外国語教育メディア学会（LET）関西支部
メソドロジー研究部会 第 15 号 報告論集
(2023 年 9 月)

発行　2023 年 9 月 30 日
編集・発行　外国語教育メディア学会(LET)関西支部
メソドロジー研究部会
代表　浦野 研
〒062-8605
北海道札幌市豊平区旭町 4－1－40
北海学園大学 経営学部
Tel: 011-841-1161
E-mail: urano@hgu.jp

Reports Vol. 15 of Japan Association for Language Education and Technology (LET),
Kansai Chapter, Methodology Special Interest Group (SIG) (pp. 1–14)
Nishina, Y. (2023).

Toward Probing into the Profile Genre: Scrutinizing the Profile Genre and the Methods of Automated Discourse Analysis

NISHINA, Yasunori
Kobe Gakuin University

Abstract

This paper presents the mechanism of the profile genre. It also suggests the automated methods that can be used in the discourse analysis of this genre instead of the qualitative move structure analyses used in previous studies such as Nishina (2021a, 2021b) and Nishina & Noguchi (2022). In order to capture the details of genre and subgenre specificity, more corpus-driven approaches should be used to elucidate the discourse studied with the two statistical methods, namely Decision Tree and Latent Semantic Analysis, which extract the features in each profile type based on linguistic properties. In this research note, I mainly summarise the literature review and the potential two methods that should be applied to the automated discourse analysis of the profile genre in detail as a first step for the research in question.

Keywords: profile genre, decision tree, latent semantic analysis

1. Introduction

Discourse analysis requires the labour-taking steady work of elucidating the structure and language properties that are consistent in the collection of texts peculiar to a particular genre/discipline through the study of words, phrases, collocations, colligations, patterns, semantic preferences, semantic/discourse prosody, semantic motif, move flow and so on (cf. Nishina, 2021a, 2021b; Swales, 1990; Tognini-Bonelli, 2001). Since lexico-grammatical patterns and move structures of discourse are closely related, the identification of such relationships at the micro level requires the labour-intensive manual work of researchers. Based on moderate corpus analysis, Nishina (2021a, 2021b) and Nishina & Noguchi (2022) qualitatively examined the discourse of the three types of profiles about artists, business executives, and companies in terms of latent move types, typical move structures, and language features used in each move (e.g., colligation, semantic preference, lexico-grammatical patterns)[1].

While these previous studies have provided specific and convincing findings,

such findings can remain within the limits of human examination and subjective interpretation by the investigators: the hidden, latent language properties and their patterns can be clarified and sometimes simplified by machine eyes and automatic analysis. Thus, the current study explains the combination of several statistical methods to potentially reveal the discourse features of the text collections of a particular genre as a hint for future studies.

2. The nature of the profile genre

2.1 The two approaches to the concept of genre

The concept of genre and its relationship to the discourse community differs, for example, between the approaches of Systemic Functional Linguistics (SFL) and the New Rhetoric School (NRS). In the SFL approach, it is generally assumed that "we are largely programmed by our societies into given ways of doing culture" (Lukin et al. 2011, p. 189) and that "genre can be defined in terms of linguistic properties alone" (Martin, 2003, p. 159) and that genre limits the choice of discourse structure (Martin, 1985; Ventola, 1987). From a pedagogical point of view, SFL practitioners recognise and teach the formal aspects of genre, such as "the functions, schematic structures and lexico-grammatical features in the texts" (Martin, 2003, p. 160), which are necessary for students to improve their input/output of the second/foreign language. In contrast to the formality-focused approach of SFL, the focus of the NRS approach is on the "sociocontextual aspects of genres" and the "social purposes or actions" fulfilled by genres (Hyon, 1996; Paltridge, 1997). For example, it examines the attitudes, values and beliefs of discourse communities using ethnographic methods, including interviews and observation. This is very different from the text-analytical approach of SFL (Hyland, 2000).

Although ESP researchers/practitioners adopt both approaches to genre study according to their research purpose, the majority in the field of ESP are strongly influenced by the SFL approach and consider the formal features of the texts to be significant. This is because such formal features (e.g. lexis, grammar, rhetorical structure) are effectively used in the teaching materials and classrooms of EFL/ESL learners. This is one of the main foci of ESP studies. However, there is a difference between SFL and ESP in terms of whether the focus is on the communicative purpose within a communicative situation (see details in Bloor, 1998; Martin, 2003).

2.2 Sub-genres of the profile

Context, (context of) situation and culture are the crucial elements in verbal

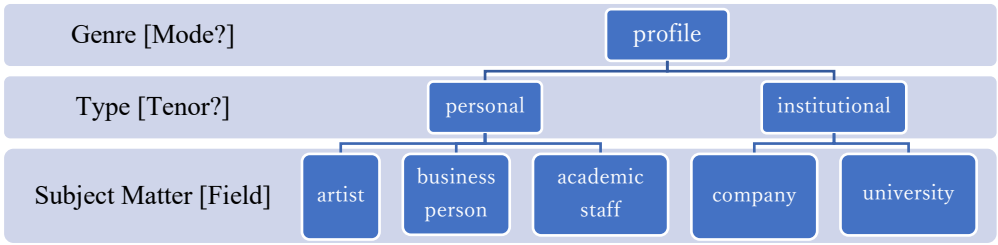
communication (Malinowski, 1923). This credo has been passed on, for example, to language studies by J.R. Firth and to SFL by M.A.K. Halliday. In the SFL framework, language takes on a higher order semiotic system, including context, semantics, lexico-grammar, expression and others in the layered system (Halliday, 1978). Context is related to field, mode and tenor (Halliday & Hasan, 1976, p. 22): The field indicates "the purposive activity of the speaker or writer", including the subject matter of the text; the mode indicates the purpose and "the function of the text", for example whether it is "spoken or written, extempore or prepared, and its genre"; The tenor indicates "the type of role interaction, the set of relevant social relations, permanent and temporary, among the participants". These three values construct the context of the situation of a text as assumed by Firth (1957). According to Hatim & Mason (1990, p. 49), genre is part of mode. This is because the mode prioritises the purpose of the text[2]. Hatim & Mason (1990) also point out that the field is not identical with the subject. More precisely, it is characterised by different subjects and is closely related in a given situation.

> Only when the subject matter is highly predictable in a given situation (e.g. a physics lecture) or when it is constitutive of a given social activity (e.g. a courtroom interaction) can we legitimately recognize a close link between field and subject matter (Hatim & Mason, 1990, p. 48)

The target genre in question is the profile genre. According to Biber et al. (1999) and Hirose (2018), the profile genre is a written mode and the main communicative purpose or content is to provide information about the target. The subgenres of the profile genre are mainly divided into personal profiles and institutional profiles. As Table 1 shows, the future study will use the three types of personal profiles (i.e. artists, business executives and academic staff) and the two types of institutional profiles (i.e. companies and universities). For your information, Figure 1 graphically illustrates the interrelationships between the five types. The profile genre (or mode) is divided into two types: personal and institutional. This level can also be interpreted as a tenor in terms of who is talking to whom. The personal profiles in this study are also divided into the profiles of art (i.e. artists), business (i.e. company CEOs) and academia (i.e. academic stuff) at the level of subject matter (or field). In contrast, the institutional profiles are divided into the two fields of business (i.e. company) and academia (i.e. university). At the level of subject matter (or field), the profiles of business people and companies are closely related, while those of academics and universities are also related.

Figure 1

The Interrelationships Between the Profile Sub-genres in the Future Study



2.3 Findings from the comparison of sub-genres

Each subgenre of the profile in Figure 1 has and may have specificity in move type, move structure and linguistic properties (see some details in Nishina, 2021a, 2021b; Nishina & Noguchi, 2022). Comparing the results of such previous studies provides some interesting insights. For example, some moves, such as *birth information, academic qualification*, and *residential/work location*, are commonly used in personal profiles regardless of the field. On the other hand, other moves are specific to the particular field. For example, the move *academic qualification* is prioritised in the personal profiles of the business field compared to those of the arts field[3]. This is probably because the work of art is all about the artist's evaluation, and artists are evaluated on the basis of their current actual performance. In other words, artists tend to take a more present and future-oriented position. On the other hand, business leaders prioritise their background, including their academic qualifications, to show that they are capable people to run businesses. Thus, business people value their history and experience in order to enhance their current bright careers.

In addition to these findings, personal profiles are also likely to contain a greater proportion of career information. This is because such information is more open to the public, whereas personal information tends to be more limited and should be open to the only closed group on SNS. Irrespective of the field, the purpose of publishing personal profiles on the websites needs to be publicly acknowledged, in terms of who s/he is, what s/he has done for the field s/he belongs to, and how s/he contributes to society. However, such career information moves are imbued with functional specificity in each field. For example, *job position* and *responsibility* are unique to the business field, while *critique* and *exhibition* are peculiar to the arts field. Discourse culture will therefore stand out by comparing the language/discourse features of different fields, and it would stand out by using the quantitative automated methods proposed in Chapter 4.

3. Proposed research questions for the future study

The main purpose of the profile genre is to introduce what it is (e.g. what the institution is or who you are) by providing information about the now and the past of the thing/person. If the purpose/function of the profile genre is similar and shared by the sub-genres of the profile, what is the difference between them? It is suggested that the difference may arise from linguistic properties and the structures that reflect them. In particular, the study of content words (i.e. nouns, verbs and adjectives) shows their apparent dissimilarity. While the use of a noun is influenced by the subject matter of the text, adjectives often indicate the writer's stance in accordance with the conventions of the specific discourse community. For example, Nishina (2021a) found that one-third of the key adjectives were common to the three subgenres of airline company profiles, but another one-third of the key adjectives were specific to each subgenre. In other words, the study of adjectives reveals cultural consistency within a sub-genre.

Looking at the verb tense reveals whether the focus in each sub-genre is on the past, present or future. As already mentioned, the profiles of businessmen prioritise the illustrious history in the past compared to artists: this is the matter at the level of the subject matter (or field). However, when comparing the profile types at the different levels, it is possible that the historical background information is prioritised more in the personal profiles than in the institutional ones. On the other hand, the current situation may be more prominent in the institutional profiles than in the personal ones. In other words, the priority of the current situation or of the history of what/whom depends on the type (or tenor?) or the subject matter (or field) of the profiles, according to Figure 1. This can be demonstrated, for example, with the corpus methods of verb tense counting.

