Kashihara, A. et al. (Eds.) (2024). Proceedings of the 32nd International Conference on Computers in Education. Asia-Pacific Society for Computers in Education

Exploring Linguistic Sophistication of Discussion Board Posts in University Learning Management Systems

Michelle P. BANAWAN[a*], Clarence James MONTEROZO[b] & Maria Mercedes T. RODRIGO[b]
[a]*Asian Institute of Management, Philippines*
[b]*Ateneo de Manila University, Philippines*
*MBanawan@aim.edu

Abstract: This study characterizes linguistic sophistication within university-based online courses using Learning Management Systems (LMS) across various academic disciplines. The research employs natural language processing tools to extract detailed linguistic features from student discussion posts and utilizes Principal Components Analysis (PCA) to identify distinct linguistic profiles. These profiles are analyzed to understand how linguistic sophistication varies across different educational contexts, specifically among various schools and courses. Subsequent cluster analysis reveals statistically significant distinct groups based on linguistic attributes. Despite the comprehensive analysis, the study did not establish significant predictive models linking linguistic sophistication to any direct educational outcomes. Instead, the findings highlight significant differences in language use across disciplines, suggesting that each academic field may have unique linguistic norms. The study emphasizes the need for further research to explore the underlying factors that influence these linguistic characteristics and their implications for educational practices.

Keywords: Linguistic sophistication, Online learing, Learning Management Systems, Discussion fora

1. Introduction

The evolution of online learning platforms has necessitated a deeper understanding of the factors contributing to student success in such environments. This exploration is critical as educational institutions increasingly rely on virtual classrooms to deliver their curricula. Previous studies have extensively analyzed student engagement and linguistic characteristics within Massive Open Online Courses (MOOCs), highlighting certain behaviors and language features as predictors of course completion and academic achievement. However, the generalizability of these findings to other online learning contexts, such as Learning Management Systems (LMS) used by universities, remains uncertain. For instance, a significant body of work, including that by Andres et al. (2017), has identified sophisticated linguistic patterns in written work as indicators of greater engagement and higher likelihood of course completion in MOOCs. Yet establishing if these indicators are present in academic discussion forums and if language sophistication translate to success metrics in LMS environments used within traditional university settings still needs further investigation most especially within underrepresented populations like the Philippines .

By situating our study within the broader discourse on language use in education—spanning from the implications of linguistic features in MOOCs to their potential influences in formal learning settings—we aim to provide comprehensive insights that could benefit educational researchers, curriculum developers, and e-learning technologists alike. This paper aims to bridge this gap by examining whether linguistic sophistication is related to academic success in university-based online courses. The motivation stems from previous findings suggesting a strong link between linguistic sophistication and educational outcomes in MOOCs and other informal online learning platforms.

1.1 Research Questions

The research questions below focus on the understanding the linguistic profiles and their distribution across different educational contexts.

> RQ1: *What principal components of linguistic sophistication characterize student discourse in university-based online courses?*
>
> RQ2: *How do these principal components of linguistic sophistication vary across different academic schools and specific courses?*
>
> RQ3: *What clusters of linguistic sophistication can be identified among students, and how do these clusters distribute across different courses and schools?*

2. Prior Work

Various dimensions of linguistic features were found to be related to positive learning outcomes, e.g. MOOC completion and academic success. A significant body of research has attempted to identify success indicators in MOOCs, given their vast and diverse participation. Crossley et al. (2016) found that participants who employed more sophisticated language were more likely to complete their courses. This supports the notion that linguistic sophistication—characterized by advanced vocabulary and complex syntactic structures—can predict course completion. Subsequent studies have explored these findings further. Andres et al. (2017) replicated the association between linguistic sophistication and MOOC completion rates. However, Monterozo and Rodrigo (2023) noted that these indicators did not hold in university LMS contexts, suggesting a divergence in success factors between MOOCs and structured LMS settings. This discrepancy highlights the need for further investigation into the dynamics of linguistic interactions across different online learning environments. In a related study, Banawan et al. (2021) utilized natural language processing tools to evaluate linguistic sophistication in student discussions on Math Nation, an online learning platform. Their results indicated that students with higher academic outcomes demonstrated more sophisticated language use, suggesting that linguistic ability significantly contributes to academic performance in math courses. These studies highlight the pivotal role of linguistic sophistication in online learning environments, emphasizing the need for deeper exploration into how language use impacts learner engagement and success.

