Multidiszciplináris Tudományok, 15. kötet, 4. sz. (2025) pp. 48–57. ***https://doi.org/10.35925/j.multi.2025.4.5***

EXPLORING LEXICAL VARIATION THROUGH SYNONYM SETS IN HUMAN AND AI-WRITTEN SCIENTIFIC TEXTS

Erika B. Varga
associate professor, Institute of Informatics, University of Miskolc
3515 Miskolc-Egyetemváros, e-mail: erika.b.varga@uni-miskolc.hu

Attila Baksa
associate professor, Institute of Applied Mechanics, University of Miskolc
3515 Miskolc-Egyetemváros, e-mail: attila.baksa@uni-miskolc.hu

Abstract

We propose a synonym set-based framework to detect stylistic and conceptual features that distinguish native scientific writing, non-native texts, and purely AI-generated texts. Using WordNet and POS-aware synonym clustering, we analyzed 12 aligned text pairs across four concept-level metrics: synonym-set coverage, lexical reduction ratio, collapsed type-token ratio, and Jaccard similarity. Native texts consistently exhibited higher conceptual overlap (Jaccard scores between 0.217–0.344 at moderate thresholds) with their AI-generated counterparts than non-native ones. Coverage was slightly richer in native texts (mean difference ≈+0.03), while non-native texts showed more vocabulary redundancy (mean reduction ≈0.03; mean rise in redundancy ≈0.05). These patterns suggest that non-native writings show lower lexical variety and higher redundancy. Our method enables researchers to identify lexical tendencies that help differentiate human-authored and AI-written texts in academic contexts.

Keywords: *AI detection in scholarly texts, lexical variation metrics, synonym clustering, synonym set-based analysis*

1. Introduction

In the context of increasing AI integration in academic writing, our previous study investigated lexical and stylistic patterns in scientific texts with a focus on surface-level similarity metrics such as Jaccard and Cosine overlap (Varga and Baksa, 2025). Building on this foundation, the present research extends that work by introducing a new, semantically richer layer of analysis: the use of synonym set-based lexical profiling.

The motivation for the research stems from repeated personal experiences as reviewers of scientific manuscripts authored by non-native English speakers. In recent years, we have frequently encountered unusually elaborate lexical choices: terms that, while grammatically correct, deviate from the conventional style of technical communication in STEM fields. Words like intricate, nuances, realm, for instance, are now appearing with a frequency and stylistic emphasis that was rarely observed in earlier scientific writing. These patterns raised questions about the evolving norms of academic English and the potential role of AI-powered tools in influencing them.

To explore these phenomena more systematically, the present study focuses on identifying and analyzing synonym-level lexical variation in scientific texts. Rather than examining individual words in isolation, we aggregate semantically related terms into synonym sets, which serve as conceptual units

for analysis. This allows us to assess not just how many words two texts share, but whether they express the same ideas using different lexical choices. In fact, the excessive use of synonyms may be an indicator of AI involvement.

The metrics introduced in this study (Synonym Set Coverage, Lexical Reduction Ratio, and Collapsed Type-Token Ratio) offer insight into how diverse or condensed a text's vocabulary becomes when synonymic variation is normalized. These measures help reveal whether a text prioritizes stylistic richness over informational clarity. In parallel, pairwise Jaccard similarity is adapted to compare synonym group usage between texts, enabling a concept-level evaluation of lexical overlap.

By applying this method to an aligned corpus of human-written, AI-assisted, and fully AI-generated scientific texts, this study aims to refine our understanding of lexical variation and redundancy in scholarly communication. The synonym set-based approach offers a more refined detection of stylistic divergence, potentially supporting future guidelines for responsible AI use in academic publishing.

2. Related work

The widespread use of large language models (LLMs) in academic writing has raised serious concerns about authorship and originality. Many recent studies investigate how to distinguish AI-generated text from human-authored text by examining linguistic features and testing detection tools.

