European Journal of Computer Science and Information Technology
Vol.10, No.4, pp.52-62, 2022
Print ISSN: 2054-0957 (Print),
Online ISSN: 2054-0965 (Online)

Text genres, readability and readers' comprehensibility

Wenchao Li
Zhejiang University

Citation: Wenchao Li (2022) Text genres, readability and readers' comprehensibility, *European Journal of Computer Science and Information Technology*, Vol.10, No.4, pp.52-62

ABSTRACT: *This study applies mathematical linguistics to explore the association between text genres, readability and readers' comprehensibility. The target readers are students from Korea, Thailand and China, studying in Japanese universities, pursuing 20 different disciplines. Japanese is their third language, and all students have passed the Japanese-Language Proficiency Test Level 1. The textbook's readability and readers' comprehensibility are measured at two levels using two metrics. Mean dependency distance (MDD) is employed for measuring syntactic diversity; moving-average morphological richness (MAMR) and moving-average mean size of paradigm (MAMSP) are calculated for measuring lexicon diversity. The findings indicate that in terms of lexical diversity, textbooks of humanities seem simpler than natural science and engineering. In syntactic complexity, the textbook of informatics shows the simplest structure while that of social welfare presents the highest. Text genres relate to a textbook's readability and eventually influence readers' comprehensibility. Moreover, lexical diversity is not corelated to syntactic complexity*.

KEYWORDS: lecture genres, reading comprehensibility, Japanese lecture, lexical diversity, syntactic complexity

INTRODUCTION

In academic lectures, the role of textbooks is to offer primary knowledge. The readability of textbooks refers to the text itself and does not vary based on readers (Chiang et al. 2008; Anderson 2013; Bargate 2012)[1]. Comprehensibility indicates readers' ability to understand the meaning of the text (Davidson 2005). Therefore, readability and comprehensibility orient distinctly. Incorporating this, readability can be measured by syntactic complexity and lexical diversity. Comprehensibility is linked to several factors, e.g. readers' interest, background, accessibility to prior knowledge, ability to read and working memory (Kendeou et al. 2016).

[1] Text readability also concerns the text's presentation, e.g. font size, line length, etc. However, this study does not consider these factors but focuses on syntactic and lexical complexity of the text.

This paper deals with readability and investigates whether textbooks of different disciplines present distinct readabilities, which in turn could lead to education efficiency. The outcome of the study may help with the selection of textbooks in terms of disciplines and help improve education efficiency. In this article, Section 2 addresses previous studies on the measures of readability of textbooks. Section 3 outlines the methodology (including the framework, corpora, syntactic parser, and MDD calculation), Section 4 addresses results and discussions, and Section 5 presents the conclusion.

Previous studies

Readability has been argued to be associated with vocabulary size (Grabe 2009; Hu and Nation 2000; Stæhr 2008). Vocabulary is particularly import in academic text reading. Several attempts have been made to measure readability, i.e. (a) eye movements, (b) word difficulty, (c) semantic richness (Lee et al. 2021), (d) N-gram analysis (Xia et al. 2016) and (e) cognitive-motivated features (Feng et al. 2009). Readability has been measured through a variety of metrics such as word length, sentence length. Liu (2008) departs from the perspective of quantitative linguistics, computing and calculating language comprehension difficulty via dependency distance. Inspired by Liu (2008), this study applies mathematical linguistics to explore the association between the 20 genres of lectures, their readability and comprehensibility.

DATA AND METHODS

Data

The data were textbooks of 20 disciplines, drawn from the Reading Corpus of Non-native Japanese Language Learners. The target readers are students (Chinese, Korean and Thai) at Japanese universities pursuing different disciplines, with Japanese as the third language. They have passed Japanese-Language Proficiency Test Level 1, meaning their Japanese reading comprehension is advanced. The textbooks include humanities, natural science and engineering, i.e. sociology, education, literature, philosophy, human geography, cognitive science, politics, tourism, Japanese studies, linguistics, social welfare, business studies, material engineering, life science, science technology, civil engineering, informatics, commercial science. All textbooks are academically written, i.e. in plain form. Detailed information of the study data is provided in Table 1.