3. Methods

3.1 Participants and Study Structure

This study was conducted in a university in Quezon City, Philippines, with a population of approximately 12,000 students. The data was collected during the COVID-19 pandemic when classes were conducted online. The current study used LMS logs recorded from August 25, 2021 to December 18, 2021 which comprises one fully online semester. The dataset used comprised of 3,429 different classes with 6,439 unique students enrolled. Furthermore, the classes in the dataset were limited to five schools of the university: education, humanities, management, science and engineering, and social sciences. The university's office in charge of research ethics reviewed the study protocol and the qualifications of the research team and gave the research team clearance to proceed.

3.2 Natural Language Processing

Fine-grained language sophistication features were extracted from the individual posts (original post and replies). This study employed two advanced tools: TAASSC (Tool for the Automatic Analysis of Syntactic Sophistication and Complexity) and TAALED (Tool for the Automatic Analysis of Lexical Diversity). Both tools were adapted from their original implementations (downloaded from their Github repositories) to suit the specific corpus of our

LMS discussion posts. Further preprocessing steps were undertaken to prepare the data for subsequent analysis. One of the primary preprocessing steps involved the exclusion of posts with an insufficient number of words (n<40), corresponding to the 25th percentile of the dataset. This cutoff was chosen because linguistic indices such as type-token ratio (TTR) can yield spurious or misleading results when applied to very short texts. The rationale is that with fewer words, there is less opportunity for lexical diversity, and the metrics that rely heavily on word count variability become less meaningful and potentially deceptive. This step resulted in a refined dataset of 196,056 entries from the original 203,477. The preprocessing steps reduced the total number of language sophistication indices from 335 to 82 indices.

The NLP approach employed in this study included Principal Component Analysis (PCA), clustering analysis, and variability analysis of the emergent components derived from the fine-grained linguistic indices. To further explore the linguistic profiles, a clustering analysis was performed on the scores of the principal components derived from the PCA. The optimal number of clusters was determined using the elbow method and silhouette analysis to ensure meaningful segmentation (see Figure 1).

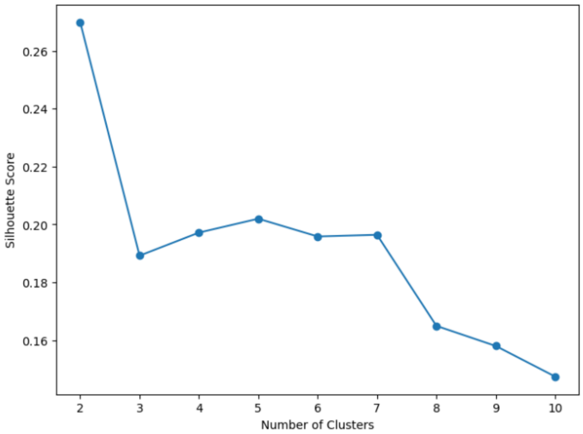

Figure 1. Silhouette Scores by Number of Clusters

4. Results and Discussion

4.1 PCA

Six principal components explained 35% of the variance in the data, with each component reflecting distinct aspects of language usage and sophistication in the discussion board texts. Based on the constituent fine-grained indices with significant loadings (greater than 0.30) per principal component, the following descriptions are presented:

1. *Component 1: Lexical Richness and Syntactic Complexity*. This component suggests a use of elaborate language and complex sentence structures, often associated with academic writing where detailed explanations and sophisticated argumentation are required.
2. *Component 2: Narrative and Explanatory Depth*. The indices that loaded in this component point to a capacity for detailed descriptions and explanations, which are critical for disciplines requiring extensive contextual or background information.
3. *Component 3: Discursive and Argumentative Complexity.* This component reflects argumentative or persuasive writing, highlighting a propensity to use language that supports claims or arguments with substantial evidence.
4. *Component 4: Clarity and Conciseness.* This component is characterized by negative loading on Complex Syntactic Structures which suggests simplicity and accessibility in language use. This component favors straightforward and accessible language, indicating a preference for clear and concise communication, which is essential in instructional settings.
5. *Component 5: Interpersonal and Functional Language Use*. This component suggests a focus on relational and functional aspects of language, typical in conversational or

collaborative academic settings where interpersonal communication is key.

6. *Component 6: Emotional and Evaluative Language*. This component is characterized by indices pertaining to the Frequency and Complexity of Verbs and Lemmas Associated with Evaluative or Emotional Content which indicates the use of expressive language to convey attitudes, values, and emotions. This component captures the expressive and evaluative aspects of language, which are crucial in fields like literature, the arts, or any discipline where subjective analysis and personal reflection are integral.

