A Principal Component Analysis of AI-Generated Medical English Texts: Examining Vocabulary Co-occurrence Patterns across Multi-Disciplinary Healthcare Domains

Emi Iwasaki
Faculty of Health and Welfare, Prefectural University of Hiroshima, Japan
1-1 Gakuen-cho, Mihara, Hiroshima, 723-0053 Japan
eiwasaki@pu-hiroshima.ac.jp

Abstract

In English for Medical Purposes (EMP) education, semi-technical materials accessible across diverse disciplines are critically needed. While Generative AI is increasingly used to design learning materials, empirical validation of how AI controls academic vocabulary across proficiency targets remains limited. This study evaluated an EMP corpus of 13,215 words (46 Gemini-generated dialogues) balanced across medical fields and three TOEIC target levels. Academic Word List items underwent Principal Component Analysis with Varimax rotation; factor scores were compared across levels using one-way ANOVAs with Dunnett's T3 post-hoc tests. PCA extracted two components accounting for 30.4% of total variance. A "Professional Base and Specificity" component showed no significant level-related difference, whereas a "Clinical Practice and Occupational Load" component varied significantly, $F(2, 43) = 14.65$, $p < .001$, $\eta^2 = .405$. Post-hoc tests revealed no difference between the 400 and 500 levels, but a significant surge at the 600 level. A student survey (N=72) supported this non-linear pattern, as perceived difficulty progression varied markedly by discipline (51.7%–90.5% agreement). These findings indicate that Generative AI adjusts lexical complexity non-linearly, with direct implications for prompt engineering in EMP materials development.

Keywords: English for Medical Purposes (EMP), Generative AI-generated materials, Learner perception

1. Introduction and Background

English for Medical Purposes (EMP) instruction faces a persistent design challenge: producing semi-technical reading and dialogue materials that remain accessible across a wide range of healthcare disciplines while still reflecting the interprofessional reality of clinical workplaces, where core professionals interact with allied health and support-service roles (Dudley-Evans and St John, 1998). This study encompasses a wide range of health and social care disciplines - such as nursing, rehabilitation, and social work—with dialogues also depicting the broader cast of roles (doctors, nurse assistants, supporting staff) who populate these settings. Because these fields share a core of general academic vocabulary (Coxhead, 2000) while diverging sharply in domain-specific terminology, EMP materials writers must balance lexical accessibility against professional authenticity.

Generative AI (GenAI) tools, particularly large language models, are increasingly used to generate levelled reading passages and dialogues at speed and scale (Knoth et al., 2024). Proponents highlight the reduced time cost of producing parallel texts across proficiency bands, while critics note that the internal criteria governing "difficulty" calibration remain largely opaque to end users. This contrasts with traditional readability formulas, which rely on transparent, quantifiable variables such as sentence length and syllable count (Flesch, 1948). Recent work assessing the large language model output directly has confirmed this gap, arguing that the versatility of language models makes a one-size-fits-all readability assessment difficult, since different generated texts may call for different evaluative criteria altogether (Marulli et al., 2024).

Empirical validation of how GenAI structurally controls and contextualizes academic vocabulary across proficiency targets therefore remains limited. Whether AI-generated texts distribute lexical difficulty smoothly or shift abruptly at particular thresholds has direct implications for how confidently practitioners can rely on GenAI output for graded EMP materials.

Equally important, but rarely examined alongside corpus-level analysis, is how the intended learners themselves perceive and respond to these AI-generated materials: a design that is structurally well-graded at the level of vocabulary co-occurrence may still be experienced by students as uneven or mismatched to their proficiency, making direct investigation of learner reception a necessary complement to corpus-based validation.

1.1 Purpose of the Study

This study addresses this gap by empirically examining the lexical co-occurrence structure of a corpus of AI-generated EMP dialogues spanning several healthcare disciplines and three TOEIC-referenced proficiency targets. Rather than relying on surface-level readability metrics, the study applies Principal Component Analysis (PCA) to uncover latent semantic components underlying academic vocabulary use, and tests whether these components differentiate systematically across proficiency levels. In addition, the study incorporates a supplementary survey of the students who used these materials to examine how learners across different healthcare disciplines subjectively perceived their difficulty, usefulness, and naturalness, providing a preliminary check on whether corpus-level patterns are reflected in learners' own experience.

2. Research Questions

RQ1: How does Generative AI structurally co-locate and contextualize academic vocabulary when generating multi-disciplinary EMP texts?

RQ2: To what extent do these AI-generated lexical patterns statistically differentiate between target proficiency levels (TOEIC 400, 500, and 600)?