Table 1. Study data

Number	Lecture	Total words	Genres	Writing style
R-01	sociology	8901	Academic	Plain form
R-02	education	6710	Academic	Plain form
R-03	philosophy	5790	Academic	Plain form
R-04	literature	9802	Academic	Plain form
R-05	human geography	8956	Academic	Plain form
R-06	cognitive science	5701	Academic	Plain form
R-07	politics	8990	Academic	Plain form
R-08	tourism	7804	Academic	Plain form
R-09	Japanese studies	7002	Academic	Plain form
R-10	linguistics	6043	Academic	Plain form
R-11	social welfare	7432	Academic	Plain form
R-12	music	6572	Academic	Plain form
R-13	business studies	4997	Academic	Plain form
R-14	economics	5908	Academic	Plain form
R-15	material engineering	4003	Academic	Plain form
R-16	life science	5979	Academic	Plain form
R-17	science technology	4204	Academic	Plain form
R-18	civil engineering	5230	Academic	Plain form
R-19	informatics	6602	Academic	Plain form
R-20	commercial science	6988	Academic	Plain form

Procedure

Syntactic complexity and lexical diversity of 20 types of textbooks written in academic Japanese (i.e. the plain form) were analysed to examine whether the genres of lectures are linked to readability, and whether they influence readers' comprehensibility. The following procedures are carried out:

Step 1: Draw raw data from the reading corpora

Step 2: Analyse the syntactic complexity and lexical diversity of textbooks of each genre

Step 3: Analyse the syntactic complexity and lexical diversity of students' repeat of textbook content.

Analysis

Lexicon diversity and syntactic complexity are employed as metrics of the textbooks and students' reveal. Specifically, mean dependency distance (MDD) is employed for measuring

syntactic diversity. Moving-average morphological richness (MAMR) and moving-average mean size of paradigm (MAMSP) are calculated for measuring lexicon diversity. The MDD, MAMR, MAMSP are computed using self-written computer programme scripts.

Syntactic complexity

Dependency distance is a concept under the framework Dependency Grammar (Tesnière 1959; Yngve 1960; Hudson 2007; Liu 2009b), referring to the distance between the governor and the dependent. The governor refers to the core linguistic element in a sentence, such as verb, predicate. The dependent refers to the subject, object, oblique, adverb, post/prepositional phrase, etc. Liu, Hudson, and Feng (2009) propose measuring the dependency distance (DD) via |governor − dependent|. The MDD of the whole sentence would be

$$\mathrm{MDD} = \frac{1}{\mathrm{n}}\sum_{i=1}^{n} |\mathrm{DD}_i|$$

Lexical diversity

Moving-average morphology richness (Cech and Kubat 2018) are employed as the metric. It is suggested by Covington and McFall (2010), Yan and Liu (2021), Li, Liu and Li (2022) that, using a moving window can obtain a better average type-token ratio (TTR). The moving window of TTR in terms of word form can be obtained by the following formula:

$$MATTR\ \ (W)_{word\ \ form} = \frac{\sum_{I=1}^{N-W+1} F_i}{\mathrm{W}\ \ (\mathrm{N}-\mathrm{W}+1)}$$

The moving window of TTR in terms of lemma can be obtained in the same way, i.e.

$$MATTR\ \ (W)_{lema} = \frac{\sum_{I=1}^{N-W+1} F_i}{\mathrm{W}\ \ (\mathrm{N}-\mathrm{W}+1)}$$

Lexical diversity is obtained by $\frac{\sum_{I=1}^{N-W+1} F_i}{\mathrm{W}\ \ (\mathrm{N}-\mathrm{W}+1)} - \frac{\sum_{I=1}^{N-W+1} F_i}{\mathrm{W}\ \ (\mathrm{N}-\mathrm{W}+1)}$. Essentially, the higher the MAMR, the greater the lexical sophistication.