Nishina (2021a, 2021b) and Nishina & Noguchi (2022) attempted to quantitatively and qualitatively elucidate the discourse features of several profile types using semi-manual corpus methods. The disadvantages of the methods used are, for example, the work/time involved in manually examining the entire discourse instances with their eyes and making the decision, including the (original) semantic categorisation. The future study, however, takes a more automated approach with one or both of the two statistical approaches of Decision Tree (hereafter, DT) and Latent Semantic Analysis (hereafter, LSA). The DT is one of the machine learning methods to represent the differences between the profile genre types with meaningful quantitative information. At the same time, LSA is a distributional semantic method to extract similarities among the documents studied. An overview of these two methods is presented in the following two sections. Here are the research questions for future studies:

(1) What are the meaningful quantitative differences between the different types of profiles?
(2) Which semantic themes are quantitatively extracted from each profile type?

With regard to RQ1, it is recommended to use DT to separate the data of linguistic properties by each profile type and to identify their quantitative characteristics. Regarding RQ2, the more semantic specificity of each profile type will be revealed by extracting the semantic topics with LSA and comparing them among the five profile types. Both (1) and (2) will contribute to the elucidation of the profile discourse in terms of formalities and meanings.

4. Proposed statistical methods to be used in the future study

4.1 The proposed method 1: Decision Tree (DT)

In the future study, DT may be conducted to identify the quantitative information for classifying profile types based on language data. DT is one of the machine learning methods of supervised learning. The results calculated by DT are easy to read due to the graphical visualisation with scores and priorities: DT creates the tree structure by repeating the classification of the input data based on the specific algorithm. Furthermore, DT can handle both quantitative and qualitative variables and its result is not affected by the outliers. Since the future study will use both types of variables, namely the profile types (e.g. personal or institutional, business or academic) as qualitative variables and the language data (e.g. the ratio of a part of speech) as quantitative variables, it is selected as the first method in the future study.

As detailed in Nishina (2023), the algorithm calculated in DT varies, such as AID, CHAID, exhaustive CHAID, QUEST, CART, ID3, C4.5 and C5.0. Among these options, the future study may choose the CART model (Breiman et al. 1984). This is because it avoids overlearning, handles both classification and regression, uses both qualitative and quantitative variables for both explanatory and objective variables, and is biantennary. These features are specific to the CART model compared to other algorithms. Shinmura (2002) also demonstrated that CART is a successful algorithm compared to others such as CHAID (Chi-squared Automatic Interactive Detector), Exhaustive CHAID and QUEST. C&RT produces a binary tree structure which is easier to interpret than the multi-branch trees produced by CHAID, C4.5(5.0). Shinmura (2002) suggested that misclassifications occurred more frequently in the multi-branch tree algorithms than in the binary tree algorithms because branching often stopped at the

upper node in the former. This is another reason why CART was chosen in this study.

Among several linguistic studies using DT, Tamaoka (2006) used it in Japanese linguistic research to reveal the position of the three types of Japanese connective particles that co-occur with the seven types of adverbs in a sentence. Okada (2007) also showed that the ambiguous pronunciations of the Japanese words 'Funiki' and 'Fuinki', both equivalent to English 'atmosphere', are uniquely classified based on attributes (e.g. year of birth, gender). Ishikawa (2013) also used the free software WEKA, invented by Waikato University, to investigate how lexical indices of English essays discriminate their writers. The DT, based on the C5.0 algorithm, analyses the type of writer in terms of lexical difficulty, lexical variety and sentence structure (for details, see Ishikawa, 2013). See also Nishina (2023).

4.2 The proposed method 2: Latent Semantic Analysis (LSA)

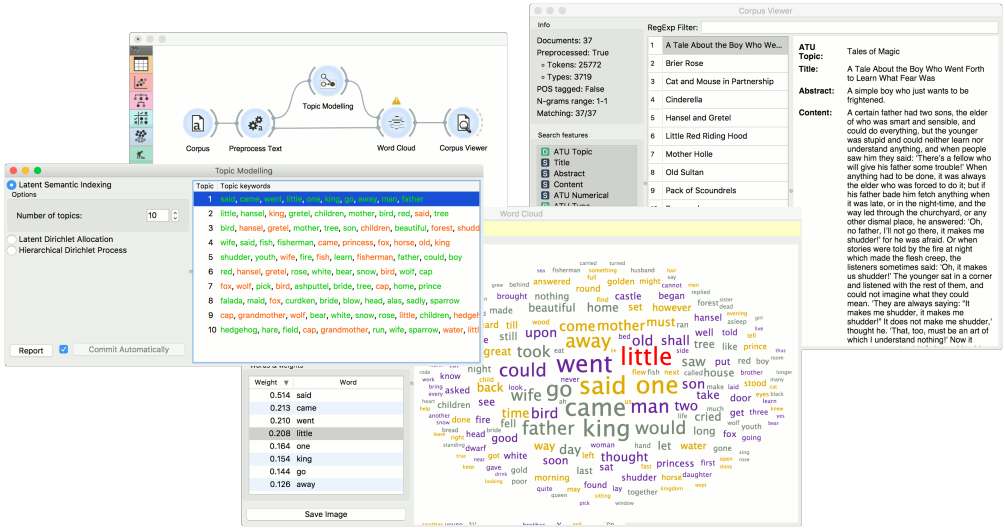
LSA[4] is an NLP technique in distributional semantics used for text summarisation and classification. It enables the detection of the underlying semantics of words in multiple texts by constructing topics related to words and texts. As detailed in Landauer et al. (2014) and summarised in Nishina (2023), LSA uses a document-term matrix in which rows correspond to texts and columns correspond to words. The advantages of LSA are to discover the similarity between a collection of texts based on linguistic data and to analyse certain relationships between words contained in a set of documents. Although some multivariate analyses, such as correspondence analysis, principal component analysis or cluster analysis, would probably outperform LSA in terms of visualising the relationships between variables and samples, LSA outperforms them in terms of semi-automatic topic discovery. Also, Latent Dirichlet Allocation (LDA), known as an extended and developed version of LSA, has also been used recently as a powerful analytical method, although the effectiveness of LSA over LDA has been partially confirmed in Fu et al. (2013) and Cvitanic et al. (2016). In this context and for this reason, LSA will be used as a second method in future studies. For more details on LSA, see Landauer et al. (2014).

To add, topic modeling strategies such as LSA and LDA can also be performed in software other than R. For reference, the following is a list of current software other than R that can perform topic modeling. First, XLSTAT, developed by Lumivero, is a software that allows statistical analysis of Excel data as it is, with LSA implemented in XLSTAT Marketing and Premium (https://www.xlstat.com/en/solutions/features/latent-sementic-analysis-lsa). Secondly, the software WordStat, developed by Provalis Research, implements topic modeling based on factor analysis

(https://provalisresearch.com/products/content-analysis-software/). Indeed, as noted in Peladeau & Davoodi (2018) and Peladeau (2022), it is more coherent, clearer and more successful in extracting a greater variety of topics when based on factor analysis than when based on LDA or neural network techniques. The third software presented here is Orange (https://orangedatamining.com/widget-catalog/text-mining/topicmodelling-widget/), which includes Latent Semantic Indexing (LSI), LDA and Hierarchical Dirichlet Process (HDP). Next, the Stanford Topic Modelling Toolbox (https://downloads.cs.stanford.edu/nlp/software/tmt/tmt-0.4/) is presented (the latest version 0.4.0 as of 25 June 2023). It supports spreadsheet-type data, such as Excel, and can perform LDA, Labelled LDA and Partically Labeled Dirchlet Allocation (PLDA). It is also possible to perform topic model analysis such as LDA in the Java-based MALLET (MAchine Learning for LanguagE Toolkit) (https://mimno.github.io/Mallet/topics.html) as well.

Figure 2

Sample Screen of Orange (extracted from https://orangedatamining.com/widget-catalog/text-mining/topicmodelling-widget/)



5. Compilation of profile corpora

Next, the DIY profile corpora were compiled from the texts extracted from the websites, as shown in Table 1. The main data of the profile corpora are also summarised in Table 2. The profiles of artists, CEOs and companies have been semi-manually

analysed by the author and the co-researcher in previous studies (e.g. Nishina, 2021a, 2021b; Nishina & Noguchi, 2022) in terms of move analysis (esp. move types, move structures and the language properties). In terms of artist profiles, only 23 profiles were examined in Nishina & Noguchi (2022), but the results of this study actually cover most of the generality in this discourse (e.g. move types and their structure) at around 60-80% from rough manual checking by the author. However, the size of the corpus is too small to be examined with the automatic method. For this reason, the new 152 profiles will be added to the artist profile sub-corpus. In addition to these corpora, the profiles of academic staff and universities are newly compiled for future study. The profiles of 152 artists will also be added to the existing corpus of artist profiles.

In particular, the academic staff corpus was compiled from the academic staff profiles on the Swansea University website in the UK[5]. The corpus of university profiles was also compiled from the SI-UK website (https://www.studyin-uk.com/): UK institution profiles page of the website, the profiles containing 'university' in the name of the institution were selected[6]. The parts of the texts entitled 'overview' were then extracted and compiled as the university profile corpus[7].

Table 1

Information About the Profile Corpora

Profile Type	Animacy	Field	Website	Note
Artists	Personal	Art	Kaikai Kiki Gallery (http://en.gallery-kaikaikiki.com/category/artists/); Art Profile (http://www.artprofile.co.uk/Index.asp)	175 artists (23 from the KaiKai Kiki Gallery and 152 from the Art Profile website)
Business Persons	Personal	Business	Star Alliance (https://portal.staralliance.com/employees/members)	28 airline CEOs
Academic Staff	Personal	Academics	Swansea University (https://www.swansea.ac.uk/staff/)	62 academic staff from Swansea University
Companies	Institutional	Business	Star Alliance (https://www.staralliance.com/); Oneworld (https://www.oneworld.com/); SkyTeam (https://www.skyteam.com/)	61 airlines
Universities	Institutional	Academics	SI-UK (https://www.studyin-uk.com/)	157 UK universities, excluding colleges and business schools

Table 2

Primary Data on the Profile Corpora

	Texts	Tokens (Ave)	Types (Ave)	TTR (STTR)	Sent (Ave)	Para (Ave)	AWL
Artists	175	3,043 (132.30)	2,141 (93.09)	70.36 (72.12)	136 (5.91)	70 (3.04)	4.99
CEOs	28	5,980 (213.57)	3,295 (117.68)	55.10 (57.32)	304 (10.86)	141 (5.04)	5.17
Academia	62	10,519 (169.66)	6,035 (97.34)	57.37 (61.43)	451 (7.27)	188 (3.03)	5.39
Companies	61	7,735 (126.80)	5,183 (84.97)	67.01 (70.67)	346 (5.67)	151 (2.48)	5.16
Universities	157	28,796 (183.41)	17,707 (112.78)	61.49 (62.92)	1,272 (8.10)	632 (4.03)	5.26

The future study will also adopt the assumption treated in the previous studies that if the members of the discourse community share a common knowledge and culture in a particular area, then regularly occurring linguistic features (e.g. words, phrases, collocations or patterns) will also be shared in each community, and that the quantitative differences of such features between genres and sub-genres will be provided: language and community are closely related. Based on this assumption, the future study will investigate the dissimilarity of profile genres and sub-genres using automated methods.