For example, Elkhatat et al., (2023) evaluated five AI content detectors (OpenAI, Writer, Copyleaks, GPTZero, CrossPlag) on paragraphs generated by ChatGPT-3.5 vs. ChatGPT-4 and on human-written controls. They found all tools were more accurate on content from the older ChatGPT-3.5 model than from GPT-4, and notably the detectors often misclassified genuine human-written text as AI-generated. This unreliability (false positives on human text) underscores the need for more robust detection methods as AI text grows more sophisticated. Similarly, Weber-Wulff et al., (2023) tested 14 detection systems and reported that while human-written documents were correctly identified over 80% of the time, AI-generated documents were flagged only about 50 to 88% of the time. They also observed inconsistent biases: some detectors tended to label nearly everything as AI-written or vice versa. These studies highlight that current detection tools are far from reliable, especially given the rapid advancement of LLM capabilities.

Beyond detection software, researchers have developed custom machine learning classifiers to identify AI versus human writing by using stylometric and lexical features. Islam et al., (2023) built a dataset of 10,000 texts (about half human-written from news sources and half ChatGPT-3.5 generated) and compared 11 classification algorithms. An Extremely Randomized Trees model performed best with approximately 77% accuracy distinguishing ChatGPT-3.5 text. Interestingly, they found that removing stop-words from the text actually hurt classification accuracy, suggesting that function words carry stylistic signals that differentiate AI and human writing. This aligns with other stylometry findings that slight differences in word usage can reveal authorship. For instance, Berriche and Larabi-Marie-Sainte, (2024) extracted various fundamental writing-style features (such as vocabulary richness, sentence length, part-of-speech patterns, etc.) to detect ChatGPT-written student essays. Their XGBoost classifier achieved 100% accuracy on a test set, dramatically outperforming simpler TF-IDF baselines. They further showed strong results even when AI and human-written paragraphs were mixed within the same document. These works indicate that classic stylometric approaches, when carefully tuned, can be very effective in identifying AI-generated text by capturing minor linguistic deviations. In fact, LLM-generated text often follows certain stylistic tendencies – e.g. being overly fluent, neutral in tone, and logically structured – that differentiate it from the variability of human writing. Recent linguistic

analyses have quantified such differences. Rosenfeld and Lazebnik, (2023) compared outputs from GPT-3.5, GPT-4, and Google Bard, and found significant divergences in vocabulary distribution, part-of-speech usage, and dependency structures across these models. Using these linguistic markers, they could even attribute a given text to the correct AI model with about 88% accuracy. These results underscore that AI-generated texts in general are linguistically distinct from human texts and even from each other, opening the door for automated detection and source attribution.

Several studies have focused specifically on scientific writing and the challenges of AI involvement in that domain. Hakam et al., (2024) examined the quality and detectability of AI-generated scientific abstracts in orthopedics by rewriting published abstracts using ChatGPT and another AI, then asking experts and software to identify which were AI. Their striking finding was that neither the human researchers nor the AI-detection tool could reliably tell the machine-written abstracts apart from the real ones. This suggests that in a specialized, formal domain like academic abstracts, current detectors and even domain experts may be easily misled by fluent AI rephrasing. Similarly, Taloni et al., (2024) explored ChatGPT's ability to produce convincing scientific text. They found that GPT-4 could paraphrase real published abstracts into new wording with low plagiarism scores (around 10% overlap) while still scoring extremely high on AI-generated text detectors (over 90% "AI-likelihood"). Interestingly, when those GPT-4 paraphrased abstracts were further "humanized" by an AI rephrasing tool (Undetectable.ai), the detector's confidence dropped dramatically (down to 28%). In other words, an AI can rewrite its own output (or another AI's output) to avoid detection, at the cost of introducing small errors.

While most prior work has focused on fully AI-generated content, fewer studies have examined AI-rephrased or AI-translated texts: cases where a human-authored text is rewritten by an AI. Such texts pose a greater detection challenge because they retain much of the original human-like content and structure. Mindner et al. (2023) directly addressed this by constructing a multilingual corpus of original vs. AI-rephrased texts in four languages. They engineered 37 features to train classifiers to detect either fully AI-written passages or human passages that had been rephrased by ChatGPT. The results show a sharp contrast: their classifiers could detect pure AI-generated text with over 96% F1-score, but detecting AI-rephrased text was far harder, with best F1 around 78%. Still, this significantly outperformed off-the-shelf detectors like GPTZero, which barely reached 28% F1 on rephrased texts. This reflects that detecting AI involvement requires identifying fine linguistic fingerprints that survive paraphrasing, or the use of semantic inconsistencies that an AI might introduce even when the text appears fluent.

