Journal of Information and Knowledge, Vol 60(3), June 2023, p.159-167

ISSN (Print) : 0972-2467
ISSN (Online) : 0976-2477
DOI: 10.17821/srels/2023/v60i3/171027

Identifying Stylometric Characteristics of Domain Specific Texts using Classification Algorithms: A Study of Library Science Articles Published in 2020

Mousumi Saha* and Saptarshi Ghosh

Department of Library and Information Science, University of North Bengal, Raja Rammohunpur – 734013, West Bengal, India; rs_mousumi@nbu.ac.in, sghosh@nbu.ac.in

Abstract

Academic writing has played an essential role in communicating the cognitive aspects of the human mind. Natural Language Processing (NLP) tools enable us to examine linguistic knowledge. However, writing patterns and applicable linguistic characteristics differ geographically. The study's primary purpose is to understand the global writing pattern and linguistic diversities of research articles in the LIS domain. The corpus was identified from four SCOPUS-enrolled open-access libraries and information science journals. The journals published in India and outside India were selected for the study in 2020. The syntactic complexity in 147 text documents was measured using the Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TASSAC). The corpus was further examined using the Structural Equation Model (SEM) to determine the causal relationship among independent variables such as syntax features and readability scores. The results depict the differences in the patterning of syntactic features at both the global and national levels. Furthermore, the study allows us to see how linguistic diversity is underplayed in research writings and helps to understand writing patterns through cross-country comparisons. Furthermore, the paper employs model-based reasoning to identify global and national latent variables.

Keywords: Corpus Linguistic, Noun Phrase Complexity, Readability, Structural Equation Model, Syntactic Sophistication

1. Introduction

Natural Language Processing (NLP) combines linguistics knowledge, machine learning, and computer science to analyze massive amounts of natural language data to generate practical applications. Linguistic knowledge is represented as a system of constraints, a grammar, which defines all and only the possible sentences of the language (Haegeman, 2001). The structural Equation Model (SEM) is a comprehensive analytical framework incorporating various statistical techniques. These techniques are used in social and behavioural sciences to evaluate theories regarding the causal effects of one or more independent variables on one or more dependent variables and among those dependent variables themselves (Larsson *et al.*, 2021). Geographically, writing patterns and linguistic characteristics differ. This study attempts to comprehend global writing patterns and linguistic knowledge of research articles in the LIS domain. The study allows seeing how linguistic knowledge is used in research articles and how that linguistic knowledge is used in different contexts and enables us to understand writing patterns more effectively. This study attempts to comprehend the global writing pattern and applied linguistic variedness of research articles in the LIS domain. The study seeks to examine how linguistic knowledge (Grammar, Linguistic sophistication and Noun phrase complexity and other characteristics) is used in research articles and different contexts and enables us to understand writing patterns more effectively. Therefore, the study focuses on the following objectives: (i) To identify and compare linguistic contexts such as inflectional morphology and lexicography knowledge. (ii) To understand

**Author for correspondence*

cross-country academic writing patterns and provide comparative shreds of evidence for linguistic diversities in domain-specific texts. (iii) To impose SEM for identifying the latent variables used globally that provide concrete evidence for understanding linguistic attributes with multiple regressions of the datasets.