6. Concluding remarks: How will the analysis be conducted in future studies?

In conclusion, I would like to mention the future analysis using the profile corpora in Table 2. The future study will use DT and LSA. DT will be used to quantitatively identify the characteristics and differences between the five type profiles. Based on the cross-tabulation of the nine average scores per profile, including tokens, types, sentences, paragraphs, STTR (Standard Type-Token Ratio), AWL (Average Word Length) and the relative frequency of content words such as ADJ, ADV, N and V[8], the quantitative picture of each profile type will be presented by the tree structure using the algorithm CART model. However, it is naturally assumed that the higher the average number of tokens per profile, the higher the average number of types, sentences and paragraphs. For this reason, STTR, AWL and the relative frequency of the four parts of speech will be prioritised in the DT analysis.

The results will also be presented by LSA in order to identify the characteristics among the five type profiles. In LSA, the cross-tabulation of lemmatised term types and the number of documents for each profile type has to be created in order to perform LSA[9]: 2,141 terms x 175 documents (artists profiles), 3,295 terms x 28 documents (CEOs profiles), 6,035 terms x 62 documents (academics profiles), 5,183 terms x 61 documents (companies profiles), 17,707 terms x 157 documents (universities profiles). The type of clustering used in this study is 'fuzzy' to perform the classification in the newly created semantic space, where each element (term/document) can belong to several topics at the same time to represent a class (soft clustering). The results of the LSA will be obtained through these processes in future studies.

In summary, this paper has attempted to show the detailed future direction of profile corpora analysis by exploring the nature of profile genre, two statistical methods, extended profile corpora compilation and preparation for future analysis. I am confident that this study will be successful.

Acknowledgements

This work was greatly supported by the 2022 Kobe Gakuin University Long-term Overseas Research Fellowship Programme. I would like to thank Prof. Bjarke Frellesvig at the University of Oxford for his support during my research stay in the UK.

Appendices

1. Previous studies (e.g. Nishina, 2021a, 2021b; Nishina & Noguchi 2022) have found that some moves are constructed from relatively fixed language expressions, while others are relatively fuzzy. As the identification of moves by a single researcher may thus reflect his/her subjective view of the interpretation of language and discourse, these studies were double-checked by another experienced researcher, Prof. Judy Noguchi of Kobe Gakuin University. While the results reflect the value of such labour-intensive manual work, the alternative automated method is highly desirable.
2. Barr (2015) points out that both "[m]ode and register provide a means to identify formality in language" (p. 367). As mentioned earlier, a mode mainly indicates the means, primarily written or spoken. On the other hand, a genre/register is "a variety of language used in a particular social or economic setting" (Van Herk, 2012, p. 110), such as newspapers, legal texts, casual conversation, academic papers, etc.
3. As another finding on the move *academic qualification* in CEO profiles, those in Asian countries are more likely to emphasise the brilliance of his/her academic career with adjective collocates (e.g. *prestigious*). Social/ethnic factors, including regional

differences between Asian and Western companies, may have led to the differences in word choice in the profiles.

4. LSA is also known as Latent Semantic Indexing (LSI).
5. The corpus of academic staff profiles is mainly extracted from the College of Arts and Humanities websites, especially American Studies, Education, English Language, TESOL, Applied Linguistics, Modern Languages, Translation and Interpreting, from 15 March to 15 May 2020.
6. This study did not include the profiles of colleges and business schools for postgraduate students only.
7. Since many sections in the profiles have a specific section heading, such as "Services for International Students", "Ranking", "Accommodation" and "Location", i.e. the topics are already fixed in many sections, I tried to extract the general parts of the profiles by focusing on the "Overview" section.
8. TTR is affected by the size of the corpus. Therefore, STTR should be used in future studies.
9. Yasumasa Someya's lemma list is used to lemmatise the words in the list (https://lexically.net/wordsmith/support/lemma_lists.html). This list was created in 1998 and contains 40,569 words (tokens) in 14,762 lemma groups.

References

Barr, B. W. B. (2015). Chase or pursue: A corpus-based study. In P. Clements, A. Krause & H. Brown (Eds.), *JALT2014 Conference Proceedings* (pp. 364–377). JALT.

Biber, D., Johanson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English.* Longman.

Bloor, M. (1998). English for specific purposes: The preservation of the species (some notes on a recently evolved species and on the contribution of John Swales to its preservation and protection). *English for Specific Purposes, 17*, 47–66. https://doi.org/10.1016/S0889-4906(97)00044-6

Breiman, L., Friedman, J. H., Olshen, R. A., & Stone, C. J. (1984). *Classification and regression trees*. Chapman & Hall/CRC.

Cvitanic, T., Lee, B., Song, H.I., Fu, K., & Rosen, D. (2016). LDA v. LSA: A comparison of two computational text analysis tools for the functional categorization of patents. *Proceedings of the ICCBR 2016 Workshops*, 41-50. https://dblp.org/db/conf/iccbr/iccbr2016w.html

Firth, J. R. (1957). *Papers in linguistics 1934-1951*. Oxford University Press.

Fu, K., Cagan, J., Kotovsky, K., & Wood, K. (2013). Discovering structure in design

databases through functional and surface based mapping. *Journal of Mechanical Design, 135*(3), 031006-1–13. https://doi.org/10.1115/1.4023484

Halliday, M. A. K. (1978). *Language as social semiotic: The social interpretation of language and meaning*. Edward Arnold.

Halliday, M. A. K., & Hasan, R. (1976). *Cohesion in English*. Longman.

Hatim, B., & Mason, I. (1990). *Discourse and the translator*. Longman.

Hirose, K. (2018). Daiwa hyoushiki to gengo shiyouiki nitsuite. *The Center for Foreign Language Education, Shimane University, 13*, 1–15. https://ir.lib.shimane-u.ac.jp/en/list/journals/J/O-JCF/15/--/item/48986

Hyland, K. (2000). *Disciplinary discourse: Social interactions in academic writing*. Pearson Education.

Hyon, S. (1996). Genre in three traditions: Implications for ESL. *TESOL Quarterly, 30*, 693–720. https://doi.org/10.2307/3587930

Ishikawa, S. (2013). Lexical difficulty, lexical variation, and sentence structuredness: Which best discriminates between native and non-native writers? A decision tree analysis based on learner corpus data. *A Statistical Approach to Language Data, The Institute of Statistical Mathematics Cooperative Research Report, 290*, 107–124. https://www.ism.ac.jp/kyodo/index_i.html

Landauer, T. K., McNamara, D. S., Dennis, S., & Kintsch, W. (2014). *Handbook of latent semantic analysis*. Routledge.

Lukin, A., Moore, A. R., Herke, M., Wegener, R., & Wu, C. (2011). Halliday's model of register revisited and explored. *Linguistics and the Human Sciences, 4*(2), 187–213. https://doi.org/10.1558/lhs.v4i2.187

Malinowski, B. (1923). The problem of meaning in primitive languages. In C. K. Ogden & I. A. Richards (Eds.), *The Meaning of Meaning* (pp. 296–336). K. Paul, Trend, Trubner.

Martin, J. (1985). *Factual writing: Exploring and challenging social reality*. Deakin University Press.

Martin, P. (2003). Genre and discourse community. *ES: Revista de filología inglesa, 25*, 153–166. http://uvadoc.uva.es/handle/10324/17297

Nishina, Y. (2021a). Corpus-assisted discourse studies in airline company profiles: Through the lens of moves and adjectives. *English Corpus Studies, 28*, 1–25. https://jaecs.com/journal28.html

Nishina, Y. (2021b). Is an impressive background so important to a CEO? Investigating the move structure of personal profiles in the business field. *Kansai LET Collected Papers, 19*, 59–82. http://www.let-kansai.org/htdocs/?page_id=49

Nishina, Y. (2023). *Aspects of parallel corpus linguistics*. Kaitakusha.

Nishina, Y., & Noguchi, J. (2022). How artists describe themselves: The procedure and application of English language material for future artists in a Japanese university setting. *English, Media and Communication, 12*, 7–33. https://james.or.jp/gakkaisi/3005/

Okada, S. (2007). Validity of limit to "searching the corpus of spontaneous Japanese for now". *Ryukoku International Center Research Bulletin, 16*, 59–80. https://ndlonline.ndl.go.jp/#!/detail/R300000002-I8744306-00

Paltridge, B. (1997). *Genre, frames and writing in research settings*. John Benjamins.

Peladeau, N., & Davoodi, E. (2018). Comparison of latent dirichlet modeling and factor analysis for topic extraction: A lesson of history. *Proceedings of the 51st Hawaii International Conference on System Sciences 2018*, 615–623. http://hdl.handle.net/10125/49965

Peladeau, N. (2022). Revisiting the past to reinvent the future: Topic modeling with single mode factorization. *Natural Language Processing and Information Systems: 27th International Conference on Applications of Natural Language to Information Systems, NLDB 2022, Valencia, Spain, June 15–17, 2022, Proceedings*, 86–97. https://doi.org/10.1007/978-3-031-08473-7_8

Shinmura, S. (2002). The comparison between IP-OLDF and decision tree [the translation by the author]. *Japanese Society of Computational Statistics*, 64–67. https://doi.org/10.20551/jscstaikai.16.0_64

Swales, J. M. (1990). *Genre analysis: English in academic and research settings.* Cambridge University Press.

