. However, the task-specific variations in correlation suggest that the nature of the source texts or the strategies employed by learners may influence the alignment between CS scores and human ratings, necessitating further exploration to refine its application.

Table 6: Inter-rater Reliability in SE.

Task 1	Task 2	Task 3	Task 4	Task 5
0.67	0.61	0.44	0.57	0.54

Note: All significant at .01 level (2-tailed)

5 DISCUSSION

The current study shed light on figuring out the suitable metrics for automated evaluation in paraphrasing assessment. A total of 1,000 paraphrases from 100 Korean EFL learners were under evaluation. There were five paraphrasing tasks and ten original sentences consisting two sentences assigned to one task. Two human raters who have professional knowledge in second language acquisition (SLA) were in charge of the rating process. The inter-rater reliability between two human raters was .866 (p < .01). The study adopted metrics from Kim and Kim (2024) to assess key components of paraphrasing: Syntactic Complexity (SC) using Dependency Distance (DD), Lexical Complexity (LC) using Jaccard Distance (JD), and Semantic Equivalency (SE) using Cosine Similarity (CS).

5.1 Low Inter-Rater Reliability in SC

The results showed that among three target metrics used in the study Cosine Similarity (CS) and Jaccard Distance (JD) revealed moderate to high inter-rater reliability with statistical significance. In contrast, dependency distance (DD), which was planned to measure the syntactic complexity (SC), showed weak inter-rater reliability compared to the other metrics.

The reason for this relatively low inter-rater reliability can imply that DD does not measure the syntactic complexity of the paraphrases. One possible explanation might be due to the differences in the sentence length between test-takers. As Jiang and Liu (2015) pointed out, DD is sensitive to sentence length, which means that longer sentences naturally result in higher dependency distances. Therefore, variations in sentence length across test-takers may have skewed the DD scores, leading to weak correlation with human ratings.

Moreover, DD's reliance on dependency trees may not fully account for the types of syntactic variation that human raters would typically notice, such as coordination, subordination, or restructuring, which contribute to the perceived complexity of a sentence. To address this limitation, future studies could consider combining DD with other syntactic metrics, such as phrase or clause complexity, to provide a more comprehensive measure of syntactic complexity that better aligns with human judgment.

5.2 Reliable Metrics for Automatic Paraphrasing Evaluation

The results demonstrated that among three target metrics, Jaccard Distance (JD) and Cosine Similarity (CS) reached moderate and high inter-rater reliability for human raters as denoted by Pearson Correlation Coefficient (PCC). The average PCC for lexical complexity using JD was .68 (p<.01) while semantic equivalency assessed with CS reached .57 (p<.01). These correlation figures imply that both metrics employed for rating lexical complexity and semantic equivalency have considerable potential to reflect the evaluative judgment of human raters in paraphrase assessment.

These findings are significant as they provide foundational support for the development of an effective educational tool for both teachers and students. An automated evaluation model for EFL learners' paraphrases, based on the metrics explored in this study, could significantly reduce the burden on educators by streamlining the assessment process, particularly in large-scale essay grading. By providing quick, consistent, and objective evaluations, such a tool would allow teachers to allocate more time to individualized instruction and feedback, ultimately enhancing the learning experience.

6 CONCLUSION AND FUTURE WORK

This study aimed to validate the evaluation metrics in assessing EFL learners' paraphrases. Despite the growing importance of paraphrasing in academic writing and natural language processing (NLP), paraphrase assessment lacks objective and sophisticated scoring rubrics, which could be a research gap in the related field. Moreover, some previous studies attempted to use computational metrics for essay scoring and paraphrasing for L2 learners, but there was limited research conducted on investigating suitable metrics for paraphrasing solely.

With this rationale, the current study examined the feasibility of three target NLP metrics for paraphrasing assessment and revealed the Cosine Similarity and Jaccard Distance can be the basis for automatic paraphrasing scoring for semantic similarity and lexical complexity.