2. Literature Review

The stylometric approach identifies the changes in authorship attribution and writing patterns of several authors by applying several models like the vector space model, Classical multidimensional scaling, and so on (López-Escobedo *et al.*, 2013, Gómez-Adorno *et al.*, 2018). In the LIS domain, the degree of complexity of user queries, the (a) synchronous communication mode, the application of information service guidelines and manuals, and the overall characteristics and quality of a specified digital information service for librarians may all interact with stylometric features such as lexical richness, structural clarity, and interpersonal support (Park *et al.*, 2022). The study by Jeetpraneechai (2019) found that native English speakers and non-native English speakers, both groups of writers heavily relied on attributive adjectives, nouns as premodifiers, and prepositional phrases as postmodifiers, and there were no significant differences in the use of prenominal modifiers between both groups of students for the most part. Computer-based corpus linguistics in LIS research is described as the main techniques used to assess strengths and weaknesses of authentic texts with linguistic evidence, exemplify the value of corpus linguistics in the context of controlled vocabularies, reduce ambiguity, and improve retrieval and authorized terms (Bowker, 2018). Structural Equation Modeling, precisely measured variable path models in SEM, is flexible and allows testing of a priori hypotheses about causal relationships between multiple independent and dependent variables (Larsson *et al.*, 2021). Partial Least Squares Path Modeling (PLS-PM) of structural equation modeling to identify causal relationships between the accuracies of intrinsic and extrinsic evaluation of word embedded. PLS-PM models can be used to investigate the effect of hyper-parameters such as the training algorithm, corpus, dimension, and context window, as well as to validate the effectiveness of intrinsic evaluation (Han *et al.*, 2020).

3. Methods

The data corpus was identified from four different SCOPUS-enlisted open-access journals on library and information science to fulfill the objectives of this paper. *College and Research Libraries, Liber Quarterly, Annals of Library and Information Studies*, and *DESIDOC Journal of Library and Information Technology* are the journals selected for the study in 2020. Data are collected in PDF format from their respective websites. Selected data were converted to plain text, and the corpus was modified by removing author details, references, and journal details. The corpus consists of 147 text documents, and the syntactic complexity was measured using the *Tool for the Automatic Analysis of Syntactic Sophistication and Complexity (TAASSC).* TAASSC uses advanced natural language processing technology to reveal several fine-grained clausal and phrasal syntactic structures (Kyle, 2016). The result was further analyzed with the Structural Equation Model (SEM). Finally, the authors developed a path model based on the latent variables of the text corpus

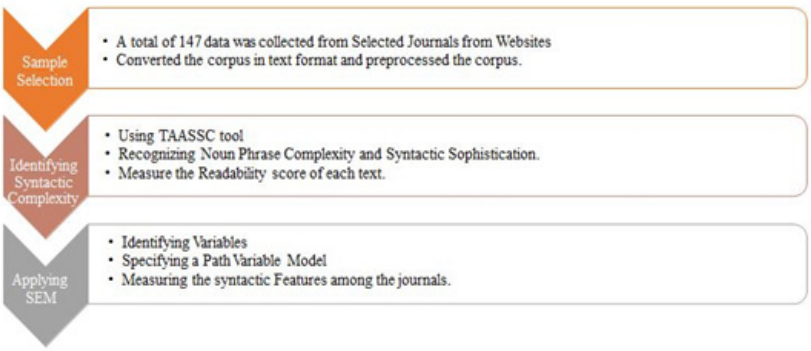

Figure 1. Diagrammatic representation of the method.

to identify the causal relationship between the dependent and independent variables. The study has identified three independent variables, viz., 'Readability score', 'Noun Phrase Complexity', and 'Syntactic Sophistication', which help to measure the writing pattern in articles in journals published in India and outside India in the LIS discipline. The diagrammatic representation of the methodology is shown in Figure 1.

The 'Readability score' indicates how simple or difficult the text is to read, and it varies from person to person. To assess the degree of accessibility of writing patterns, the readability score integrates statistical modeling, theoretical linguistics, and psychology theory (Dubay, 2004). The study used 'Flesch Reading Ease' and 'Flesch Kincaid Grade Level' as indicators of a dependent variable, the readability formula with the most applicability (Hamat *et al.*, 2022). TAASSC identified nine components related to 'Noun Phrase Complexity' and 'Syntactic Sophistication'. 'Noun Phrase Complexity' consists of noun Components such as "NP elaboration", "Nouns as modifiers", "Determiners," and "Possessives" describes the usage of nouns. 'Syntactic Sophistication' comprises five Components such as "Verb-VAC frequency" (Verb Argument Construction), "VAC frequency and direct objects" indicates the frequency of VACs, "Association strength" includes main verb lemma-VAC combinations, "Diversity and Frequency" and "Frequency" (Kyle, 2016).