Tamaoka, K. (2006). Possibility of 'decision tree' analysis on collocation frequencies: In the case of conjunctive particles *kara, node* and *noni* co-occurring with adverbs at the middle and end of sentences. *Journal of Natural Language Processing, 13*(2), 169–179. http://dx.doi.org/10.5715/jnlp.13.2_169

Tognini-Bonelli, E. (2001). *Corpus linguistics at work.* John Benjamins.

Van Herk, G. (2012). *What is sociolinguistics?* Wiley-Blackwell.

Ventola, E. (1987). *The structure of social interaction: A systemic approach to the semiotics of service encounters.* Pinter.

外国語教育メディア学会 (LET) 関西支部メソドロジー研究部会 報告論集 第 15 号
菅井 康祐 (pp. 15–21)

実験研究の収集データの質を高める一工夫
―短期記憶課題における統制課題の例―

菅井康祐
近畿大学

Enhancing the Quality of Experimental Research Data: An Example of Control Tasks in Short-Term Memory Tasks

SUGAI, Kosuke
Kindai Univesity

Abstract

In studies encompassing psycholinguistics, applied linguistics, and the application of psychological experimental methods, quality of raw data is the most crucial factor influencing the success of the research. Uninterpretable noise in the collected raw data hinders a correct interpretation of the results, no matter how sophisticated statistical methods are employed. To prevent these situations, it is essential to rigorously control conditions during the experimental design stage. This helps minimize the impact of extra noise on the dependent variables. This report explores the impact of introducing additional tasks to control conditions in experiments involving data from a classroom-based short-term memory task.

Keywords: 実験研究，条件統制，統制課題，ノイズ，データの精度

1. はじめに

心理言語学・応用言語学を始め，心理学的実験手法を用いる研究において，精度の高いデータ，言い換えればできるだけノイズの少ないデータを収集することは研究の成否を左右するもっとも重要な要因の一つである。収集した元データに解釈のできない要因（ノイズ）が含まれてた場合，いくら高度な統計手法を用いて分析したところで，その結果が何を反映したものかを解釈することは非常に難しくなってしまう。このようなことを避けるために，実験をデザインする段階で可能な限り従属変数に余分なノイズが入らないように条件統制をすることが実験研究の基本である（浦野他，2016）。

E-Prime[1] や SuperLab[2] のような心理実験専用のツール・ソフトウェアを使えば比較的容易かつ正確にこのような条件統制をすることは可能である。しかし，そのようなツールが使用できない場合や，教室・オンライン授業のように，集団を対象とする環境で実験を行う場合には，実験に合わせてノイズを可能な限り小さくする工夫が必要になる。

この報告では，データの精度を高める一例として、教室環境で実施された短期記憶課題（音韻スパンテスト）の実験データをサンプルに，条件統制用の課題を追加し、ノイズの可能性が高いデータを取り除く方法を紹介する。

2. データ収集

2.1 実験協力者

日本語を母語とする大学 1，2 回生 138 名。

2.2 刺激

近畿方言話者の男性が発話した日本語の 5 母音/i/, /e/, /a/, /o/, /u/を録音し（SONY 社製コンデンサマイクロフォン ECM360，OLYMPUS 社製 PCI レコーダ LS-11 使用），Praat (5.3.39)を用いてそれぞれの母音を 190 ms に調整した。この 5 母音を刺激間隔（ISI）200 ms でランダムに 10 刺激並べたものを課題の 1 セットとした（各母音が 2 回提示されるように調整)。

課題セットの例：

/i/ - ISI - /u/ - ISI - /o/ - ISI - /a/ - ISI - /e/ - ISI - /e/- ISI - /u/ - ISI - /o/ - ISI - /i/ - ISI - /a/

2.3 手順

実験課題はは教室備え付けのスピーカーおよびプロジェクタ＋スクリーン（教室によってはセンターモニタ）から提示された。実験協力者は教室備え付けの PC および，各自の端末（スマートフォン等）を用いて，Google Forms で作成されたフォームに入力する形で解答するように指示された。課題は以下の時間軸で左から右に提示された。

課題提示の時系列：

聴覚: tone (100 ms) ----- 1 sec ----- stim----- (ISI 200) ----- tone ----- (ISI) 10 sec-

視覚: No #　　(a, b, c, d, e)

図 1

画面に提示された指示文

記憶課題説明

・「あ～お」の音が 10 個 1 組で続けて聞こえます
（例：い・う・あ・え・お・い・あ・お・う・え）

・前から順番にできるだけ多くの音を覚えて下さい

・「ピッ」という音が聞こえたら画面の文字(a, b, c, d,e)を選んでから

・覚えた音をできるだけ多く答えてください。

・解答時間が短いので，書ける範囲でかまいません。

実験協力者はまず，tone に合わせて画面に表示される問題番号を確認し，スピーカーから聞こえる課題を前から順にできるだけ多く記憶する。その後，tone（「ピツ」という音）と共に画面に表示されるアルファベット（a, b, c, d, e）をフォームに入力し，続けて記憶した文字列をフォームに入力するように求められた。実験協力者はこの課題 5 セットを実施した（実際にはここではしようしていない課題を含めて 7 種類の課題があるので 5 x 7 の 35 セットプラス本課題前後のダミー5 セットを加えた 40 セット，所要時間約 10 分）。

図 2

Google Forms の入力画面

3. 分析

このデータ採集では，本課題の短期記憶課題の回答入力の前に，統制課題として画面に提示されるアルファベット（a, b, c, d, e）を入力させるという手法を用いた。この統制課題により以下の 2 点の条件を統制することを試みた。一点目は，実験協力者の最低限の集中度を確認すること，つまり，アルファベットを入力するだけの課題に答えられない場合はかなり集中力が落ちていると考えられる。二点目は，解答の入力開始のタイミングを統制すること，つまり，課題音声が提示されている最中に入力が開始（フライング）された可能性のあるデータを検出・除外することである。本節では，統制課題（アルファベット入力）を要因とし，正しく入力されているものと，入力が不正確・未入力であったものとに分類し比較することで，反応統制課題にどの程度の効果があったかを検証する。

表 1

記述統計

コード	n	Mean	S. D.	95% CI
正解	680	2.75	2.24	[2.59, 2.97]
不正解・未入力	104	2.67	2.58	[217, 3.18]

図 3

統制課題の解答による比較

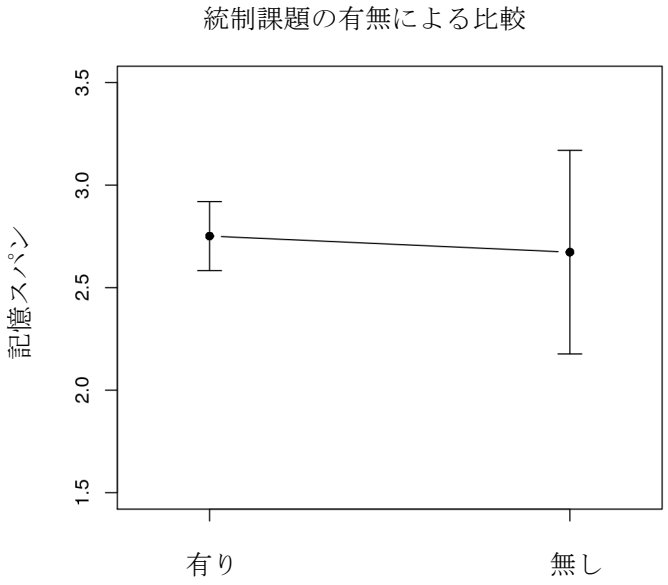


表 1・図 3 で示された通り，正しい統制課題の入力の有無によって得られるデータの平均値に有意な差はないものの（t (782) = 0.33, p = 0.74, d = 0.03, 95% CI [-0.17, 0.24]），統制課題に正解できていない試行データのばらつきがかなり大きい。このことは、統制課題によってノイズの可能性が高いデータを検出することができているということであり、これらのデータを分析対象から除外することでより精度の高い分析が可能になる。

4. おわりに

本稿では，日本語の母音を用いた短期記憶課題というかなりシンプルな実験の事例を用いて，統制課題による条件統制の効果を検証した。このようなシンプルな課題でも一つの統制課題を加えることでこれだけデータからノイズを取り除くことができるということは，応用言語学などのように複雑な実験パラダイムを用いる分野においては，適切な統制課題・条件統制にはより大きな効果があると考えられる。本稿はあくまでも一つの事例ではあるが，さまざまな実証研究において考案された条件統制の方法を共有することで，それぞれの分野においてより質の高いデータ収集・データ分析の質の向上に繋がることが期待される。

注

1. 米国 Psychology Software Tools 社が制作・販売している心理学実験パッケージ。2023 年 9 月時点で version 3.0 が公開されている（https://pstnet.com/）。（https://pstnet.com/）
2. 米国 Cedrus 社が制作・販売している心理学実験パッケージ。2023 年 9 月時点で version 6.0 が公開されている（https://cedrus.com/index.htm）。

参考文献

Boersma, P & Weenink, D. (2013). Praat: doing phonetics by computer [Computer program]. Version 5.3.39, retrieved 5 February 2013 from http://www.praat.org/

Mizumoto, A. (n.d.). Langtest. Retrieved from http://langtest.jp/#app

浦野研・亘理陽一・田中武夫・藤田卓郎・髙木亜希子・酒井英樹 (2016)『はじめての英語教育研究―押さえておきたいコツとポイント―』研究社

Reports Vol. 15 of Japan Association for Language Education and Technology (LET),
Kansai Chapter, Methodology Special Interest Group (SIG) (pp. 22–43)
Spring, R. (2023).

A Human-AI Integrated Rating Scheme for Improving Second Language Writing: The Case of Japanese Learners of English for General Academic Purposes

SPRING, Ryan
Tohoku University

Abstract

In order to solve the problem of teachers not assigning and evaluating student writing but not completely trusting AI raters, I created and tested a rating scheme in which an AI model would rate students' language use based on understandable criteria and humans would quickly check the AI responses while rating content and structure. Teachers tried the scheme and improvements were made based on new data and newly available research. An online practice tool was also created for students so that they could understand how the AI would rate their language use and practice accordingly. The AI rating models were improved over the course of three semesters based on student data and the ratings of external professional raters. As a result, an increasing number of teachers used the rating scheme, the number of students that practiced writing and were evaluated increased university-wide, and reasonable levels of fairness assessment were maintained.