RQ3: How do students across different healthcare disciplines perceive the difficulty, usefulness, and naturalness of these AI-generated EMP materials?

3. Materials

Text materials: The dialogues were systematically generated using an AI model (Gemini) with carefully controlled parameters to match specific proficiency levels and healthcare contexts. For instance, a prompt targeting a lower-intermediate level (TOEIC 400) and its corresponding AI-generated dialogue were structured as follows:

Prompt:
healthcare conversation between the healthcare staff and a patient. Toeic level 400, word count =100. Give me an exact word count at the end.

Gemini:
Ben: Good morning, Mrs. Miller. I am your new nurse for the day shift. I am taking over your care now.
Mrs. Miller: Hello. It is nice to meet you. Is my health okay today?
Ben: Yes, your condition is stable. You have a physiotherapy session today at 3:00 p.m. Do you need anything before we start?
Mrs. Miller: May I have a glass of water, please?
Ben: Of course. I will get that for you immediately. I am here if you need any help.
Mrs. Miller: Thank you very much for your kind assistance.

Word Count: 95 words. (Generated on Feb.2, 2026)

Furthermore, the materials covered 12 core-weekly topics distributed across healthcare disciplines, as listed below:

1. Greetings in Healthcare fields
2. Food & Drink
3. Asking and Giving Directions
4. Weather & Clothing
5. Sleep and Relaxation
6. Everyday Problems
7. Illness and Diseases
8. Injuries and Trauma
9. Activities of Daily Living (ADLs)
10. Exercise and Mobility
11. Signs and Symptoms
12. Computer and the Internet

4. Data and Corpus Construction

Following McEnery and Hardie's (2012) framework for specialized mini-corpora—which posits that collections of roughly 10,000 running words can yield reliable lexical patterns for well-defined Language for Specific Purposes domains—a purpose-built EMP dialogue corpus was constructed using Gemini over a semester. The resulting corpus comprises 46 AI-generated short dialogues, totaling 13,215 running words. To ensure breadth, the dataset was systematically counterbalanced across three target proficiency levels (TOEIC 400, 500, and 600) and various clinical roles across healthcare disciplines, centered on nursing, physiotherapy, and speech therapy alongside supporting medical staff.

To prevent discipline-specific vocabulary from confounding comparisons across proficiency levels, text generation was organized using a randomized rotation matrix spanning several generation cycles (weeks). Each target level was systematically paired with a different discipline in each week—for example, pairing TOEIC 400 with nursing in week 1, physiotherapy in week 2, and speech therapy in week 3—so that no single level was persistently associated with any one clinical domain. This counterbalancing design allowed subsequent statistical comparisons across levels to be attributed to level-related lexical control rather than to topical or disciplinary bias. A 6-week sample of the full 16-week rotation matrix is presented in Table 1.

	TL	WC	wk1	wk2	wk3	wk4	wk5	wk6
Q1	400	100	NS	PT	ST	Dr	Ms	NS
Q2	500	130	ST	Dr	Ms	NS	PT	ST
Q3	600	150	-	Ms	PT	ST	Dr	Ms

Table 1. Sample Rotation Matrix (6-week Excerpt from the 16-week Cycle) NB.: Abbreviations – TL (TOEIC level), WC (word count), wk (week), NS (Nursing), PT (Physiotherapy), ST (Speech Therapy), Dr (Doctor), Ms (miscellaneous)

The 46 dialogues were distributed across the three target levels as follows: TOEIC 400 (n = 15), TOEIC 500 (n = 15), and TOEIC 600 (n = 16). All texts were generated using an identical base prompt structure that specified the target TOEIC level, the healthcare discipline, and a target word count that increased incrementally with proficiency level (see Table 1); no additional stylistic or lexical constraints were imposed beyond the proficiency label and target length. This design choice was deliberate: the study aimed to capture how the model behaves under the kind of minimal, level-only prompting that a time-pressed materials developer might realistically use, rather than under heavily engineered prompts.

5. Methodology

- **Vocabulary selection:** Using AntConc, a corpus analysis toolkit, the corpus was cross-referenced against Coxhead's (2000) Academic Word List (AWL). To ensure that only vocabulary with sufficient distributional presence were retained for co-occurrence analysis, a minimum frequency threshold of five occurrences was applied. Within this 13,215-word corpus, exactly 14 AWL items met this threshold (*physical, stress, focus, job, professional, specific, primary, significant, relax, medical, virtual, network, technique, schedule*). Lowering the threshold (e.g., to

three occurrences) yielded no additional AWL types, confirming that these 14 words represented the complete set of sufficiently recurrent academic vocabulary in the dataset.