4. Results

After identifying the variables, the study illustrates the path model for linguistic inquiries. Using the Partial Least Squares (PLS) path modeling method, the study compares the writing pattern statistically and examines the relationship between readability and syntactic features. The path model is measured using SmartPLS 4 (Ringle *et al.*, 2022), and models from the perspectives of journals published in India and those published outside India are compared.

The list of journals used for the study is shown in Table 1. The Annals of Library and Information Studies (NISCAIR, India) and the DESIDOC Journal of Library and Information Technology (DRDO, India) of Indian origin, and the College and Research Libraries (USA) and Liber Quarterly (Association of European Research Libraries, Europe) foreign publications, have been chosen for the application of the path model to identify the written pattern. 761976 words were analyzed from the 147 selected journal articles.

The PLS-derived measurement of the dependent variables across the text corpora is shown in Table 2. The highest standard deviation as measured by Flesch Reading Ease (FRE) of Readability is 8.749. Verb-Verd Argument Construction Frequency values of 'Syntactic Sophistication' and Noun Phrase Elaboration of 'Noun Phrase Complexity', with respective values of 6.744 and 5.280, are subsequent.

Table 3 shows the Reliability and validity construction of the studied corpus to provide an understanding regarding the consistency of the evaluation method of the linguistic data. Conrbach's Alpha and Composite Reliability is the most acceptable test for Reliability and validity construct with the recommended standard value of 0.70. Table 3 depicts Conrbach's Alpha of all variables as weak and establishes lower construct reliability. Composite Reliability (rho_a) indicates the highest validity in readability with a value of 4.928. Average Variance Extracted (AVE) measures the level of correlation between multiple indicators of the same construct. AVE is greater than or equal to 0.5, confirming the convergent validity (Shrestha, 2021). The analysis revealed that 'Noun Phrase Complexity' and 'Readability'

Table 1. List of journals studied

Name of the Journals	No. of Documents	No. of Words
Annals of Library and Information Studies	26	110052
DESIDOC Journal of Library and Information Technology	51	184846
College and Research Libraries	53	373493
Liber Quarterly	17	93585

Table 2. Statistical measure of variables of All LIS journals studied

	Variables	Mean	Standard deviation
Readability Scores	Flesch Reading Ease (FRE)	25.027	8.749
	Flesch Kincaid Grade Level (FKGL)	15.057	1.888
Noun Phrase Complexity	Noun as Modifier (nouns_as_modifiers)	-2.040	3.641
	Noun Phrase Elaboration (np_elaboration)	-1.22	6.744
	Determiners	6.122	3.451
	Possessives	6.802	2.739
Syntactic Sophistication	Association Strength (association_strength)	1.360	2.355
	Diversity and frequency (diversity_and_frequency)	-2.416	2.559
	Frequency	2.721	2.251
	Verd Argument Construction Frequency (vac_frequency)	2.040	2.611
	Verb-Verd Argument Construction Frequency (verb_vac_frequency)	6.122	5.280

Table 3. Overview of construct reliability and validity of LIS research

Variables	Cronbach's Alpha	Composite Reliability (rho_a)	The Average Variance Extracted (AVE)
Noun Phrase Complexity	0.053	0.78	0.516
Readability	-10.5	4.928	0.881
Syntactic Sophistication	-0.19	0.22	0.257

Table 4. Latent variables correlation among LIS research

Variables	Noun Phrase Complexity	Readability	Syntactic Sophistication
Noun Phrase Complexity	1	0.261	0.517
Readability	0.261	1	0.185
Syntactic Sophistication	0.517	0.185	1

are equal to or higher than the standard value, indicating Reliability than 'Syntactic Sophistication'.

Table 4 shows the correlational matrix of the latent variables of studied text corpora. The highest correlation is identified between 'Noun Phrase Complexity' and 'Syntactic Sophistication' (0.517) in the LIS journals studied. 'Readability' and 'Syntactic Sophistication' indicate a lower correlational value of 0.185 among the LIS journals Studied.