Keywords: Automated Rating, Human-AI Integration, CAF Measures

1. Background

1.1 Educational Context and Problem

In 2020, Tohoku University, a university in Japan with a high national ranking and strong focus on science and engineering, initiated a new general education English as a Foreign Language (EFL) curriculum for its students based on the principles of English for General Academic Purposes (EGAP). As part of the curriculum, the university created its own in-house textbook, *Pathways to Academic English*[1], which outlined the skills that students are expected to learn in their general education EFL classes and detailed the exact points that they should focus on. Teachers were asked to use the textbook and teach the skills according to the book but were given much freedom regarding how to best teach the details and enhance students' skills. Practice materials and end of semester tests

were provided to teachers, but their use was not mandated. The practice materials consisted of worksheets, videos and audio files that matched the contents of the textbook. The end of semester tests consisted largely of multiple-choice questions, but also included speaking and writing questions, depending on the content of the course.

Amongst the skills outlined in the textbook were two writing skills: summary writing and paragraph writing. The former refers to a type of source writing in which students read a long passage of about 400 words and rewrite the passage in an abbreviated form (i.e., between 25 and 40% of the original length, according to the textbook) without over copying from the reading passage. The latter refers to an independent writing task in which the students are expected to write about their opinion using an appropriate paragraph structure while including as much supporting evidence as they can in a short time. The textbook indicates that when writing these paragraphs, students should use specific discourse markers to indicate evidence and supporting details for their main points and also use a wide variety of vocabulary.

After the first iteration of the curriculum in the 2020 academic year, I noticed that many teachers used the multiple-choice questions from the provided end of semester tests but did not use the writing questions. After an informal inquiry, teachers said that they did not use the writing questions because they had too many students and that it would take too long to grade all of their responses. I suggested an AI rating system, but many teachers responded that they could not trust AI raters because they presented a "black box" problem, i.e., they had no idea how the AI would rate the students and therefore were unsure that the AI would be rating students according to what they taught in their classes. However, if teachers did not ask their students to actually write and never evaluated student writing, I find it unlikely that students actually developed any writing ability.

1.2 The Proposed Solution: A Human-AI Integrated Rating Scheme

In order to remedy the problem of teachers not evaluating writing, I worked to create an integrated human-AI rating scheme. When doing so, I had to create it in such a way that teachers would trust the rating scheme and find it time-saving (so that they would try using it), but also had to make the scheme as trustworthy as possible in order to ensure fairness in grading. Therefore, I created the scheme based on the following premises:

1. The use of human-AI rating scheme should reduce the time needed for grading student writing responses.
2. Teachers should have control over the final scores to increase their trust in the

integrated rating scheme.

3. The AI model should be crated in-house based on data from students at the university it will be implemented at and aimed at skills that the students are specifically asked to learn to increase fairness in scoring and trust in the scheme.
4. The human-AI integrated rating scheme should increase rating fairness university-wide, i.e., they will be graded the same way on the same points by the same AI model, which should reduce the effect of teachers' human bias (e.g., Fang & Wang, 2011; Schneck & Daly, 2012).

1.3 Research Questions

Based on the aforementioned problems and the proposed solution, this study seeks to answer some very basic preliminary research questions related to implementing the human-AI integrated rating scheme at Tohoku University. Specifically, this paper reports on the creation of the scheme while answering the following questions:

1. Can a human-AI rating scheme be created and implemented for judging student writing in a very specific educational context?
2. What challenges are there when implementing a human-AI rating scheme?
3. How do students and teachers react to the implementation of a human-AI rating scheme?

2. Creating the Human-AI Integrated Rating Scheme

2.1 Determining which Aspects to Judge Via AI

In order to determine what aspects of writing the models should be based on, I first took summary ($N = 165$) and paragraph ($N = 136$) writing samples from students, with their permission to use the data for research purposes. I also asked students for their TOEFL ITP® scores, as this test is considered a gold-standard for EGAP, although the test does not contain an actual production section (it contains a structure and written expression section but uses multiple choice questions). I hired five professional writing raters to rate the students' writing and provided them with rubrics. The summary writing rubric was based on Li (2014) and Sawaki (2020) and included four sub-categories to be rated: (1) main idea coverage – i.e., the ratio of main ideas included in the summary, (2) integration – i.e., the logical order and global interpretability of the statements, (3) language use – i.e., the complexity and accuracy of the summary, and (4) source use – i.e., to what degree the summary is written correctly and in the writer's own words. The paragraph writing rubric was based on the Educational Testing Service independent writing task rubric for the TOEFL iBT® test, which contains four subcategories: (1)

content – i.e., how well the writing addresses the topic, (2) structure – i.e., how well the writing is organized, (3) coherence – i.e. how understandable the writing is, and (4) language – i.e., the variety and complexity of vocabulary and its usage (e.g., ETS, nd). The raters were asked to provide a score from one to five for each category and were subsequently asked which categories they felt were difficult to judge. In order to determine which areas of judgement were most problematic for human raters, I used the judges' responses about which areas they felt were difficult, but also checked for the amount of correlation between raters' scores using both Cronbach's alpha for inter-rater reliability across all raters, and simple one-to-one Pearson's correlation analyses to check for correlation between raters. A greater magnitude of correlation between rater scores would suggest trustworthiness in the scoring, so trends in the data were observed.

Table 1 shows the Cronbach's alpha and range of correlation magnitudes between rater scores for each category in the two writing tasks. According to the data, there seems to be a solid trend that the raters had much more agreement on concept-based rating, i.e., main idea and integration for summary writing, and content and structure for paragraph writing, than they did for language-usage-based rating, i.e., language use and source use for summary writing and coherence and language for paragraph writing. Furthermore, the raters themselves mentioned that it was difficult to judge language use, because it was difficult to know what could be considered complex or advanced, which made them have to re-read the responses several times. The raters also noted that it was difficult to judge source use for the summary writing task because they often forgot exactly what was written in the source text, and also had difficulty judging how much copying was 'too much.'

Table 1

Cronbach's Alpha and Range of Correlation Magnitudes for Rater Scores

Writing Task	Rubric Score	Cronbach's α	Range of Rater Correlation (*r*)
Summary Writing	Main Idea	.81	.54~.76
	Integration	.63	.34~.55
	Language Use	.45	-.11~.27
	Source Use	.31	-.09~.44
Paragraph Writing	Content	.96	.78~.83
	Structure	.91	.58~.71
	Coherence	.79	.32~.50
	Language	.75	.27~.40

Details of data set available in Appendix 1

The results from Table 1 and the raters' comments suggested that the areas that were most problematic for humans were the language-related domains, and that the content-related domains were much easier for them to judge accurately and quickly. Based on these findings, I endeavored to create two AI models for writing rating: one for summary writing that checks for language and source use, and one for paragraph writing that checks for language use. Human raters would then be left to only judge the content and structure of the responses, which the aforementioned data suggests that they can do much more accurately and readily. Furthermore, it should be noted that the raters mentioned that summary writing was much more difficult to judge, and the lower amounts of correlation in their scoring seem to match this notion.

Based on the results of Table 1, I decided to create an AI model that could judge language use and source use for summary writing and coherence and language for paragraph writing. Upon observing previous studies of AI essay-rating models, I discovered that most relied heavily on looking for keywords and n-grams (sequences of particular words) and their likelihood of appearing in a highly rated essay (e.g., Li, 2021). While this technique does result in high accuracy, it essentially attempts to check content and is therefore highly topic-specific. Furthermore, creating a similar model would also require thousands of previously graded essays. Since the writing questions on the tests at Tohoku University would change yearly and have no previous responses of the same topic on which to build a model, I needed more generalizable metrics. Therefore, I

decided to use CAF (complexity, accuracy, and fluency) metrics and genre-specific features that other studies have reported to be associated with proficiency (e.g., Lambert & Kormos, 2014; Lu, 2010; 2012; Kyle, 2016; Kyle & Crossley, 2017; 2018; Spring, 2023) and which also are aimed at measuring language use and coherency, specifically. Furthermore, I created my own model due to the suggestion that the way in which CAF measures are used in a second language varies greatly depending on the first language and levels of the learners (Lu & Ai, 2015), and the students at Tohoku University represent a homogenous first language population with a comparatively narrow range of EFL skill.

2.2 CAF Measures

A number of second language acquisition studies have pointed out that the complexity, accuracy, and fluency of second language learners tends to increase as they become more proficient in their target language (e.g., Lambert & Kormos, 2014; Ortega, 2003; Skehan, 2009; Wolfe-Quintero et al., 1998). In the past decade, a number of tools have become available to automatically calculate many of the CAF metrics that previous studies have indicated as indicative of second language writing proficiency and second language proficiency in general, e.g., the second language syntactic complexity analyzer (L2SCA; Lu, 2010), the lexical complexity analyzer (LCA; Lu, 2012), the tool for the automatic analysis of syntactic complexity (TAASC; Kyle, 2016), and the tool for the automatic analysis of lexical sophistication (TAALES; Kyle, et al., 2018). In order to create a single model that could both analyze various CAF measures and assign a score based on these metrics and previous data taken from Tohoku University, I created my own version of these tools using Python 3.9 and the SpaCy (Honnibal & Motani, 2017) "en_core_web_lg" pipeline for part of speech and dependency tagging, which can then be used to calculate the various CAF measures from the aforementioned tools[1]. These settings were used because they were found to produce CAF measures that showed the most correlation to general second language proficiency and human-rater scores of second language writing (Spring & Johnson, 2022). The selection of particular CAF measures for inclusion in the AI model are described below.