However, in terms of syntactic complexity, future work is necessary for finding out alternative metrics. While DD is an objective and widely used metric, the weak inter-rater reliability observed in this study suggests that it may not fully capture the nuances of syntactic complexity as perceived by human evaluators.

These results in the current study can be employed in building automatic scoring model for paraphrasing for future work. After finding an alternative scoring metric for syntactic complexity for EFL paraphrases, a composite model that integrates all three scoring dimensions of paraphrasing could be developed.

Automatizing the process of paraphrase assessment could reduce the time and effort required for teachers particularly in large classrooms, allowing them to focus on providing individualized feedback and instruction. Furthermore, such tools could offer students immediate, objective, and consistent feedback on their paraphrasing performance, fostering greater autonomy in learning.

REFERENCES

Andrews, M., & Witteveen, S. (2019). Unsupervised Natural Question Answering with a Small Model. *In Proceedings of the Second Workshop on Fact Extraction and VERification* (FEVER) (pp. 34-38). Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-6606

Attali, Y. & Burstein, J. (2004), AUTOMATED ESSAY SCORING WITH E-RATER® V.2.0. *ETS Research Report Series*, 2004: i-21. https://doi.org/10.1002/j.2333-8504.2004.tb01972.x

Awajan, M., & Alian, A. (2020). *Paraphrasing identification techniques in English and Arabic texts*. In *The 11th International Conference on Information and Communication Systems* (pp. 155–160), Irbid, Jordan.

Besta, M., Kanakagiri, R., Mustafa, H., Karasikov, M., Rätsch, G., & Hoefler, T. (2020). Communication-efficient Jaccard similarity for high-performance distributed genome comparisons. In *Proceedings of the 2020 IEEE International Parallel and Distributed Processing Symposium (IPDPS)* (pp. xx-xx). IEEE. https://doi.org/10.1109/IPDPS47924.2020.00118

Chen, M. H., Huang, S.T., Chang, J. S., & Liou, H.C. (2015). Developing a corpus-based paraphrase tool to improve EFL learners' writing skills. *Computer Assisted Language Learning*, 28(1), 22–40. https://doi.org/10.1080/09588221.2013.783873

Dolan, W. B., & Brockett, C. (2005). Automatically constructing a corpus of sentential paraphrases. In Proceedings of the Third International Workshop on Paraphrasing (IWP2005). *Association for Computational Linguistics*. https://aclanthology.org/I05-5002

Eberle, M. E. (2013). Paraphrasing, plagiarism, and misrepresentation in scientific writing. *Transactions of the Kansas Academy of Science (1903-)*, *116*(3/4), 157–167. https://www.jstor.org/stable/42636364

Han, T., Li, D., Ma, X., & Hu, N. (2022). Comparing product quality between translation and paraphrasing: Using NLP-assisted evaluation frameworks. *Frontiers in Psychology, 13*, 1048132. https://doi.org/10.3389/fpsyg.2022.1048132

Hawes, K. (2003). *Mastering academic writing: Write a paraphrase sentence*. Memphis, TN: University of Memphis.

Hegde, C., & Patil, S. (2020). Unsupervised Paraphrase Generation using Pre-trained Language Models. arXiv. https://doi.org/10.48550/arXiv.2006.05477

Hong, Y. & Nam, H. (2021). Evaluating score reliability of automatic English pronunciation assessment system for education. *Studies in Foreign Language Education*, 35(1), 91-104.

Hudson, R. (1995). *Measuring syntactic difficulty*. Draft manuscript.

IBM Corp. (2021). *IBM SPSS Statistics for Windows, Version 28.0*. Armonk, NY: IBM Corp.

Jaccard, P. (1908). Nouvelles recherches sur la distribution florale. *Bulletin de la Société Vaudoise des Sciences Naturelles, 44*, 223–270.

Jayawardena, L., & Yapa, P. (2024). ParaFusion: A large-scale LLM-driven English paraphrase dataset infused with high-quality lexical and syntactic diversity. arXiv. https://doi.org/10.48550/arXiv.2404.12010

Ji, N. (2018). Investigation into Validity of Paraphrasing Task as a Writing Performance Test Item for EFL Learners. *Modern English Education, 19(2),* 20-29.