Table 5 depicts the direct relationship between variables. The use of 'Noun Phrase Complexity' (β = 0.366) and 'Syntactic Sophistication' (β = 0.233) in the context of the 'Readability score' demonstrates a significant level of direct effect. However, publications published outside India have a negative impact on 'Noun Phrase Complexity' (β = -0.461) and 'Syntactic Sophistication' (β = -0.101), indicating a problematic level of writing pattern for the readers.

Table 6 indicates the indirect effect between 'Readability' and 'Syntactic Sophistication' through the 'Noun Phrase Complexity.' The 'Readability score' (β = 0.032) towards the writing pattern influences the 'Syntactic Sophistication' intermediate by 'Noun Phrase Complexity' of the journals published in India. However, publication outside India negatively impacts the 'Readability score' (β = -0.255), indicating that the reader is less familiar with syntactic relations and that phrase ambiguity tends to increase the degree of reading difficulty (Von Glasersfeld, 1970).

Figures 2 and 3 depicted the Heterotrait-Monotrait ratio (HTMT) for evaluating variable discriminant validity. Discriminant validity assesses the variables that should not be related hypothetically to each other. Henseler *et al.* (2015) suggested acceptable levels of discriminant validity below 0.90. The result obtained in this study has established discriminant validity between the variables with the value of <0.09 in both Indian-published journals and journals published outside India.

Table 7 reveals the collinearity statistics used in the model to evaluate correlational variables in a linear

Table 5. Direct effect among variables

Variables (Publication from India)	**Path Coefficients**
Noun Phrase Complexity -> Syntactic Sophistication	0.087
Readability -> Noun Phrase Complexity	0.366
Readability -> Syntactic Sophistication	0.233
Variables (Publication Outside India)	Path C oefficients
Noun Phrase Complexity -> Syntactic Sophistication	0.553
Readability -> Noun Phrase Complexity	-0.461
Readability -> Syntactic Sophistication	-0.101

Table 6. Indirect effects among variables

Variables (Publication from India)	**Specific Indirect Effects**
Readability -> Noun Phrase Complexity -> Syntactic Sophistication	0.032
Variables (Publication Outside India)	
Readability -> Noun Phrase Complexity -> Syntactic Sophistication	-0.255

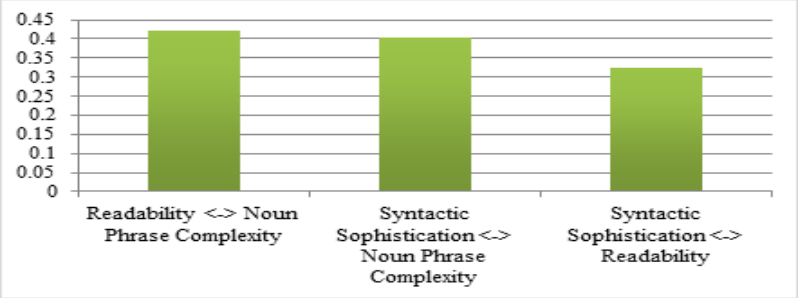

Figure 2. Heterotrait-Monotrait ratio (HTMT) publication from India.

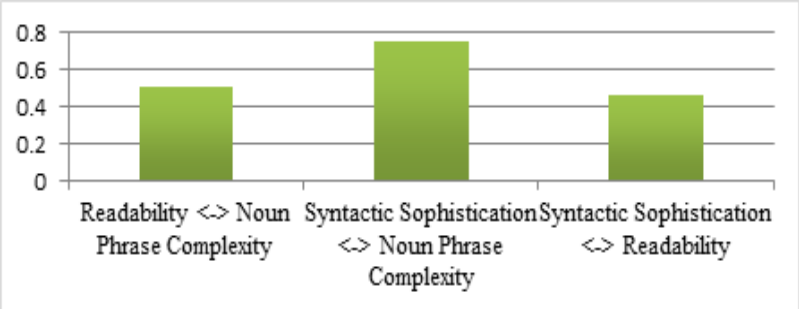

Figure 3. Heterotrait-Monotrait ratio (HTMT) publication outside India.