One promising approach is to analyze changes in word choice and synonym usage introduced by AI rephrasing. Since a common strategy in plagiarism and AI integration is to substitute words with synonyms or alter phrasing while preserving meaning, the examination of synonyms can increase the performance of plagiarism detection tools. By applying synonym databases, researchers can map different words to their underlying synonym sets and detect when an apparently different text is semantically equivalent to another. Prior works in plagiarism detection have utilized WordNet, the lexical database that groups English words representing the same concept into sets of synonyms (called synsets) to catch paraphrased content (Chen et.al, 2010; Thomson, 2017; Álvarez-Carmona, 2019).

In our context, analyzing synonym patterns can reveal AI involvement: an AI-translated or AI-rewritten text might use unusual or overly formal synonyms that a human author would not. To systematically capture this, we generate synonym sets for words in the text using Pyhton's NLTK interface and WordNet. Rather than blindly merging all transitively connected synonyms which can lead

to overly broad groups connecting weakly related terms, we apply a graph-based clustering method that conservatively groups words by strong direct synonymy. This approach avoids the pitfalls of full transitive closure in a synonym graph, which would erroneously equate words that share an intermediate synonym but differ in sense. This way, we preserve more precise semantic signals and can detect higher-level lexical choices that may suggest AI involvement.

3. Research method

3.1. Corpus transformation into synonym sets

For conducting our experiments, we have created a corpus of long academic texts (3-10 thousand words) written in English in Computer Science domain. The corpus is divided into 4 groups. The first group contains papers authored by native speakers from 2015 (before the release of large language models, LLMs). The second group contains texts generated by ChatGPT 4 aligned with the topics in the first group. In the third group there are papers and working drafts of non-native speakers from the era when LLMs are used to help constructing scientific publications. These texts are selected in a way that guarantees moderate use of AI, therefore this group is referred to as translated papers. Finally, the fourth group includes texts generated by ChatGPT 4.o deep research function aligned with the topics in the third group.

To explore lexical variation between human- and machine written texts, we have transformed the text corpus into a corpus of synonym sets. In this process, as illustrated in *Figure 1*, the unique lemmas were collected first for each text together with the frequency of their occurrences and the part-of-speech (POS) tags they appear in. Next, WordNet was used to search for true synonyms of each lemma that also appear inside the same text and that share at least one identical part-of-speech tag. This preprocessing step may result in overlapping synonym sets. To consolidate these sets, we propose an iterative Local Threshold-based Synonym Set Merging (LTSSM) method.

Our LTSSM algorithm starts with a collection of synonym sets, each represented as a dictionary of words with their corresponding frequencies and part-of-speech tags. Each synonym set is compared pairwise with the remaining sets to determine potential merging candidates. The similarity between two sets is defined as the ratio of their shared words to the size of the smaller set. A dynamic similarity threshold is computed for each pair based on the minimum size of the two sets. This threshold is defined as:

$$Threshold = \begin{cases} 1.1, & \text{if } \min(|S_1|, |S_2|) \leq 1 \text{ (to preserve singleton set)} \\ 1 - \dfrac{1}{\min(|S_1|, |S_2|)}, & \text{otherwise} \end{cases} \quad (1)$$

This ensures that larger sets require a higher degree of overlap for merging. On the other hand, singleton sets (containing only one word) are explicitly preserved and excluded from merging.

Two sets are merged only if their similarity ratio meets or exceeds the dynamically computed threshold. After each merge, the algorithm iteratively re-examines the updated sets until no further merges are possible. By enforcing local threshold checks at each merge, this method avoids the unintended chaining effects common in global transitive closure algorithms.