4.2 Cluster Analysis

The cluster analysis used the scores of the six principal components (PCs) identified through PCA, employing the K-means clustering algorithm to group similar instances into clusters based on their PC scores. Two distinct clusters emerged from this analysis, reflecting different levels of linguistic sophistication and usage patterns in the online academic discussions. The first cluster is characterized by high linguistic sophistication, featuring complex vocabulary and detailed language structures but less emphasis on interaction and conversational elements. This cluster likely represents more formal, structured, and detailed linguistic usage typical of academic or professional discourse. The second cluster, on the other hand, features simpler linguistic usage with some attention to interactive and conversational language elements but less overall structure and coherence. This cluster could represent more informal or conversational contexts where interaction is more common but with less complexity and detail in language use.

4.3 Variability Analysis of the Linguistic Profiles (Principal Components) between Schools and Courses

We used the Kruskal-Wallis test to evaluate the differences in principal components and average score percentages. Significant differences were found across schools and courses, with varying degrees of linguistic sophistication and academic performance metrics. Additionally, average score percentage showed highly significant differences, suggesting variability in grading standards or performance levels across different educational settings. In our analysis, we found that the lexical sophistication of academic texts reflects distinct disciplinary differences, consistent with findings from disciplinal differences of Crossley, Kyle, & Römer (2019) and conventions of rhetorical moves specific to the disciplines (Simanjuntak, 2022). Specifically, the variability analysis of principal components such as PC1, PC2, and PC6 revealed substantial differences both across schools and individual courses. Conversely, other components like PC3, PC4, and PC5 showed more pronounced variability among individual courses rather than between schools, highlighting the nuanced ways in which language use can vary within educational contexts.

4.4 Investigating the Relationship between Linguistic Sophistication Features with Academic Performance

Correlation analysis was performed to explore the relationships between the principal components (PCs) and the average score percentage, providing insights into how different linguistic aspects captured by the PCs relate to academic performance.

The correlation analysis reveals nuanced interactions between linguistic complexity and academic performance as captured by the principal components derived from discussion board texts. Notably, PC1, which encapsulates lexical richness and syntactic complexity, shows a modest positive correlation with average score percentage. This suggests that a certain level of sophistication in vocabulary and sentence structure might enhance comprehension and engagement, thereby potentially improving academic performance. Such features are likely appreciated in settings where analytical depth and precision in language are valued, such as in advanced humanities or social science courses.

Conversely, PCs that represent narrative depth, clarity, and argumentative complexity (PC2, PC4, and PC5) exhibit negative correlations with academic performance. This could indicate that while complexity adds depth, it may also introduce ambiguities or distract from the core message, particularly in examinations or assignments where conciseness and directness are rewarded. This is particularly salient in disciplines where clear and precise communication is critical, such as in the sciences and technical fields, where overly elaborate language might obscure essential scientific concepts. PC3, which shows the strongest negative correlation, highlights a critical educational insight: excessive argumentative complexity may not only fail to aid but actively hinder academic performance. This could reflect scenarios where students, in attempting to construct sophisticated arguments, may sacrifice clarity and coherence (Crossley & McNamara, 2012; Leung, Davison, & Mohan, 2014), leading to confusion and a dilution of key points in their communication (Tabari & Johnson, 2023). In contrast, PC6, which involves emotional and evaluative language, presents a strong positive correlation. This indicates that the ability to effectively convey emotions and evaluations might engage readers more deeply (Foolen, 2012; Puglisi & Ackerman, 2019), fostering a clearer understanding and stronger alignment with the rhetorical and evaluative aspects of coursework, particularly in subjects that value persuasive and affective communication.

4.5 Regression Analysis

Regression analyses were conducted to quantify the impact of each principal component on academic performance, adjusted for school-specific variations. Significant indicators of academic performance only emerged for the humanities and the science and engineering. A regression model with PC5 as a predictor shows a slight negative impact on performance for the humanities (Coefficient = -0.3827 , R^2 = 0.0034), suggesting that overemphasis on interpersonal and functional language use might reduce effectiveness in humanities disciplines, which require a balance of depth and interaction. PC3 and PC4 negatively affect performance in science and engineering, highlighting issues with excessive complexity. Conversely, PC6 positively influences performance, underscoring the value of clear and evaluative language in enhancing understanding and engagement in scientific contexts.

The regression analyses conducted provide insight into how specific components of linguistic behavior potentially influence academic performance within diverse educational settings. It is important to acknowledge from the outset that the explanatory power of these models, as indicated by very small R-squared values, suggests that linguistic sophistication alone does not predominantly drive academic performance. However, the analyses still reveal nuanced insights that contribute to our understanding of academic communication within specific disciplines. Despite the overall low explanatory power of these models, the significant results for specific components highlight the subtle yet relevant roles that certain linguistic features can play in academic settings (Galloway & Jenkins, 2009; Johnston, 2023), particularly linguistic sophistication (Kim et al, 2018; Kyle & Crossley, 2016).