Table 7. Collinearity statistics (VIF)

Variables	Publication from India VIF (Variance Inflation Factor)	Publication Outside India VIF (Variance Inflation Factor)
Flesch Reading Ease (FRE)	3.589	5.433
Flesch Kincaid Grade Level (FKGL)	3.586	5.433
Noun as Modifier (nouns_as_modifiers)	1.1	1.211
Noun Phrase Elaboration (np_elaboration)	1.206	1.307
Determiners	1.588	1.309
Possessives	2.22	1.519
Association Strength (association_strength)	1.328	1.825
Diversity and frequency (diversity_and_ frequency)	1.515	2.318
Frequency	1.106	1.317
Verd Argument Construction Frequency (vac_frequency)	1.387	1.491
Verb-Verd Argument Construction Frequency (verb_vac_frequency)	1.74	1.226

Table 8. Comparison of fit indicators between models

Model Fit Test	Journal Published outside India	Journal Published in India	Required Value for Good fit
Standardized Root Mean Square Residual (SRMR)	0.134	0.126	<0.8
Normed Fit Index (NFI)	0.499	0.49	>0.9

relationship. According to accepted collinearity rules, if VIF is five or higher, it indicates a potential collinearity problem (Hair *et al.*, 2011). However, the higher value of collinearity between formative indicators found in journals published outside of India needs to be revised because it affects the variables' weight estimation and statistical significance.

The model fit test evaluates how the path models fit with the linguistic corpus of the LIS domain. SmartPLS4 recommended several indicators for evaluating the models. However, the study only focuses on SRMR and NFI indicators. As an absolute measure of the (model) fit criterion, SRMR enables evaluation of the average magnitude of the discrepancies between observed and expected correlations (Hu *et al.*, 1998). NFI calculates and compares the proposed model's Chi^2 value to a relevant standard (Bentler *et al.*, 1980). Due to the vast difference in datasets, Table 8 shows that the models fit as SMRM indicators. However, NFI rejects the models for both publications. For the journals published in India, the Noun Modifier (-0.841), Noun Phrase Elaboration (-0.692), Diversity and Frequency (-0.687), Verd Augmented Frequency (-0.189), and FKGL (-0.948) are the determining factors for the NFI test's output. For journals that were published outside of India, the following variables had an impact on the results of the NFI as Possessive (-0.701), Association strength (-0.610), Verb-VAC frequency (-0.027), Frequency (-0.610), and FKGL (-0.971).

5. Discussion

In this paper, we have examined the performance of stylometric-based features for detecting writing style changes in selected LIS journal articles published in India and outside India. Effective writing enhances the quality and reputation of journals among researchers. A crucial component of effective writing is the linguistic feature. Using the structural equation model, the study briefly discussed the quality of publication strength based on linguistic features. The structural equation model shows latent variables as circles (Figures 4 and 5), and dependent variables are shown as rectangles. The arrows indicate a relationship between the variables. Statistical significance was observed using SEM in both the cases and among the variables. The models are based on R-square statistics that explain the variance in the dependent variables explained by the independent variables. For example, 'Noun Phrase Complexity' is influenced by the four above-mentioned dependent variables with an R-Square value of 0.134 in Indian journals and 0.212 in journals published outside India.

Regarding 'syntactic sophistication', the R-Square value is 0.077 in journals published from India and 0.368 in journals published outside India. Cohen (1988) suggested that R-square values for endogenous latent variables are assessed as follows: 0.26 (substantial), 0.13 (moderate), and 0.02 (weak). It is evident that noun phrase complexity (0.136) is moderate and syntactic sophistication (0.077) is weak in journals published in India (Figure 4) as measured through Cohen's prescription. However, 'noun phrase complexity' (0.212) and syntactic sophistication (0.368) are close to 'substantial' in journals published outside India (Figure 5), as the pre-suggested R-square value. The relationship between 'Noun Phrase Complexity' and 'Syntactic Sophistication' is higher in journals published outside India, with a difference of 0.466. The higher value of 'Noun Phrase Complexity' indicates that the writing style is more informative than wordy (Maestre, 1998) for articles published outside India. Moreover, phrasal traits of such noun phrases are considered crucial elements of academic writing in native and non-native contexts (Staples *et al.*, 2016). The study shows that the LIS research publications published in India and outside India significantly considered 'Noun Phrase' complexity.