RESULTS AND DISCUSSION

With the methodology highlighted above, this section proceeds to an assessment of readability of the 20 genres of textbooks, assessing lexical diversity and syntactic complexity. Section 4.1 addresses text genres and their readability at the lexical level (measure via MAMR and MAMSP). Section 4.2 addresses text genres and their readability at the syntactic level (measure via MDD). Section 4.3 carries out a Euclidian distance clustering of text genres and their readability, examining the association of genres, readability and readers' comprehensibility.

Text genres and readability at lexical level (by MAMR and MAMSP)

The readability of textbooks of diverse genres at the lexical level is provided in Table 2. The linguistics textbook is found to bear the least lexical diversity, while commercial science is the most lexically diverse. It appears that textbooks of humanities, such as life science, literature, business studies, music, informatics, tourism, cognitive science, education, sociology, social welfare, philosophy, human geography and economics, bear relatively less MAMR than textbooks of natural science and engineering, e.g. material engineering, civil engineering, science and technology, commercial sciences.

Table 2. Text genres and readability measured by lexical diversity

Textbook genres	MAMR	MAMSP
linguistics	0.0027	1.0039
life science	0.0031	1.0038
literature	0.0056	1.0076
business studies	0.0059	1.0082
music	0.0061	1.0081
informatics	0.0079	1.0107
tourism	0.0083	1.0114
cognitive science	0.0104	1.0149
education	0.0106	1.0142
sociology	0.0108	1.0151
social welfare	0.0131	1.0197
philosophy	0.0137	1.0198
human geography	0.0137	1.0199
economics	0.0146	1.0201
material engineering	0.0158	1.0207
civil engineering	0.0165	1.0233
Japanese studies	0.0178	1.0251
politics	0.0184	1.0238
science technology	0.0212	1.0311
commercial science	0.0345	1.0477

Text genres and readability at syntactic level (by MDD)

Mean dependency distance of textbooks is measured to observe their syntactic complexity. The finding reveals that the textbook of informatics presents the simplest structure complexity, while social welfare bears the highest.

Table 3. Text genres and readability measured by syntactic complexity

Genres	**MDD**
informatics	3.0276
politics	3.2433
cognitive science	3.2927
Japanese studies	3.4588
tourism	3.5177
science technology	3.5856
sociology	3.6537
material engineering	3.6758
human geography	3.7754
literature	3.8022
music	3.8645
life science	3.9242
commercial science	4.0000
philosophy	4.0851
economics	4.2534
linguistics	4.4358
education	4.7556
business studies	4.9052
civil engineering	5.1166
social welfare	5.2519

A Euclidian clustering analysis was carried out and Figure 1 presents the outcomes.