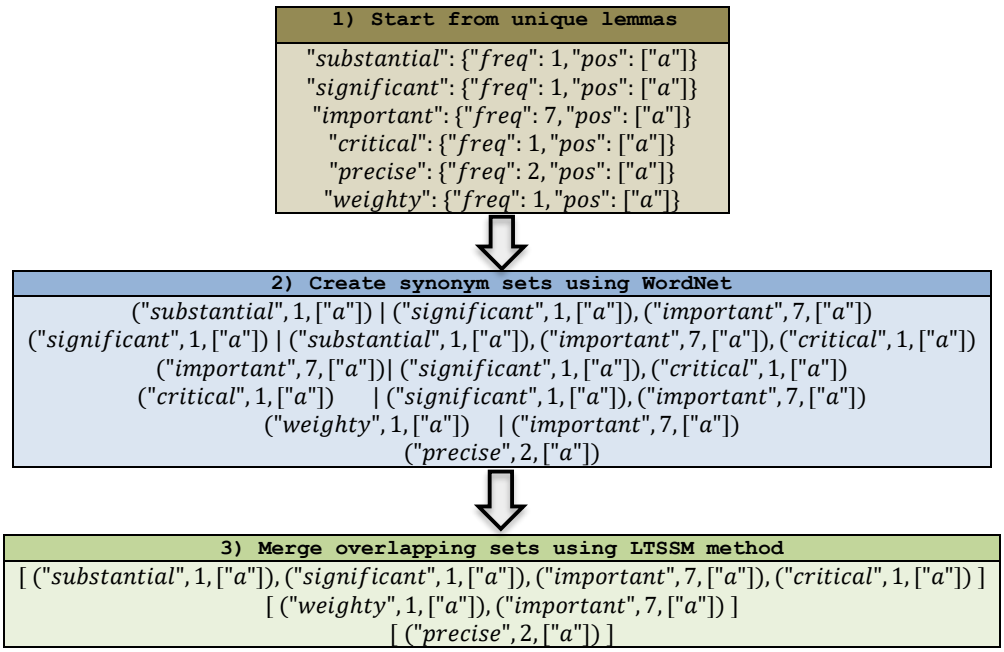

Figure 1*. Steps of creating and merging synonym sets*

3.2. Lexical variation metrics

3.2.1. Synonym set coverage

This measures the proportion of a text's vocabulary that belongs to at least one synonym set. High coverage indicates that the text contains more words with recognizable semantic overlap, which suggests homogeneity in vocabulary. The formula to compute coverage is:

$$Coverage = \frac{\sum_{w \epsilon V_{syn}} f(w)}{\sum_{w \epsilon V} f(w)}, \tag{2}$$

where V is the set of all lemmatized words (types) in the text; $V_{syn} \subseteq V$ is the subset of words included in synonym sets of size greater or equal to 2; and $f(w)$ is the frequency of word w.

The complementary to the synonym set coverage is uncovered frequency ratio, which is the proportion of words not covered by any synonym set.

3.2.2. Lexical reduction ratio

It estimates how much the text's vocabulary can be reduced by collapsing synonyms into a single representative term per set and calculated as:

$$reduction\_ratio = \frac{|V_{collapsed}|}{|V|}, \tag{3}$$

where V is the set of unique words (types) in the original vocabulary, and $V_{collapsed}$ is the number of unique representatives after synonym merging.

Lower values suggest higher lexical redundancy, meaning the text uses many synonymous expressions. For example, the synonym set

[("substantial", 1, ["a"]), ("significant", 1, ["a"]), ("important", 7, ["a"]), ("critical", 1, ["a"])]

can be collapsed to one representative, "substantial", so the reduction ratio for this set is 1/4.

3.2.3. Collapsed type-token ratio (TTR)

This metric assesses lexical diversity after synonym collapsing by measuring how varied the text remains when synonyms are treated as equivalent. High values indicate rich and diverse vocabulary even after synonym merging, while low values suggest conceptual repetition or redundant phrasing. This is computed as:

$$collapsed\_ttr = \frac{|Types_{collapsed}|}{|Tokens_{collapsed}|}, \tag{4}$$

where $Types_{collapsed}$ is the number of unique representative forms after collapsing. $Tokens_{collapsed}$ is the total number of tokens (including repetitions), with all synonyms replaced by their representative. For example, the synonym set

[("substantial", 1, ["a"]), ("significant", 1, ["a"]), ("important", 7, ["a"]), ("critical", 1, ["a"])]

can be collapsed to one representative, "*substantial*", and considering also the frequency of the tokens, the collapsed TTR is 1/10.