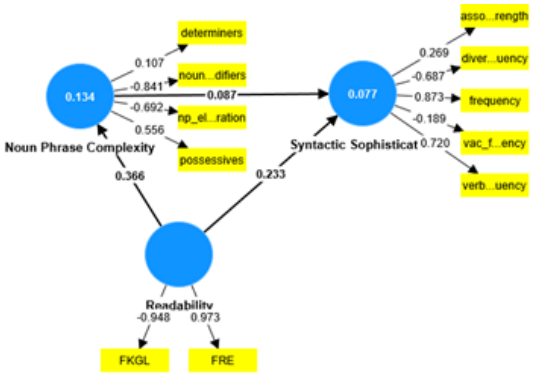

Figure 4. Path model (Journal Published from India).

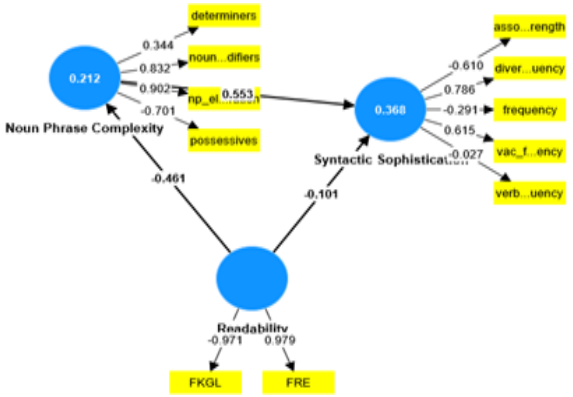

Figure 5. Path model (Journals published outside India).

Additionally, we discovered that for articles published in India, the causal relationship between 'syntactic sophistication' and 'readability' is more significant, with a score of 0.233, indicating a cultural influence (Ruijsscher, 2017) on writing style. Readability is a complex cognitive concept influenced by prior knowledge, motivation, and reading experiences. If the text's readability level is higher than the readers' ability level, readers may need help understanding the author's intended meaning (Eslami, 2014). In journals published outside India, syntactic sophistication causes difficulties in the readability score with a negative value. The study also found that inflectional morphology and lexicography knowledge are used in Indian journals to make them more understandable to their readers. Finally, the study found that complex noun phrases are widely employed in scholarly work published outside India to indicate meaning concerning complex syntactic structures.

6. Conclusion

The study adopts model-based reasoning using path models of SEM techniques that enable understanding linguistic analysis, in-depth knowledge of the literature, and linguistic interpretation. The study is the foundation for sophisticated techniques in the linguistic corpus of the LIS domain. The study indicates the unique structures of writing patterns statistically for understanding correlational aspects of linguistic features. It further recognized changes in writing style vis-à-vis other features, such as document embedding, which will be one of the directions for future work. The study's findings allow researchers to identify what needs to be done to improve academic writing skills, fill gaps, and increase syntactic complexity in writing.

7. References

Bentler, P. M. and Bonett, D. G. (1980). Significance tests and goodness of fit in the analysis of covariance structures. Psychological Bulletin, *88,* 588-606. https://doi.org/10.1037/0033-2909.88.3.588

Cohen, J. (1988). Statistical power analysis for the behavioral sciences. 2nd ed. New York: Routledge, p. 206.

de Ruijsscher, J. A. (2017). Cultural influences on the report readability of US-listed Asian Companies.

Dubay, W. H. (2004). The principles of readability. Costa Mesa: Impact information, p. 3.