- **Principal Component Analysis (PCA):** A PCA with Varimax (orthogonal) rotation was applied to the co-occurrence patterns of these 14 words in order to extract underlying semantic components. Varimax rotation was selected to maximize the interpretability of loadings by producing components with a small number of strongly loading variables.

- **Statistical testing:** A series of one-way ANOVAs examined whether component (factor) scores differed significantly across the three TOEIC target levels. Because Levene's test indicated a violation of the homogeneity-of-variance assumption ($p < .001$), Dunnett's T3 post-hoc procedure, which does not assume equal variances, was used for all pairwise comparisons.

- **Effect size:** Eta-squared (η^2) was calculated to quantify the proportion of variance in component scores attributable to proficiency level, supplementing the significance testing with a standardized measure of practical magnitude.

- **Learner perception survey:** To complement the corpus-based analysis with learner-perspective data, an anonymous questionnaire (Likert-scale items plus open-ended comments) was administered via the Moodle learning management system at the end of the semester to students who had used the AI-generated materials in class. Of 92 invited students, 72 responded (78.3% response rate), representing the primary enrolled student disciplines in the course (nursing, physiotherapy, and speech therapy). Although enrolled in different discipline-specific courses, all respondents completed the same test content (the same set of AI-generated dialogues), so that differences in survey responses reflect variation in learner perception rather than variation in material content.

This combination of PCA and inferential statistical testing allowed the study to move beyond a purely descriptive account of vocabulary frequency and to test, quantitatively, whether latent semantic structures were deployed differentially across proficiency targets.

6. Results

The PCA extracted two interpretable components, jointly accounting for 30.4% of total variance in vocabulary co-occurrence. Table 2 summarizes the components, their defining vocabulary, and the results of the associated ANOVAs.

Statistic	Component 1: Clinical Practice & Occupational Load	Component 2: Professional Base & Specificity
Variance Explained	15.3%	15.1%
High-Loading Words	*physical, stress, focus, job*	*professional, specific, primary, significant*
***F*(2, 43)**	14.65	not significant
p	< .001	> .05
η²	.405	-

Table 2: PCA Components and Level-Related Differences

6.1 Component 1: Clinical Practice and Occupational Load

Component 1 (15.3% of variance) was defined by high loadings on the words *physical*, *stress*, *focus*, and *job*. It reflects language associated with the physical and psychological demands of clinical work, evoking an occupational register centered on workload and bodily/mental exertion.

6.2 Component 2: Professional Base and Specificity

Component 2 (15.1% of variance) was defined by high loadings on *professional*, *specific*, *primary*, and *significant*. It reflects a more general register of professional and academic precision common across disciplines, independent of any particular occupational scenario.

Component 2 factor scores did not differ significantly across the three TOEIC levels ($p > .05$), suggesting that this general professional-register component is deployed comparably by the AI regardless of the targeted proficiency level.

6.3 The Non-Linear Surge in Component 1

By contrast, Component 1 factor scores varied significantly across levels, $F(2, 43) = 14.65$, $p < .001$, $\eta^2 = .405$, indicating that occupational-load vocabulary explained a very large share (over 40%) of the variance in factor scores associated with proficiency level. Dunnett's T3 post-hoc comparisons, summarized in Table 3, revealed a striking, non-linear pattern: there was no significant difference between the TOEIC 400 and 500 levels ($p = .155$), but the TOEIC 600 level scored significantly higher than both TOEIC 400 ($p = .001$) and TOEIC 500 ($p = .011$).

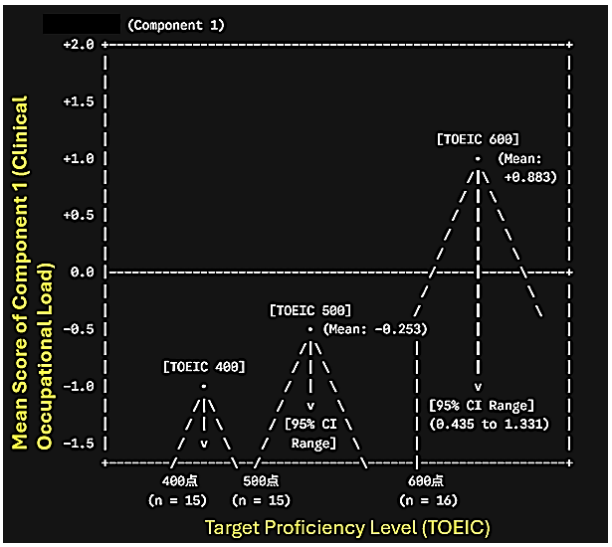

Figure 1: Mean Component 1 (Clinical Practice & Occupational Load) Scores Across TOEIC Proficiency Levels