Complexity is the most heavily researched area of CAF measures with regards to writing. This is likely due to the fact that complexity is a multi-faceted aspect of writing, many measures can be automatically calculated with high precision, and many of the automatically calculated measures of complexity show significant correlation to both general second language proficiency and to second language writing scores (e.g., Jiang et al., 2019; Lu, 2010; 2012; Kyle, 2016; Kyle et al., 2018; 2021; Kyle &

Crossley, 2017; 2018; Spring & Johnson, 2022). First, there is a general division between lexical complexity, i.e., complexity at a word-unit level, and syntactic complexity, i.e., complexity at a grammatical or structural level. However, there are further distinctions, as measures of both lexical and syntactic complexity can include counts of "difficult" units, the frequency with which difficult units are used, and the variety of units that are used. Furthermore, there is another distinction between fine-grained and large-grained measures of complexity. In general, Lu's (2010; 2012) tools tend to look at larger-grained measures of complexity, such as type-token ratios (e.g., the number of different words divided by the total number of words), whereas Kyle's (2016) tools tend to also provide fine-grained measures (e.g., the number of prepositions that are the dependents of prepositional objects). Several studies have suggested that when making a model to predict rater scores of second language writing, combining several fine-grained measures can lead to a more accurate model than one that is comprised of several large-grained measures, although large-grained measures can often, individually, show stronger correlation to second language writing rating (e.g., Lu & Hu, 2021; Kyle & Crossley, 2017; 2018; Spring, 2023). Unfortunately, I was unaware of Kyle's tools in the first iteration of my human-AI integrated rating system, and thus the measures provided by Kyle's tools were not considered until the second iteration.

Accuracy is one of the less studied domains within CAF and automatically calculated measures are not used very much when creating models predictive of rater scores. One potential reason for this is that slight errors with accuracy often do not impede communication, and thus the number of total errors is not necessarily indicative of communicative ability (e.g., Tavakoli & Skehan, 2005; Wolfe-Quintero et al., 1998). Another potential reason is that learners often tend to make more errors when attempting to use new vocabulary and linguistic structures, and thus, accuracy often does not follow a straight upward path, but rather exhibits a curved u-shaped path, which would diminish correlation to rater-scoring or language proficiency (Vercellotti, 2017; Wolfe-Quintero et al., 1998). While some works have noted that counting the number of errors that impede communication, or the ratio of error-free language units to total language units can be indicative of learner proficiency (e.g., Robinson, 2001; Thai & Boers, 2016; Vercellotti, 2017), current software is generally unable to differentiate between errors that impact meaning and those that do not, so many automatically calculated measures of accuracy do not correlate to rater scores. After trying several different free online grammar accuracy checkers available in Python 3.9 with the two data sets presented in Table 1, I found that none of the measures or transformations were correlated with general English proficiency or rater scores, and thus did not consider them in my AI

model when creating the human-AI integrated rating system.

In the realm of second language writing, there is some argument as to what constitutes fluency, but some works (e.g., Lu, 2010; Wolfe-Quintero et al., 1998) consider the number of language units, i.e., words, clauses, t-units, sentences, etc., written in a timed-writing task to be indicative of written fluency. Since the writing tasks at Tohoku University are both timed, and several counts of the number of language units produced correlate highly with proficiency and rater scores (e.g., Lu, 2010; 2011; 2012; Kyle, 2016; Wolfe-Quintero et al., 1998), the various counts of language units provided by the L2SCA and TAASC tools were considered. As previously mentioned, the first iteration of the tool only considered those provided by the L2SCA due to my lack of awareness of the TAASC until the second iteration.

2.3 Genre and Context Specific Measures

Certain genre-specific considerations were also required for the Human-AI integrated rating systems at Tohoku University. Specifically, source writing, as defined by works such as Li (2014) and Sawaki (2020), and summary writing as defined by the curriculum at Tohoku University, requires that students do not over-copy from the source reading passage. Furthermore, the curriculum at Tohoku University requests that students use particular words and phrases to mark the evidence and supporting details for their main ideas to aid in coherence. Therefore, a metric for source-text copying and a metric for use of the supporting detail markers were created.

In order to create the metric for source-text copying, I first considered the *Pathways to Academic English*[1] textbook at Tohoku University which forbids five or more consecutive words to be copied directly from the source text. I then created a simple Python 3.9 script that would check for the number 2-grams, 3-grams, and 4-grams (i.e., two, three, and four consecutive words) that were copied directly from a source text[2]. I then used the tool to calculate the number of matched n-grams and the percentage of copied n-grams to total number of n-grams in the summary writings in my first data set (see Table 1). I then calculated the correlation to the professional raters' averaged source-use scores, and students TOEFL ITP® scores, the results of which are presented in Table 2. According to these results, the percentage of 3-grams copied from the source text exhibited the greatest magnitude of correlation to rater scores and none of the measures was significantly correlated to TOEFL ITP® scores, so the percentage of copied 3-grams was used as a metric of copying, along with the number of 5-grams, which were expressly forbidden by textbook.

Table 2

Correlation Between Source-Copying Metrics, Rater Scores, and TOEFL ITP® Scores

Metric	Correlation to Rater Scores	Correlation to TOEFL ITP®
copied 2-grams	-.24	.06
% of copied 2-grams	-.35	.06
copied 3-grams	-.39	.05
% of copied 3-grams	-.58	.03
copied 4-grams	-.43	.02
% of copied 4-grams	-.57	.01

In order to create the metric for evidence and supporting detail markers, I created a simple Python 3.9 script[2] that checks for the use of supporting detail markers that were given in the Tohoku University textbook. I also created a number of transformations based on the frequency of use per language unit and checked the correlation between these metrics and both rater scores and TOEFL ITP® scores for the first data set of paragraph writing (see Table 1). I found that a simple count of the supporting detail markers exhibited the greatest correlation to both rater and TOELF ITP® scores (Spring, 2023; results partially repeated in Table 3) and thus used the pure counts in the AI model.

Table 3

Correlation Between Supporting Detail Markers, Rater Scores, and TOEFL ITP® Scores

Metric	Correlation to Rater Scores	Correlation to TOEFL ITP®
number of markers	.28	.21
markers per sentence	.09	.17
markers per clause	.09	.10

Data repeated partially from Spring (2023)

2.4 Designing the Human-AI Integrated Rating Scheme

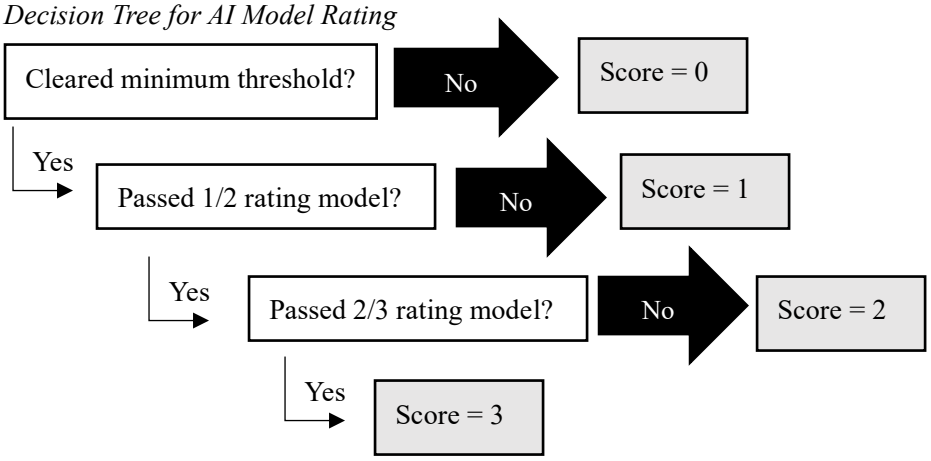
The first step in designing the Human-AI integrated rating scheme for the two writing assignments (summary writing and paragraph writing) was to determine the point layout of each. Because most students at Tohoku University belong to one of three CEFR[3] levels, I surmised that an AI model could be made to divide students on a three-point scale. Based on informal talks with colleagues at Tohoku University, teachers suggested a three-

point scale for main idea coverage based on the idea that most pieces of writing that students summarized contained three main point with several supporting details. Therefore, for summary writing, a six-point scale was adopted: three points would be determined by teachers' evaluation of main idea coverage, and three points would be determined by an AI model based on length, percentage of copied 3-grams, and a number of complexity measures. Teachers reported that for paragraph writing, they wanted to check for paragraph structure, adherence to the topic, and strength of the argument. Therefore, for paragraph writing, a five-point scale was adopted: two points would be determined by teachers' evaluation of paragraph structure and argument strength, three points would be determined by an AI model based on supporting detail markers and CAF measures, and teachers would be expected to overturn the AI score and assign a score of 0 if the paragraph was not written about the assigned topic. In the rating scheme, AI scores are provided first, and teachers are allowed to overturn AI scores if they feel them to be inappropriate. This allows for a final check and to assuage the fears of raters and students who might be distrustful of AI.

The AI models were created based on two premises. First, I did not assume that all metrics of writing ability would develop linearly. Therefore, I developed one model to distinguish between a score of one and a score of two and another to distinguish between a score of two and three. If a response passed the first model and received a score of two, it was then checked against the second model and in the event that it passed the second model as well, it received a score of three. Failure at the first model resulted in a score of one and failure at the second model resulted in a score of two. Furthermore, cut-offs were created which resulted in an automatic score of zero, which the students were made aware of. Specifically, a response of less than 50 words resulted in a score of zero for paragraph writing, and two or more instances of 5-grams copied directly from the source text resulted in a score of zero for summary writing. This process is visualized in Figure 1.

Figure 1

Decision Tree for AI Model Rating



Second, I did not think that any one metric should overly punish or reward responses. Therefore, I created a series of relative metric scores (RMS) that were used for rating. RMSs were created for each metric that was used in the final AI models based on the medians and standard deviations (SD) of previous data sets. Specifically, scores one SD above the median were given the maximum RMS of 3, scores one SD below the median were given the minimum RMS of 1, and all other scores were calculated as two plus the response metric minus the median divided by the SD (see formula below). This prevented students from trying to game the AI rating system by superficially improving just one metric, e.g., from achieving a score of 3 by erroneously increasing their word count with meaningless series of words.

Formula for Relative Metric Scores within +/- One Standard Deviation of the Median

$$\text{RMS} = 2 + (\text{User Metric Score} - \left(\frac{\text{Metric Score Median}}{Metric\ SD}\right))$$

In order to create the AI models, I first used average rater scores to classify writing samples as worthy of a score of one, two, or three. Writing samples that did not meet the minimum requirements and received a score of zero were not considered, as they were considered outside of the rules. First, the model to distinguish between a score of one and two was created by observing the raw correlation between each automatically calculated measure described in sections 2.2 and 2.3 and averaged rater score (i.e., one or two), as well as between each measure and general English proficiency (i.e., TOEFL ITP® scores). All measures that were correlated at a threshold of $r >= 0.2$ were

considered for the model. Next, a stepwise model was created by removing all automatically calculated measures that did not exhibit homoscedasticity or had a correlation of $r >= 0.7$ with other measures. When two measures exhibited such multicollinearity, the one with the greater magnitude of correlation to rater scores was kept, and the other was eliminated, following Kyle and Crossley (2018). Then a logistic regression analysis with dominance analysis refactored as relative weight was conducted, following Mizumoto (2023), to determine the weight each measure should carry in the model. In the final analysis conducted by the AI rater, each RMS was multiplied by the relative weight as suggested by the regression analysis, these scores were summed, and then a cutoff point for rejection was determined by finding the cutoff point at which the maximum number of writing samples would be correctly categorized. The same process was carried out for the model that distinguished between a score of two and three.