Eslami, H. (2014) The effect of syntactic simplicity and complexity on the readability of the text. Journal of Language Teaching and Research, *5,* 1185-1191. https://doi.org/10.4304/jltr.5.5.1185-1191

Gómez-Adorno, H., Posadas-Duran, J. P., Ríos-Toledo, G., Sidorov, G. and Sierra, G. (2018). Stylometry-based approach for detecting writing style changes in literary texts. Computacion y Sistemas, *22,* 47-53. https://doi.org/10.13053/cys-22-1-2882

Haegeman, L. (2001). International encyclopedia of the social and behavioral sciences. In Linguistics: Theory of Principles and Parameters, 8957-8961. https://doi.org/10.1016/B0-08-043076-7/02956-9

Hair, J. F., Ringle, C. M., and Sarstedt, M. (2011). PLS-SEM: Indeed, a silver bullet. Journal of Marketing Theory and Practice, *19*: 139-152. https://doi.org/10.2753/MTP1069-6679190202

Hamat, A., Jaludin, A., Mohd-Dom, T. N., Rani, H., Jamil, N. A., and Aziz, A. F. A. (2022). Diabetes in the News:

Readability analysis of Malaysian diabetes corpus. International Journal of Environmental Research and Public Health, *19,* 6802. https://doi.org/10.3390/ijerph19116802

Han, N., Hayashi, K. and Miyao, Y. (2020). Analyzing word embedding through structural equation modeling. In LREC 2020 - 12th International Conference on Language Resources and Evaluation, Conference Proceedings, May 2020, Marseille, France, p. 1823-1832.

Henseler, J., Ringle, C. M. and Sarstedt, M. (2015). A new criterion for assessing discriminant validity in variance-based structural equation modeling. Journal of the Academy of Marketing Science, *43*, 115-135. https://doi.org/10.1007/s11747-014-0403-8

Hu, L.-T. and Bentler, P. M. (1998). Fit indices in covariance structure modeling: Sensitivity to underparameterized model misspecification. Psychological Methods, *3,* 424-453. https://doi.org/10.1037/1082-989X.3.4.424

Jitpraneechai, N. (2019). Noun phrase complexity in academic writing: A comparison of argumentative English essays written by Thai and Native English University students. LEARN Journal: Language Education and Acquisition Research Network, *12,* 71-88.

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine-grained indices of syntactic complexity and usage-based indices of syntactic sophistication.

Larsson, T., Plonsky, L. and Hancock, G. R. (2021). On the benefits of structural equation modeling for corpus linguists. Corpus Linguistics and Linguistic Theory, *17,* 683-714. https://doi.org/10.1515/cllt-2020-0051

López-Escobedo, F., Méndez-Cruz, C.-F., Sierra, G. and Solórzano-Soto, J. (2013). Analysis of stylometric variables in long and short texts. Procedia - Social and Behavioral Sciences, *95,* 604-611. https://doi.org/10.1016/j.sbspro.2013.10.688

Maestre, M. D. L. (1998). Noun phrase "complexity" as a style marker: an exercise in stylistic analysis. Atlantis, *20,* 91-105.

Park, J. R., Poole, E. and Li, J. (2022). Stylometric features in librarian's responses to user queries: implications for user interaction in digital information services. Global Knowledge, Memory, and Communication, ahead-of-print(ahead-of-print). https://doi.org/10.1108/GKMC-03-2022-0055

Ringle, C. M., Wende, S. and Becker, J.-M. (2022). SmartPLS 4 (No. 4). SmartPLS GmbH.

Shrestha, N. (2021). Factor analysis as a tool for survey analysis. American Journal of Applied Mathematics and Statistics, *9,* 4-11. https://doi.org/10.12691/ajams-9-1-2

Staples, S., Egbert, J., Biber, D. and Gray, B. (2016). Academic writing development at the university level: Phrasal and clausal complexity across level of study, discipline, and genre. Written Communication, *33,* 149-183. https://doi.org/10.1177/0741088316631527

von Glasersfeld, E. (1970). The problem of syntactic complexity in reading and readability. Journal of Literacy Research, *3*, 1-14. https://doi.org/10.1080/10862967009546930