Statistic	400 vs. 500	400 vs. 600	500 vs. 600
Mean Difference	small	large	large
p (Dunnett's T3)	.155	.001	.011
Significant?	No	Yes	Yes

Table 3. Dunnett's T3 Post-Hoc Comparisons for Component 1

In other words, rather than increasing steadily across the three levels, occupational-load contextualization remained essentially flat between the lower two levels and then surged sharply at the 600 level, forming a pattern more consistent with a step function than with a linear progression.

7. Discussion: The “Threshold” Effect

These results indicate that Generative AI does not scale lexical complexity in EMP texts along a smooth, linear gradient. Instead, the data point to a threshold, or "gear-shift," effect: once text generation crosses into the TOEIC 600 target, the AI system markedly tightens the co-occurrence of vocabulary associated with clinical workload and occupational stress, effectively simulating an authentic professional workplace register. Below that threshold, at the 400 and 500 levels, the AI appears to treat occupational-load vocabulary in a largely undifferentiated manner, producing comparable contextual density regardless of the intended proficiency gap between these two levels.

This pattern suggests a form of AI "tone-deafness" to fine-grained increments in lower-to-intermediate proficiency, even while the model is capable of sharply differentiating a higher target level. One plausible explanation is that the model's internal representation of proficiency labels such as "TOEIC 400" and "TOEIC 500" may not be sufficiently distinct in training-derived associations, whereas a marker like "TOEIC 600" may be more reliably linked in training data to descriptions of workplace-ready or professional-level English, triggering a qualitatively different generation strategy rather than a proportionally adjusted one.

For material developers, this means that simply specifying a numeric proficiency target may be insufficient to guarantee a proportionate adjustment in contextual and occupational density. Explicit prompt engineering, specifying not only target vocabulary but also desired levels of contextual density and occupational load, may be necessary to achieve genuinely graded materials at the lower-intermediate range. Moreover, the fact that Component 2 (professional base and specificity) remained stable across levels suggests that certain lexical registers may be comparatively robust to level-based prompting, while others, such as occupational-load vocabulary, are highly sensitive to a specific threshold. This distinction implies that materials developers cannot assume uniform AI behavior across all semantic dimensions of a text; some aspects of register may need to be manually adjusted even when others are handled adequately by the model.

7.1 *Preliminary Learner Perception Data*

To address RQ3, an anonymous end-of-semester survey (72 of 92 invited students, 78.3% response rate) was analyzed across nursing (n = 22), physiotherapy (n = 21), and speech therapy (n = 29). Overall, student evaluations were overwhelmingly positive regarding the usefulness and naturalness of the AI-generated materials, while revealing key nuances in perceived difficulty.

Regarding usefulness and professional relevance, an overwhelming 97.2% of respondents agreed (Likert 4–5) that learning scenarios from other professions helped them understand interprofessional team medicine. Furthermore, 91.7% reported high interest in content related to their own major, 88.9% appreciated the balance between technical and everyday language, and 84.7% expressed a desire to continue using major-tailored AI materials. In terms of naturalness, 90.3% of students perceived the AI-generated dialogue scenarios as realistic and natural.

Concerning perceived difficulty and vocabulary, 83.3% agreed that the vocabulary level was appropriate, and 77.8% felt the overall material difficulty matched their proficiency level. However, perception of step-by-step difficulty progression varied significantly by discipline: while 72.2% overall agreed that difficulty escalated incrementally, agreement ranged from 90.5% in physiotherapy and 81.8% in nursing down to 51.7% in speech therapy.

Qualitative feedback reinforced these quantitative findings: open-ended comments highlighted that students gained valuable exposure to interprofessional contexts. Taken together, these perception data indicate that while Generative AI successfully delivers highly useful, engaging, and contextually realistic EMP materials, its

non-linear difficulty progression and delivery-quality variations necessitate ongoing human oversight.