3.2.4. Pairwise conceptual similarity

Jaccard index is used to quantify conceptual similarity between two texts based on their synonym set usage. Instead of comparing literal words, we compare the conceptual units (synonym groups) used in each text. Thus, A denotes the set of synonym groups found in Text 1, B denotes the set of synonym groups in Text 2 and matching is based on a thresholded group similarity function.

$$Jaccard_{approx}(A, B; \theta) = \frac{|Matched(A, B; \theta)|}{|A| + |B| - |Matched(A, B; \theta)|} \tag{5}$$

where A and B are the sets of synonym groups in the two texts. $\theta \in [0,1]$ is the minimum required overlap between two groups for them to be considered a match. $Matched(A, B; \theta)$ is the set of synonym group pairs (a, b) such that

$$\frac{|a \cap b|}{min(|a|, |b|)} \geq \theta. \tag{6}$$

Two groups are counted as matching if they share sufficient word overlap (based on a normalized threshold) and compatible POS tags. The resulting value falls between 0 (no conceptual overlap) and 1 (perfect overlap in synonym usage).

4. Results

4.1. Lexical variation

Lexical variation metrics are computed for each text as the average of the calculated values for each synonym set. In this way, these metrics characterize the text independent of its length. For comparison, we consider the pairwise differences of the calculated values. In each case, we compare the AI-generated text to the human-authored counterpart, so a positive difference means that the AI text has higher value

in terms of the given metric. Although direct comparison can only be made between topic-aligned texts, this framework allows us to infer rankings between the different authors.

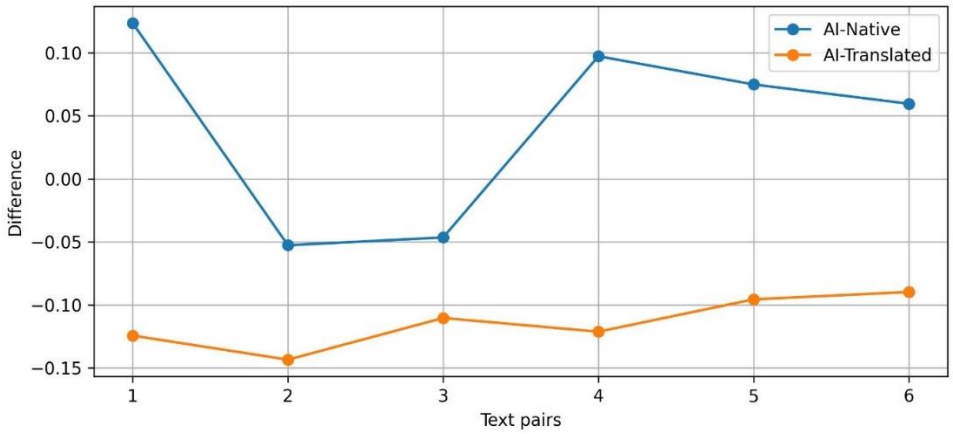

Figure 2. *Trends in Δ Coverage*

The trendlines in *Figure 2* show the results from the pairwise comparison of synonym set coverage values. As can be seen, non-native texts produce constantly higher coverage than AI-generated texts, which means that non-native speakers tend to reuse common vocabulary while AI uses more rare words. When AI texts are compared to native texts, the differences in coverage are smaller and rather positive. Thus, we can say that for synonym set coverage values the following ranking is valid among the different authors:

$$Coverage_{non-native} > Coverage_{AI} <> Coverage_{native}.$$

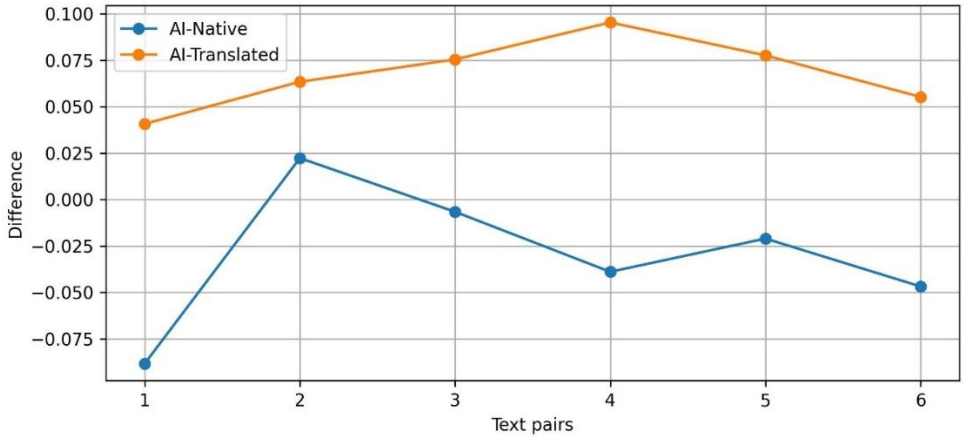

Figure 3. *Trends in Δ Reducation_ratio*

From *Figure 3* we can see the pairwise differences in lexical reduction ratio. The results show that this metric is slightly higher for AI texts than for non-native ones, which means that more words can be collapsed into synonym sets in the case of AI authorship. When comparing AI texts to native ones, the deviations are very small but rather negative. All together we can infer the following ranking among the authors in terms of lexical reduction ratio:

$$Reduction\_ratio_{native} >= Reduction\_ratio_{AI} > Reduction\_ratio_{non-native}.$$

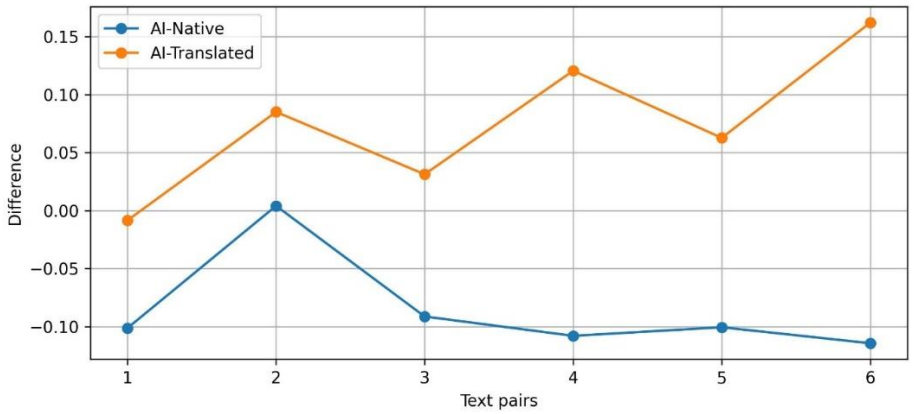

Figure 4. *Trends in Δ collapsed_ttr*

Figure 4 represents the pairwise differences in collapsed TTR. As is shown, this metric can be equal or slightly higher in the case of AI texts when compared to non-native ones, which indicates that AI uses broader synonym-based vocabulary. In contrast, native texts usually yield higher collapsed TTR values than AI ones. So the derived ranking among the authors is:

$$TTR_{native} > TTR_{AI} >= TTR_{non-native}.$$

4.2. Conceptual similarity

In this experiment, we made use of synonym sets to measure the similarity of the topic-related text pairs on a conceptual level. Using Jaccard similarity with variable thresholds, we found that very few synonym sets overlap exactly between the authors. However, partial overlaps increase as the threshold lowers which is illustrated in *Figure 5*. Notably, AI texts align more closely with native writers than with non-native ones. This highlights that AI can mimic authentic human writing quite well.

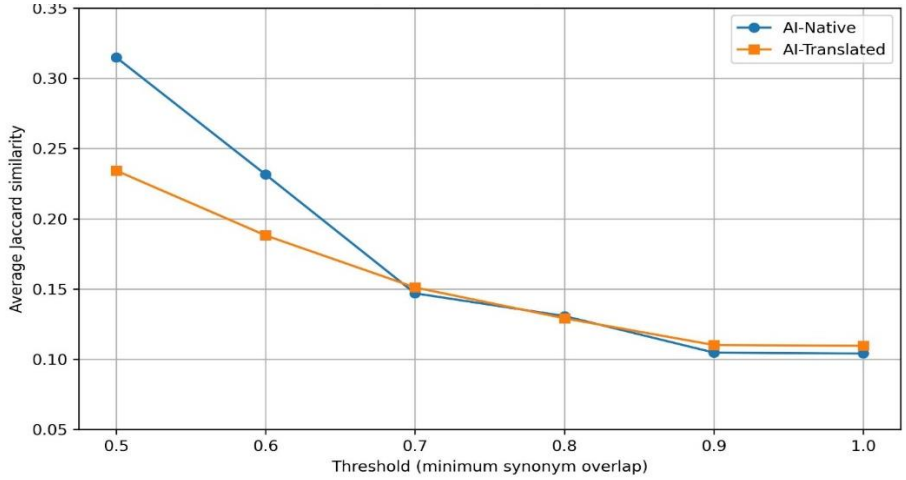

Figure 5. *Average Jaccard Similarity per Threshold*

5. Conclusion

This study introduces a novel synonym set-based methodology for profiling concept-level lexical variation in scientific writing to distinguish native, non-native, and AI-generated texts. Our key findings are as follows.