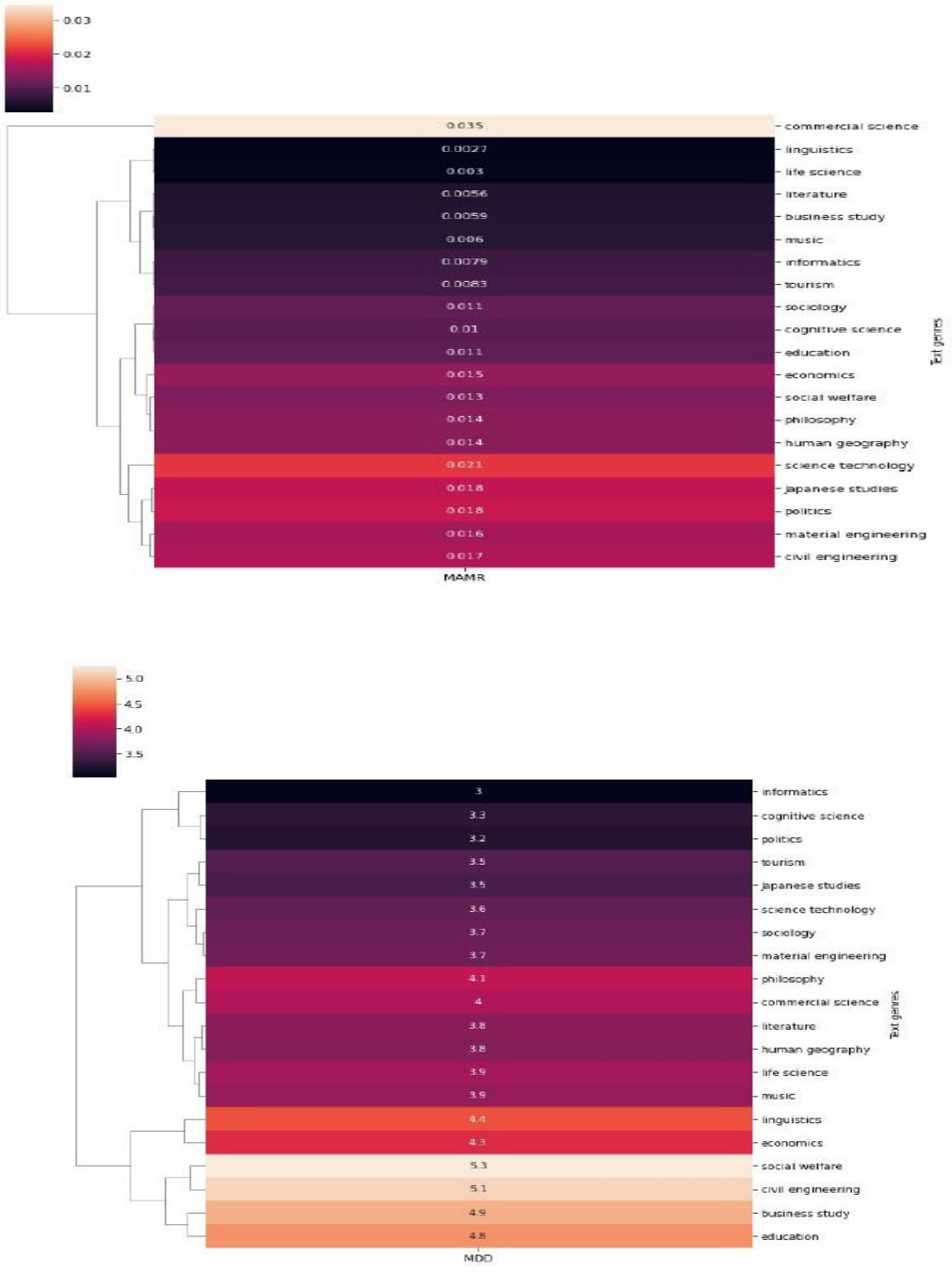

Figure 1. Euclidian distance clustering of textbooks (via MAMR (lexical) and MDD (syntactic))

Given the MDD and clustering via Euclidian distance, the 20 genres of textbooks were classified into the following three groups:

(a) the short group, with MDD ranging from 3.0276 to 3.9242 — which included disciplines
such as informatics, politics, cognitive science, tourism, science technology, Japanese studies, sociology, material engineering, human geography, literature, music, life science.
(b) the middle group, with MDD ranging from 4.0000 to 4.9052 — which included disciplines
such as commercial science, philosophy, economics, education, and business studies.
(c) the long group, with MDD ranging from 5.1155 to 5. 2519 — which included genres such as civil engineering and social welfare.
Further confirmed by corelation analysis (Figure 2), lexical diversity and syntactic complexity do not seem to be corelated.

