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

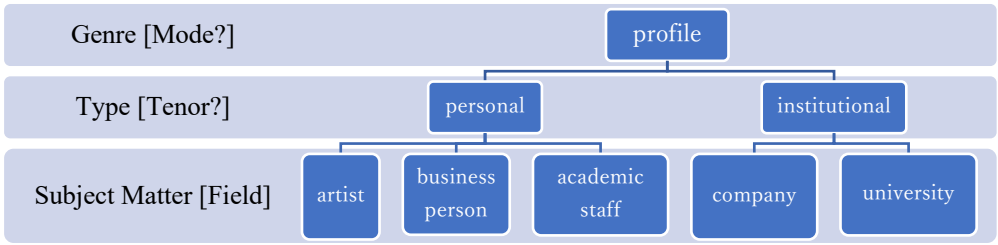
communication (Malinowski, 1923). This credo has been passed on, for example, to language studies by J.R. Firth and to SFL by M.A.K. Halliday. In the SFL framework, language takes on a higher order semiotic system, including context, semantics, lexico-grammar, expression and others in the layered system (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)

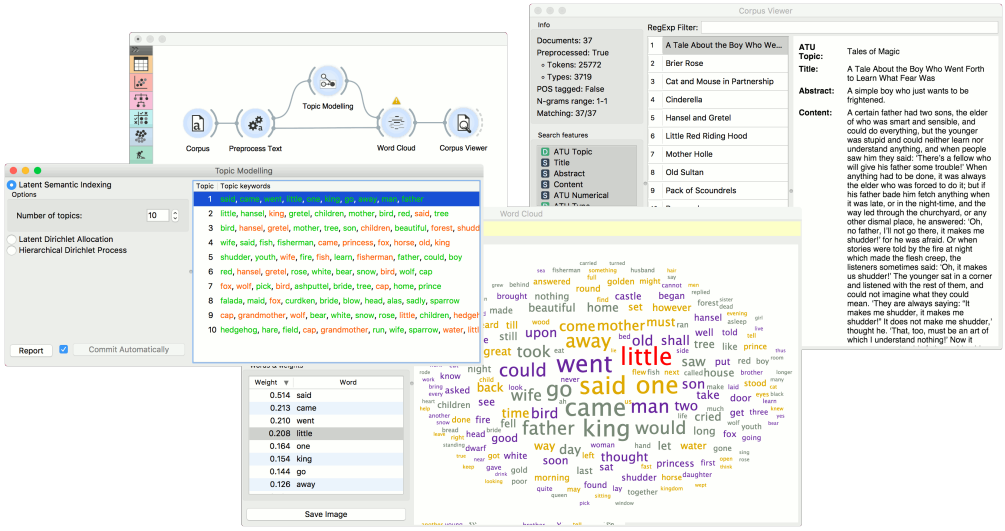
LSA[4] is an NLP technique in distributional semantics used for text summarisation and classification. It enables the detection of the underlying semantics of words in multiple texts by constructing topics related to words and texts. As detailed in Landauer et al. (2014) and summarised in Nishina (2023), 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.