Jaccard similarity calculations show that native texts maintain consistently higher overlap in synonym set usage with their AI-generated counterparts than non-native texts. This implies that AI-written texts are more closely resemble native writing than non-native one. Non-native texts produce constantly higher synonym set coverage than AI-generated texts, and these texts exhibit greater lexical simplification and repetition after synonym consolidation, reflecting a more redundant conceptual structure.

These results affirm that specific lexical metrics can effectively help distinguish authentic texts written by non-native authors from AI-assisted ones, supporting more reliable detection methods.

The limitations of our work include its reliance on WordNet's lexical granularity: the clustering algorithm may over-generalize by merging semantically distant terms. Additionally, the study is restricted to English, so our findings may not generalize to other languages or wordnets. Finally, we did not include corpora of heavily AI-rewritten non-native texts, which might exhibit distinct stylistic patterns.

References

[1] Varga, B. E., Baksa, A. (2025). A comparative analysis of human and machine-written scientific texts. *International Scientific Conference on Computer Science and Information Technology*, University of Alesander Moisiu Durres, June, 2025, Albania, CSIT.

[2] Elkhatat, A. M., Elsaid, K., Almeer, S. (2023). Evaluating the efficacy of AI content detection tools in differentiating between human and AI-generated text. *Int. J. Educ. Integr.*, 19, 17. **https://doi.org/10.1007/s40979-023-00140-5**

[3] Weber-Wulff, D., Anohina-Naumeca, A., Bjelobaba, S. et al. (2023). Testing of detection tools for AI-generated text. *Int.J. Educ. Integr.*, 19, 26. **https://doi.org/10.1007/s40979-023-00146-z**

[4] Islam, Niful, Sutradhar, Debopom, Noor, Humaira, Raya, Jarin Tasnim, Maisha, Monowara Tabassum, Md Farid, Dewan (2023). *Distinguishing human generated text from ChatGPT generated text using machine learning.* https://arXiv:2306.01761.

[5] Berriche, L., Larabi-Marie-Sainte, S. (2024). Unveiling ChatGPT text using writing style. *Heliyon*, 10 (12), e32976. PMID: 38984302; PMCID: PMC11231544. **https://doi.org/10.1016/j.heliyon.2024.e32976**

[6] Rosenfeld, A., Lazebnik, T. (2023). *Whose LLM is it anyway? Linguistic comparison and LLM attribution for GPT-3.5, GPT-4 and Bard.* https://arxiv.org/pdf/2402.14533v1.

[7] Hakam, H., Prill, R., Korte, L., Lovreković, B., Ostojic, M., Ramadanov, N., Mühlensiepen, F. (2024). Human-written vs AI-generated texts in orthopedic academic literature: Comparative qualitative analysis. *JMIR Formative Research*, 8, e52164. **https://doi.org/10.2196/52164**

[8] Taloni, A., Scorcia, V., Giannaccare, G. (2024). Modern threats in academia: evaluating plagiarism and artificial intelligence detection scores of ChatGPT. *Eye*, 38, 397–400. **https://doi.org/10.1038/s41433-023-02678-7**

[9] Mindner, L., Schlippe, T., Schaaff, K. (2024). *Classification of human- and AI-generated texts: Investigating features for ChatGPT.* https://arxiv.org/pdf/2308.05341.

[10] Chien-Ying Chen, Jen-Yuan Yeh, Hao-Ren Ke (2010). Plagiarism detection using rouge and WordNet. *Journal of Computing*, 2 (3). **https://arxiv.org/pdf/1003.4065**

[11] Thompson, V. (2017). *Methods for detecting paraphrase plagiarism.* https://arxiv.org/abs/ 1712.10309.

[12] Álvarez-Carmona, M. A., Franco-Salvador, M., Villatoro-Tello, E., Montes-y-Gómez, M., Rosso, P., Villaseñor-Pineda, L. (2015). Semantically informed distance and similarity measures for paraphrase plagiarism identification. *Journal of Intelligent & Fuzzy Systems*, 34 (5), 2983–2990. **https://doi.org/10.3233/JIFS-169483**