5. Conclusions, Recommendations, and Limitations

This study focused on the exploration of linguistic sophistication within university-based online courses facilitated by an LMS. Our key findings include linguistic variability as revealed by the significant variability in linguistic profiles across different schools and courses. A PCA identified distinct dimensions of language use, including lexical richness, narrative depth, and emotional language, which varied significantly between educational settings. The two major clusters identified, representing high and low levels of linguistic sophistication, showed different distributions across schools and courses, suggesting that linguistic characteristics are reflective of specific academic cultures and expectations. A major limitation of this study is the absence of significant predictive models connecting the academic performance measure (average score percentage) and the language sophistication features extracted from the posts. Both fine-grained features and emergent components were utilized in predictive modeling attempts, but the outcomes suggested that factors other than those measured might play a more substantial role in determining academic success.

Our recommendation is that future studies should explore additional variables that may impact academic success in LMS environments, such as psychological factors, instructor-student interactions, and external educational resources.

Acknowledgements

The authors would like to express their gratitude to the Ateneo Laboratory for the Learning Sciences (ALLS), Ateneo Research Institute for Science and Engineering (ARISE), and Accenture for the funding and support. Additionally, we acknowledge the Asian Institute of Management for providing additional resources, including computing facilities, which significantly contributed to this research.

References

Andres, J., Baker, R., Gaševic, D., Siemens, G., & Spann, C. (2017). Replicating 21 Findings on Student Success in Online Learning. Technology, Instruction, Cognition and Learning, 10, 4. https://eric.ed.gov/?id=EJ1257952

Banawan, M. P., Balyan, R., Shin, J., Leite, W. L., & McNamara, D. S. (2021). Linguistic Features of Discourse within an Algebra Online Discussion Board. In EDM.

Crossley, S. A., Kyle, K., & Römer, U. (2019). Examining lexical and cohesion differences in discipline-specific writing using multi-dimensional analysis. Multi-Dimensional Analysis: Research Methods and Current Issues, 6, 189-216.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. Journal of Research in Reading, 35(2), 115-135.

Crossley, S. A., Paquette, L., Dascalu, M., McNamara, D. S., & Baker, R. S. (2016). Combining click-stream data with NLP tools to better understand MOOC completion. Learning Analytics and Knowledge. https://doi.org/10.1145/2883851.2883931

Foolen, A. (2012). The relevance of emotion for language and linguistics. Moving ourselves, moving others: Motion and emotion in intersubjectivity, consciousness and language, 349-369.

Galloway, F. J., & Jenkins, J. R. (2009). The adjustment problems faced by international students in the United States: A comparison of international students and administrative perceptions at two private, religiously affiliated universities. NASPA journal, 46(4), 661-673.

Johnston, P. (2023). Choice words: How our language affects children's learning. Routledge.

Kim, H. Y., LaRusso, M. D., Hsin, L. B., Harbaugh, A. G., Selman, R. L., & Snow, C. E. (2018). Social perspective-taking performance: Construct, measurement, and relations with academic performance and engagement. Journal of Applied Developmental Psychology, 57, 24-41.

Kyle, K., & Crossley, S. (2016). The relationship between lexical sophistication and independent and source-based writing. Journal of Second Language Writing, 34, 12-24.

Leung, C., Davison, C., & Mohan, B. (2014). English as a second language in the mainstream: Teaching, learning and identity. Routledge.

Mancilla, R. L., Polat, N., & Akcay, A. O. (2017). An investigation of native and nonnative English speakers' levels of written syntactic complexity in asynchronous online discussions. Applied Linguistics, 38(1), 112-134.

Monterozo, C. J., & Rodrigo, M. M. T. (2023). Do the Same Rules Apply? Transferring MOOC Success Behaviors to University Online Learning. In 31st International Conference on Computers in Education. Asia-Pacific Society for Computers in Education.

Puglisi, B., & Ackerman, A. (2019). The emotion thesaurus: A writer's guide to character expression (Vol. 1). JADD Publishing.

Simanjuntak, R. R. (2022). Revealing the rhetorical moves and linguistic patterns in discipline-related undergraduate thesis. JOALL (Journal of Applied Linguistics and Literature), 7(2), 345-361.

Tabari, M. A., & Johnson, M. D. (2023). Exploring new insights into the role of cohesive devices in written academic genres. Assessing Writing, 57, 100749.