Jiang, J., & Liu, H. (2015). The effects of sentence length on dependency distance, dependency direction, and the implications – Based on a parallel English–Chinese dependency treebank. *Language Sciences, 50*, 93–104.

Kim, J. (2014). Predicting L2 writing proficiency using linguistic complexity measures: A corpus-based study. *English Teaching*, 69(4), 27-51.

Kim, M., & Kim, J. (2024). Comparing paraphrases by humans and LLMs: An analysis of syntactic complexity, semantic similarity, and lexical diversity through NLP. In *Proceedings of the 2024 Fall Joint Conference of the Korean Generative Grammar Circle and the Korean Society for Language and Information* (pp. 118-129). Korean Generative Grammar Circle & Korean Society for Language and Information.

Kim, M. (2018). Investigating Reliability of Re-modified Scoring Rubrics for EFL Paraphrasing Task. *SNU Working Papers in English Language and Linguistics, Vol.16*, pp. 36-56.

Kim, M. (2020). *Investigating the reliability and validity of scores from a paraphrasing test for Korean EFL learners* (Unpublished master's thesis). Seoul National University.

Lee, S., Lee, J., Moon, H., Park, C., Seo, J., Eo, S., Koo, S., & Lim, H. (2023). A survey on evaluation metrics for machine translation. *Mathematics*, 11(4), 1006.

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

OpenAI. (2023). GPT-4 [Large language model]. https://openai.com/gpt-4

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligenc*e, 6, 100234.

Page, E. B. (1966). The imminence of grading essays by computer. *The Phi Delta Kappan*, 47(5), 238–243. https://www.jstor.org/stable/20371545

Palivela, H. (2021). Optimization of paraphrase generation and identification using language models in natural language processing. *International Journal of Information Management Data Insights, 1(2)*, 100025. https://doi.org/10.1016/j.jjimei.2021.100025

Pecorari, D. (2023). Generative AI: Same same but different?. *Journal of Second Language Writing*, 62, 101067.

Pehlivanoğlu, M. K., Gobosho, R. T., Syakura, M. A., Shanmuganathan, V., & de-la-Fuente-Valentín, L. (2024). Comparative analysis of paraphrasing performance of ChatGPT, GPT-3, and T5 language models using a new ChatGPT generated dataset: ParaGPT. *Expert Systems, 41(11)*, e13699. https://doi.org/10.1111/exsy.13699

Python Software Foundation. (2023). *Python (Version 3.13)* [Computer software]. Retrieved from https://www.python.org

Ryu, J. (2020). Predicting second language writing proficiency in the different genres of writing using computational tools. *Korean Journal of Applied Linguistics*, 36(1), 141-170.

Salman, M., Haller, A., & Méndez, S. J. R. (2023). Syntactic complexity identification, measurement, and reduction through controlled syntactic simplification. *arXiv preprint* arXiv:2304.07774.

Salton, G., Wong, A., & Yang, C. S. (1975). A vector space model for automatic indexing. *Communications of the ACM, 18(11),* 613–620. https://doi.org/10.1145/361219.361220

Vrbanec, T., & Meštrović, A. (2021). *Relevance of similarity measures usage for paraphrase detection*. In *Proceedings of the 13th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2021) - Volume 1: KDIR* (pp. 129–138). SCITEPRESS. https://doi.org/10.5220/0010649800003064

Wahyuni, T. S., & Purwanto, K. K. (2020). Students' conceptual understanding on acid-base titration and its relationship with drawing skills on a titration curve. *Journal of Physics: Conference Series, 1440*(1), 012018. IOP Publishing. https://doi.org/10.1088/1742-6596/1440/1/012018]

Warschauer, M., Pasquier, A., & Grin, F. (2023). The affordances and contradictions of AI-generated text for writers of English as a second or foreign language. *Patterns, 4*(7), Article 100779. https://doi.org/10.1016/j.patter.2023.100779