The first iteration of both the summary-writing and paragraph-writing Human-AI integrated rating schemes were based on the initially taken data (see Table 1), but then modified based on new data after implementation in the grading of students' final exams. Specifically, several students agreed to allow the writing samples from their final exams to be used for research purposes, and these were used to adjust the AI-rating models for the following iterations. Five professional raters were asked to rate the writing from the final exams after the semester had ended, and the same basic procedures as above were taken to create a new model. It should be noted that after the first iteration, I became aware of Kyle's tools, and several measures from the TAASC program were considered for later iterations of the AI-rating model, as well as a separate phrasal complexity measure (i.e., the number of satellite-framed expressions) based on an early version of the Event Conflation Finder (Spring & Ono, 2023). After each iteration, the initial data set, as well as the writing samples from all exams up to that point were both considered, and only variables that showed steady correlation across all data sets were considered. Cutoffs for rejection in each model were created based on those which would provide the highest number of correct scores for all data sets. Furthermore, I informally canvassed teachers for their ideas for improvement and attempted to implement as many as possible to increase the number of teachers willing to use the writing questions in their final exams.

The exact formulas that were used for the two AI models, i.e., the final relative weights for the two decisions models and the values for the medians and standard deviations on which the RMSs were calculated, can be found in the GitHub repository[2], in the rater_s (for summary writing rating) and rater_p (for paragraph writing rating) subdirectories.

3. Using the Human-AI Rating Scheme

3.1 First Implementation – Paragraph Writing

The first iteration of the human-AI integrated rating scheme took place in the fall of 2021 and was used to rate paragraph writing by students on their final exam. Three teachers participated and were given a short survey asking whether or not the human-AI rating scheme saved them time and their confidence in their scores. In order to determine the accuracy of the human-AI rating scheme, the correlation between the AI-only score and the human-AI rating scores were checked against students' TOEFL ITP® scores and the average scores of five professional human raters, who later rated the essays on a scale of one to five. The results of these analyses, as well as the number of scores overturned by each teacher are summarized in Table 4. The results indicate that the AI rating model was highly correlated with both TOEFL ITP® scores and professional rater scores. Furthermore, the scores from the human-AI integrated rating scheme were correlated similarly to TOEFL ITP® scores but slightly less to professional human rater scores, but only when the raters trusted the AI rater. Specifically, teacher B overturned several scores, resulting in the final human-AI rating scheme scores to be far less correlated to both TOEFL ITP® scores and professional human rater scores. Interestingly, the less confidence the teachers had in their own ability to rate students' writing, the more positively their scores contributed to accuracy.

Table 4

Results of the First Iteration of the Human-AI Integrated Rating Scheme (Paragraph Writing)

Teacher (*N*)	Saved Time?	Confidence? (1-10)	Overturned Scores (%)	AI / TOEFL	Human-AI TOEFL	AI / PR (5)	Human-AI / PR (5)
A (79)	Yes	3	1 (1%)	.26*	.31**	.49**	.39**
B (120)	No	10	54 (45%)	.16*	.01	.67**	.09
C (40)	Yes	6	2 (5%)	.43**	.30**	.61**	.47**
Total (239)			57 (24%)	.26**	.09	.69**	.24**

**p < .05, **p < .01; part of this data is repeated from Spring (2022)*

After the first iteration, informal canvassing of teachers and students revealed that both parties were worried about the AI rater and not understanding or clearly being able

to see how it would rate various responses. In order to remedy this issue, a simple web-based tool was developed in HTML and JavaScript to mimic the over-copying and word count rating for summary writing[4], which was the writing type of the second iteration. These two features were selected because there were relatively easy to recreate with high accuracy in JavaScript, and they represented a significant portion of the AI-rating models for summary writing. The web-based tool was provided to teachers and students for practice for the final exam in iteration two. Similarly, a web-based tool was created for students and teachers to use during the third iteration that recreated some of the highly representative measures for the paragraph writing task[4]. Specifically, word count, corrected type-token ratio (CTTR; see Lu, 2012 and Spring & Johnson, 2022), counts of supporting detail markers (see Spring, 2023), and mean length of sentence could be calculated and displayed graphically along with benchmarks for students, set at one standard deviation above and below the median scores from previous data sets. Students were allowed to practice with these tools, teachers were encouraged to use them, and both were informed clearly that the AI rating model would largely draw from the representative measures displayed by the online tools.

3.2 Second Implementation – Summary Writing

The second iteration of the human-AI integrated rating scheme took place in the spring of 2022 and was used to rate summary writing by students on their final exam. Four teachers participated, two of whom also participated in the first iteration. A similar survey was given to teachers after using the scheme, and once again, correlation of both AI-rating and human-AI integrated rating was conducted against both TOEFL ITP® scores and the average scores of three professional human raters[3]. The results suggest that the AI-rating system worked extremely well and correlated more highly to professional rater scores than in the first iteration. Furthermore, the human-AI rating system exhibited greater correlation to target scores (i.e., professional rater scores and TOEFL ITP® scores) than the AI-rater alone. Furthermore, most teachers thought that the human-AI rating scheme saved them time in scoring as compared to rating alone. These results are summarized in Table 5.

Table 5

Results of the Second Iteration of the Human-AI Integrated Rating Scheme (Summary Writing)

Teacher (*N*)	Saved Time?	Overturned Scores (%)	AI / TOEFL	Human-AI / TOEFL	AI / PR (5)	Human-AI / PR (5)
A (127)	Yes	0 (0%)	.25**	.28**	.47**	.67**
C (251)	Yes	0 (0%)	.21**	.22**	.87**	.89**
D (84)	Yes	4 (5%)	.32**	.33**	N/A	N/A
E (160)	Neutral	10 (6%)	.24**	.29**	.82**	.87**
Total (622)		14 (2%)	.22**	.27**	.85**	.86**

**p < .05, **p < .01*

3.3 Third Implementation – Paragraph Writing

The third iteration of the human-AI integrated rating scheme took place in the fall of 2022 and was used to rate paragraph writing on students' final exams. Changes from the first iteration include a recalibration of the AI rating model as described in section 2.4 and the introduction of the online feedback tool described above. Seven teachers participated in the third iteration, three of whom returned from previous iterations, a similar survey was conducted afterwards, and the same correlation analyses as described above were conducted once more. The results showed that the accuracy of the AI model greatly increased and that most teachers improved the magnitude of correlation to target scores by adding their scores to the AI model. Furthermore, the correlation university-wide was greatly improved from the first iteration. The results of this iteration are summarized in Table 6.

Table 6

Results of the Third Iteration of the Human-AI Integrated Rating Scheme (Paragraph Writing)

Teacher (*N*)	Saved Time?	Confidence (1-10)	Overturned Scores (%)	AI / TOEFL	Human-AI / TOEFL	AI / PR (5)	Human-AI / PR (5)
A (117)	Yes	2	0 (0%)	.42**	.35**	.49**	.54**
C (157)	Yes	3	0 (0%)	.47**	.55**	.69**	.63**
D (84)	Yes	6	3 (4%)	.52**	.49**	.67**	.64**
F (115)	Yes	7	0 (0%)	.12*	.23**	.64**	.75**
G (41)	Yes	7	7 (17%)	.39**	.47**	.58**	.69**
H (122)	No	6	122 (100%)	.48**	.54**	.52**	.52**
I (84)	Yes	N/A	0 (0%)	.53**	.47**	.59**	.73**
Total (720)			132 (0%)	.36**	.32**	.57**	.48**

**p < .05, **p < .01*

3.4 Summative Impact on the Curriculum

Overall, the human-AI integrated rating system seems to have had the intended impact on the curriculum that it was designed to have. Specifically, it increased the number of teachers who were willing to provide writing questions on the end of semester tests and evaluate student writing, which resulted in more students being made to actually write, which most would argue is a prerequisite for acquiring writing skill. Generally, most teachers also felt that the system saved them time, and many decided to use the human-AI integrated rating scheme again after trying it once. Furthermore, both students and teachers grew to trust the AI ratings, especially once the online practice tools were made available, which made the grading system clearer and provided students with goal-centered practice and feedback. Finally, the rating accuracy of the human-AI integrated system improved over time and provided a common source of grading across the classes, which theoretically improved the curriculum-wide (i.e., intra-class) fairness of the writing scoring. The impact is summarized in Table 7.

Table 7

Summative Impact of the Human-AI Writing Scoring System on Writing over Time

Iteration	No. Teachers	No. Students	No. Teachers whose Time was Saved (%)	Correlation between AI-human and Pro Raters
1	3	239	2 (67%)	.24**
2	4	622	3 (75%)	.86**
3	7	720	6 (86%)	.48**

***p < .01; iteration numbers 1 and 3 were for paragraph writing, iteration 2 was for summary writing*

4. Discussion

Despite the recent advances in language research due to the use of Large Language Models (LLMs), there are still many stakeholders who are still skeptical of AI scoring in EFL education, i.e., teachers and students. However, as this study shows, the solution may be to introduce human-AI integrated models that are built on easy-to-understand and theoretically sound metrics that students can practice with and receive clear feedback on. By creating such a system, students were able to understand what targets they were expected to reach in their writing and could practice with the online tools and ensure that they were reaching them. Teachers could also clearly see how their students were performing, but more importantly, were left with the time and energy in the classroom to focus on features of writing that AI does not rate or provide feedback on as easily: namely structure, content, and style. Therefore, a more collaborative system that allows for more teacher freedom and clarity in the integration process might be a good way to integrate both the human element provided by teachers and the latest advances in technology provided by AI.