Dimension and Survey Item	Agree (%)
Usefulness & Relevance	
• Team medicine relevance	97.2%
• Interest in major-related content	91.7%
• Balance of technical/everyday English	88.9%
• Desire for future AI-material use	84.7%
Naturalness & Realism	
• Realistic AI dialogues	90.3%
Difficulty & Vocabulary	
• Appropriate vocabulary level	83.3%
• Overall difficulty match	77.8%
• Step-by-step difficulty progression*	72.2%

Table 4. Student Evaluation of AI-Generated EMP Materials (n = 72)
*NB.: Progression varied by discipline (PT: 90.5%, NS: 81.8%, ST: 51.7%).

7.2 *Pedagogical Implications*

- **AI tone-deafness:** GenAI output lacks fine-grained sensitivity to lower proficiency increments (400 vs. 500), even though it differentiates sharply at higher levels. Instructors should not assume that a nominal level difference in a prompt corresponds to a proportionate difficulty difference in the output.
- **Prompt engineering:** Material developers should build explicit constraints on contextual density and occupational load into prompts targeting intermediate levels, rather than relying on proficiency labels alone.
- **Contextual density over word difficulty:** Higher-level AI-generated texts do not simply substitute "harder" words; they embed vocabulary within denser, more occupationally loaded structural clusters. Vocabulary-focused readability checks alone may therefore miss important shifts in text difficulty.
- **Human review remains essential:** Given the threshold effect identified here, human review of AI-generated materials at intermediate levels is particularly important, as this is precisely the range where AI-driven differentiation appears weakest.

7.3 *Limitations*

Several limitations should be noted. First, the corpus, while consistent with McEnery and Hardie's (2012) guideline for specialized mini-corpora, remains modest in absolute size (13,215 words), and replication with larger corpora would strengthen confidence in the generalizability of the threshold effect. Second, the study examined a single GenAI model (Gemini) under a single, relatively minimal prompting strategy; different models or more elaborately engineered prompts might produce different scaling behavior. Third, the two extracted PCA components accounted for only 30.4% of total variance, indicating that a substantial portion of lexical co-occurrence structure remains unexplained by this two-component solution and may warrant further analysis with an expanded vocabulary set.

7.4 *Directions for Future Research*

Future studies could extend this design in several directions: (a) testing whether the threshold effect persists or shifts when explicit contextual-density constraints are added to prompts; (b) comparing multiple GenAI models to assess whether the TOEIC 600 threshold is model-specific or reflects a more general property of large language models; and (c) building on the preliminary learner-perception survey reported above by administering directly TOEIC-level-aligned difficulty-rating items, to more rigorously validate whether the statistically identified components correspond to perceived difficulty differences among actual EMP learners.

8. Conclusion

This study found that AI-generated multi-disciplinary EMP texts contain two lexical components, only one—an occupational-load component—varying significantly with proficiency, and non-linearly: little difference between the 400 and 500 levels, followed by a sharp surge at the 600 level. A supplementary student survey corroborated this pattern: despite high ratings for material usefulness and dialogue naturalness, perceived difficulty progression varied widely by discipline, indirectly supporting the finding that AI-generated difficulty does not scale smoothly. These results indicate that AI-driven difficulty adjustment operates as a threshold-based structural shift rather than a smooth ramp, underscoring the continued need for deliberate prompt engineering and human review in AI-assisted EMP materials development.

9. Ethical Approval

The student survey component of this study was approved by the Research Ethics Committee of Prefectural University of Hiroshima Mihara Campus (Approval No. 26MH019), and informed consent was obtained from all participants.

10. Conflicts of Interest

The authors declare no conflicts of interest.

11. Bibliographical References

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213–238. https://doi.org/10.2307/3587951

Dudley-Evans, T., and St John, M. J. (1998). *Developments in English for specific purposes: A multi-disciplinary approach*. Cambridge University Press.

Flesch, R. (1948). A new readability yardstick. *Journal of Applied Psychology, 32*(3), 221–233.

Knoth, N., Tolzin, A., Janson, A., and Leimeister, J. M. (2024). AI literacy and its implications for prompt engineering strategies. *Computers and Education: Artificial Intelligence, 6*, Article 100225. https://doi.org/10.1016/j.caeai.2024.100225

Marulli, F., Campanile, L., de Biase, M. S., Marrone, S., Verde, L., and Bifulco, M. (2024). Understanding readability of large language models output: An empirical analysis. *Procedia Computer Science, 246*, 5273–5282. https://doi.org/10.1016/j.procs.2024.09.636

McEnery, T., and Hardie, A. (2012). Corpus linguistics: Method, theory and practice. Cambridge University Press.

12. Language Resource References

Anthony, L. (2024). AntConc (Version 4.x) [Computer software]. Waseda University.

Google. (2026). *Gemini* [Large language model]. https://gemini.google.com