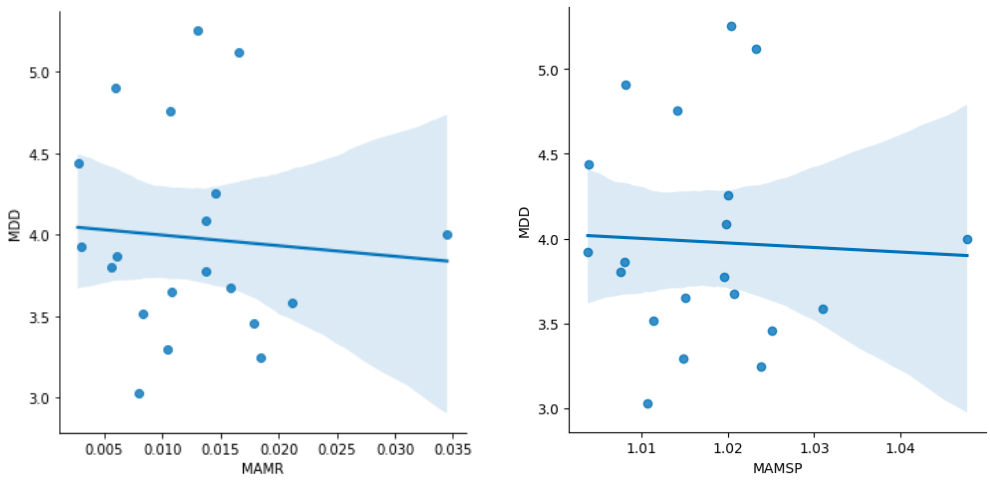

Figure 2. Corelation analysis to lexical diversity and syntactic complexity

Text readability and readers' comprehensibility

The previous section presented a picture of textbook readability by measuring lexical diversity and syntactic complexity. The finding ascertains that readability is linked to the diversity of genres. Regarding the second question as to how textbook genres might influence readers' comprehensibility, students from Chinese, Korean and Thai background, pursuing 20 different disciplines in Japanese universities are asked to read the corresponding textbooks and reveal the content in Japanese. Table 4 presents MAMR, MAMSP, MDD in textbooks and reader's repeat. It appears that the following Japanese textbooks' readability is close to readers' comprehensibility: life science, tourism, cognitive science, human geography, material geography, politics, science and technology, and informatics. The following textbooks are reportedly difficult to follow: general linguistics, Japanese literature, business

studies, music, education, social welfare, philosophy, civil engineering, Japanese studies and commercial science.

Table 4. MAMR, MAMSP, MDD in textbook and readers' repeat

Textbook genres	**MAMR_ textbook**	**MAMSP_ textbook**	**MDD_ textbook**	**MAMR_ Japanese learners**	**MAMSP_ Japanese learners**	**MDD_Japanese learners**
linguistics	0.0027	1.0039	4.4358	0.0018	1.0009	2.8987
life science	0.0030	1.0038	3.9242	0.0021	1.0027	4.2377
literature	0.0056	1.0076	3.8022	0.0289	1.0429	2.7136
business studies	0.0059	1.0082	4.9052	0.0150	1.0262	3.4022
music	0.0060	1.008	3.8645	0.0051	1.0061	2.9263
informatics	0.0079	1.0107	3.0276	0.0052	1.0095	2.8342
tourism	0.0083	1.0114	3.5177	0.0078	1.0109	4.2600
cognitive science	0.0104	1.0149	3.2927	0.0097	1.0101	4.1484
education	0.0106	1.0142	4.7556	0.0252	1.0384	3.7811
sociology	0.0108	1.0151	3.6537	0.0101	1.0116	4.5109
social welfare	0.013	1.0204	5.2519	0.0110	1.0154	2.6661
philosophy	0.0137	1.0198	4.0851	0.0102	1.0104	3.2274
human geography	0.0137	1.0196	3.7754	0.0111	1.0121	3.4481
economics	0.0146	1.02	4.2534	0.0105	1.0131	3.0604
material engineering	0.0158	1.0207	3.6758	0.0112	1.0198	3.5679
civil engineering	0.0165	1.0233	5.1166	0.0104	1.0187	4.0085
Japanese studies	0.0178	1.0251	3.4588	0.0174	1.0221	2.5336
politics	0.0184	1.0238	3.2433	0.0148	1.0201	3.7456
science technology	0.0212	1.031	3.5856	0.0210	1.0309	3.6857
commercial science	0.0345	1.0477	4.0000	0.0213	1.0385	3.3186