While this project proved somewhat successful, there are a number of areas that require future study, observation, and improvement in the future. First, the tools that provide the metrics that the AI is based on are constantly improving and future iterations should reflect these advances. For example, Crossley et al. (2019) have reported meaningful textual cohesion measures which should be explored through the Tool for the Automatic Analysis of Cohesion (TAACO), and studies such as Eguchi (2023) have shown that AI can also detect more meaning-based aspects of text, such as writer stance.

Second, as large language models such as Chat GPT 4.0 become increasingly human-like in their responses and as their neural networks develop to contain fewer hallucinations, the prospect of using such models for grading should be observed. In fact, some studies such as Mizumoto and Eguchi (2023) have already begun to show that Chat

GPT 4.0 has some capability to rate EFL student writing similarly to humans. While this presents a black-box problem, teachers may begin to trust AI more readily as advances are made, and there may be opportunities to integrate a more wholistic rating provided by Chat GPT 4.0 (i.e., Mizumoto & Eguchi, 2023) with measures such as those presented in this study to create a new AI-AI integrated system that will be more transparent to learners than a completely black-box model (i.e., having just Chat GPT 4.0 rate responses).

Finally, more work needs to be done to convince more teachers to try AI solutions, including, but not limited to, AI-human integrated solutions, such as the one suggested in this paper. While the results of this action research show that the system has attracted an increasing number of teachers to use the system, it should be noted that some teachers, albeit a minority, did not decide to use the AI-human integrated rating scheme again, saying that they did not trust the AI scores. However, as the data in the previous section suggests, such teachers might be overconfident about their ability to accurately assess student writing or may have a completely different standard from their colleagues. Either way, it should also be noted that only about 20% of teachers at Tohoku university were willing to even try the AI-human rating scheme, which speaks to the fact that much more work needs to be done to convince teachers that such solutions are valid and worth the effort. If more is not done in this area, students will, unfortunately, continue to miss opportunities to write and have their writing evaluated and receive feedback.

Acknowledgements

This paper was funded in part by a grant by the Japan Society for the Promotion Sciences (grant number 22K00810). It represents the culmination of several works, which are referenced throughout the paper, but also represents original work and data that has not been previously published. Approval for this study was granted by the Internal Review Board of the author's university and all participants gave informed consent to participate and for the results to be published. Accordingly, numerical data is available upon request, but specific responses are not.

Notes

1. There are several editions to the Pathways to Academic English series, used at Tohoku University since 2020. This study began under the 3rd edition (Spring et al., 2022) and continued into the 4th edition (Spring & Scura, 2023).
2. The code for all of the tools can be found at https://github.com/mwjohnson/autograder. An anonymous reviewer for another

paper pointed out that the SpaCy trf model is slightly more accurate than the web_lg model, but no significant changes were found in correlations between human raters and calculated measures by switching to the trf model, so I kept the web_lg model for efficiency. The supporting detail marker counter is called as a class and the source-checking script is contained in a separate folder. The raters for summary writing and paragraph writing are kept in separate folders but call the 'spacy_full.py' file and the appropriate classes and subscripts in order to produce the measurements required for the AI models.

3. The Common European Framework of Reference for Language: a commonly used scale to contextualize foreign language learners.
4. The code for both the summary writing and paragraph writing feedback generators for students and teachers can be found at https://github.com/springuistics, specifically in the "online_summary_checker" and "paragraph_feedback" projects.

References

Crossley, S. A., Kyle, K., & McNamara, D. S. (2016). The tool for the automatic analysis of text cohesion (TAACO): Automatic assessment of local, global, and text cohesion. *Behavior Research Methods, 48*(4), 1227–1237. https://doi.org/10.3758/s13428-015-0651-7

Eguchi, M. (2023, March 18-21). *Towards the automatic analysis of rhetorical strategies: Development and evaluation of a stance-taking analyzer* [Conference presentation]. AAAL 2023 Conference, Portland, OR, United States. https://www.xcdsystem.com/aaal/program/T3QFbEa/index.cfm?pgid=220

Fang, Z., & Wang, Z. (2011). Beyond rubrics: Using functional language analysis to evaluate student writing. *Australian Journal of Language and Literacy, 34*, 147–165.

Honnibal, M., & Montani, I. (2017) SpaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. https://spacy.io/

Jiang, J., Bi, P., Liu, H. (2019). Syntactic complexity development in the writings of EFL learners: Insights from a dependency syntactically-annotated corpus. Journal of Second Language Writing, 46, 1–13. https://doi.org/10.1016/j.jslw.2019.100666

Kyle, K. (2016). Measuring syntactic development in L2 writing: Fine grained indices of syntactic complexity and usage-based indices of syntactic sophistication (Unpublished doctoral dissertation). Georgia State University, Atlanta, GA.

Kyle, K., & Crossley, S. A. (2017). Assessing syntactic sophistication in L2 writing: A usage-based approach. *Language Testing, 34*(4), 513–535. https://doi.org/10.1177/0265532217712554

Kyle, K., & Crossley, S. A. (2018). Measuring syntactic complexity in L2 writing using fine-grained clausal and phrasal indices. *The Modern Language Journal, 102*(2), 333–349. https://doi.org/10.1111/modl.12468

Kyle, K., Crossley, S. A., & Berger, C. (2018). The tool for the automatic analysis of lexical sophistication (TAALES): Version 2.0. *Behavior Research Methods, 50,* 1030–1046. https://doi.org/10.3758/s13428-017-0924-4

Kyle, K., Crossley, S. A., & Verspoor, M. (2021). Measuring longitudinal writing development using indices of syntactic complexity and sophistication. *Studies in Second Language Acquisition, 43*(4), 781–812. https://doi.org/10.1017/S0272263120000546

Lambert, C., & Kormos, J. (2014). Complexity, accuracy, and fluency in task-based L2 research: Toward more developmentally based measures of second language acquisition. *Applied Linguistics, 35*, 607–614. https://doi.org/10.1016/j.system.2004.01.001

Li, J. (2014). The role of reading and writing in summarization as an integrated task. *Language Testing in Asia*, *4*(3). https://doi.org/10.1186/2229-0443-4-3

Li, M. (2021). *Researching and teaching second language writing in the digital age.* Springer Nature. https://doi.org/10.1007/978-3-030-87710-1_7

Lu, X. (2010). Automatic analysis of syntactic complexity in second language writing. *International Journal of Corpus Linguistics, 15*(4), 474–496. https://doi.org/10.1075/ijcl.15.4.02lu

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. TESOL Quarterly, 45(1), 36–62.

Lu, X. (2012). The relationship of lexical richness to the quality of ESL learners' oral narratives. *The Modern Language Journal, 96*(2), 190–208. https://doi.org/10.1111/j.1540-4781.2011.01232_1.x

Lu, X., & Ai, H. (2015). Syntactic complexity in college-level English writing: Differences among writers with diverse L1 backgrounds. *Journal of Second Language Writing, 29*, 16–27. https://doi.org/10.1016/j.jslw.2015.06.003

Lu, X., & Hu, R. (2021). Sense-aware lexical sophistication indices and their relationship to second language writing quality. *Behavior Research Methods*. https://doi.org/10.3758/s13428-021-01675-6

Mizumoto, A. (2023). Calculating the relative importance of multiple regression predictor variables using dominance analysis and random forests. *Language Learning, 73*(1), 161–196. https://doi.org/10.1111/lang.12518

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2*(2), 100050. https://doi.org/10.1016/j.rmal.2023.100050

Ortega, L. (2003). Syntactic complexity measures and their relationship to L2 proficiency: A research synthesis of college-level L2 writing. *Applied Linguistics, 24*(4), 492–518. https://doi.org/10.1093/applin/24.4.492

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a componential framework. *Applied Linguistics, 22,* 27–57. https://doi.org/10.1093/applin/22.1.27

Sawaki, Y. (2020). Developing Summary Content Scoring Criteria for University L2 Writing Instruction in Japan. In G.J. Ockey & B.A. Green (Eds.), *Another Generation of Fundamental Considerations in Language Assessment* (pp. 153–171). Springer. https://doi.org/10.1007/978-981-15-8952-2_10

Schenk, A. D., & Daly, E. (2012). Building a better mousetrap: Replacing subjective writing rubrics with more empirically-sound alternatives for EFL learners. *Creative Education, 3*(8), 1320–1325. http://dx.doi.org/10.4236/ce.2012.38193

Skehan, P. (2009). Modelling second language performance: Integrating complexity, accuracy, fluency, and lexis. *Applied Linguistics, 30*(4), 510–532. https://doi.org/10.1093/applin/amp047

Spring, R., & Johnson, M. W. (2022). The possibility of improving automated calculation of measures of lexical richness for EFL writing: A comparison of the LCA, NLTK, and SpaCy tools. *System, 106*, 770–786. https://doi.org/10.1016/j.system.2022.102770

Spring, R., & Ono, N. (2023). Creating an automated tool to assist with event-conflation studies: An explanation and argument for its importance. *Research Methods in Applied Linguistics, in Press.* https://doi.org/10.1016/j.rmal.2023.100054

Spring, R. (2022). A pilot study of an integrated AI-human writing-rater system: How I learned to stop worrying and love the machine. *The 23rd Annual International Conference of the Japanese Society for Language Sciences (JSLS) Handbook* (pp. 130-134).

Spring, R. (2023). Transformations of number of words and phrases signaling supporting details: Potential variables for automated rating. *Language Education & Technology, 60*

Tavakoli, P., & Skehan, P. (2005). Strategic planning, task structure, and performance testing. In R. Ellis (Ed.), *Planning and Task Performance in a Second Language* (pp.239–273). John Benjamins. https:/doi.org/10.1075/lllt.11.15tav

Thai, C., & Boers, F. (2016). Repeating a monologue under increasing time pressure: Effects of fluency, complex, and accuracy. *TESOL Quarterly, 50*(2), 369–393. https:/doi.org/10.1002/tesq.232

Vercellotti, M.L. (2017). The development of complexity, accuracy, and fluency in second language performance: A longitudinal study. *Applied Linguistics, 38*(1), 90–111.

https://doi.org/10.1093/applin/amv002

Wolfe-Quintero, K., Inagaki, S., & Kim, H. Y. (1998). *Second language development in writing: Measures of fluency, accuracy and complexity* (Report No. 17). Honolulu: University of Hawai'i, Second Language Teaching and Curriculum Center. https://doi.org/10.1017/s0272263101263050