CONCLUSION

This study applies mathematical linguistics to explore the association between text genres, their readability and readers' comprehensibility. The target readers are students from Korea, Thailand and China, studying in Japanese universities, pursuing 20 different disciplines. Japanese is their third language, and all students have passed the Japanese-Language Proficiency Test Level 1. The textbook's readability and readers' comprehensibility are measured at two levels using two metrics. Mean dependency distance is employed for measuring syntactic diversity; moving-average morphological richness and moving-average mean size of paradigm are calculated for measuring lexicon diversity. The findings indicate that in terms of lexical diversity, textbooks of humanities, i.e. life science, literature, business studies, music, informatics, tourism, cognitive science, education, sociology, social welfare, philosophy, human geography and economics, seem simpler than natural science and engineering, e.g. material engineering, civil engineering, science and technology, commercial sciences. As regards syntactic complexity, the textbook of informatics shows the simplest structure while that of social welfare presents the highest complexity. Although it is not very likely that lexical diversity should be corelated to syntactic complexity, text genres have to do with the textbook's readability and eventually influence readers' comprehensibility.

It should be noted that the readers tested in this study are isolated (Chinese, Thai) and agglutinative language-backgrounded Japanese learners (Korean), which limits the scope of evaluating readers' comprehensibility. A closer investigation of Japanese learners' comprehensibility from different first language backgrounds is necessary. Moreover, prior knowledge, motivation and interest also weigh a good deal in text comprehensibility. Nonetheless, this study computes and calculates textbooks' readability, and their associations with genres. This, we hope, could provide sound support for textbook selection in different disciplines.

References

Anderson, M. 2013. The Readability and Usability of Building User Guides. Wellington, NZ: Victoria University.

Bargate, K. 2012. The readability of managerial accounting and financial management textbooks. Meditari Accountancy Research 20(1): 4–20.

Chiang, W.-C., T.D. Englebrecht, T.J. Phillips Jr and Y. Wang. 2008. Readability of financial accounting principles textbooks. The Accounting Educators' Journal 18: 48–80.

Davidson, R.A. 2005. Analysis of the complexity of writing used in accounting textbooks over the past 100 years. Accounting Education 14(1): 53–74.

Feng, Lijun; Elhadad, Noémie; Huenerfauth, Matt. 2009. Cognitively motivated features for readability assessment. *Proceedings of the 12th Conference of the European Chapter of the ACL*: 229–237.

Grabe, W. 2009. Teaching and testing reading. In M. H. Long, and C. J. Doughty (Eds.), The Handbook of Language Teaching (pp. 441–462). Chichester: Wiley and Blackwell.

Hu, M., and Nation, I. S. P. 2000. Unknown vocabulary density and reading comprehension. Reading in a Foreign Language, 13(1), 403−430.

Kendeou, P., McMaster, K.L., and Christ T.J. 2016. Reading Comprehension. Policy Insights from the Behavioral and Brain Sciences, 3 (1), 62–69.

Lee, Bruce W.; Jang, Yoo Sung; Lee, Jason Hyung-Jong. 2021. Pushing on text readability assessment: a transformer meets handcrafted linguistic features. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*: 10669–10686.

Liu, Haitao. 2008. Dependency distance as a metric of language comprehension difficulty. *Journal of Cognitive Science*, 9 (2): 159-191.

Stæhr, L. S. (2008). Vocabulary size and the skills of listening, reading and writing. Language Learning Journal, 36(2), 139–152.

Xia, Menglin; Kochmar, Ekaterina; Briscoe, Ted (2016). Text readability assessment for second language learners. *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*: 12–22