International Journal of Interdisciplinary Research Methods
Vol.9, No.2, pp.37-49, 2022
Print ISSN: ISSN 2398-712X,
Online ISSN: ISSN 2398-7138

Lexical sophistication measures and writing proficiency: The case of Indonesian learners of Japanese

Wenchao Li
Zhejiang University

Wenchao Li (2022) Lexical sophistication measures and writing proficiency: The case of Indonesian learners of Japanese, *International Journal of Interdisciplinary Research Methods*, Vol.9, No.2, pp.37-49

ABSTRACT: *The present study tests two measures of lexical sophistication in writing proficiency (moving-average morphological richness and moving-average mean size of paradigm for testing the lexical diversity) and mean word length for testing writing form (plain, humble, honorification). The findings suggest that the three metrics work reliably. Regarding lexical diversity, moving-average of morphological richness (MAMR) and moving-average mean size of paradigm (MAMSP) of Indonesian Japanese learners-written texts are close to native Japanese-written data. Lexical complexity measured by word length by Indonesian Japanese learners is characterised by slightly less richness than native Japanese data but remains very close. Word length-frequency relationship in the Indonesian-written data presents outstanding fitting results to nine models, including the Poisson Model families and Binomial Model families, with 0.9918 as the lowest and 0.9987 as the highest determination coefficient R^2. It is hoped that this study's outcome may help develop an automatic evaluation of the writing proficiency of agglutinative languages with diverse writing forms.*

KEYWORDS: L3 acquisition, word length, mean dependency distance, dependency direction

INTRODUCTION

Much effort has been put into measuring writing quality in language testing research. Previously, the focus was on fluency, accuracy, and complexity (e.g., Bachman 1989, Bachman and Cohen 1998, Brindley 1998, Cartier 1980, Ishikawa 1995). Skehan and Foster (1999) and Foster et al.'s (2000) studies focus on oral proficiency, and Falhive and Snow

(1980) evaluated compositions by EFL students at the syntactic level. A different view came from quantitative linguistics, computing and calculating language acquisition quality, and language comprehension difficulty (Liu 2008). Several attempts emerged, including Chinese English learning quality (Jiang and Liu 2015; Jiang et al. 2019), Japanese English acquisition proficiency (Komori et al. 2019; Li and Yan 2021), writing proficiency by Hungarian Japanese learners (Li 2022), and oral proficiency by Turkish Japanese learners (2022). These contributions highlight that mean dependency distance can reliably measure syntactic complexity. There is, however, room for further exploration in this line of research. First, it remains to be seen if the language distance between L1 (first language) and L3 (third language) play a part in L3 language acquisition. Second, the measuring units are examined. Mean T-unit length, mean error-free T-unit length, and percentage of error-free T-units seem to have contributed significantly to indexing learning quality (Halleck 1995). Wolfe-Quintero et al. (1998) employ the T-unit complexity ratio, dependent clause ratio, and verb phrase ratio to discriminate the learning levels. Foster et al. (2000) put forward an independent clause or a subclause unit. Regarding the acquisition of the Japanese language, it seems that T-units are usually adopted when testing the syntactic complexity of oral proficiency. Representative work includes Kanakubo et al.'s (1993) investigation of the usage of Japanese in university classes and Ishida's (1991) study on the learning process in Japanese for French-speaking university students. A third issue to be explored is whether language distance involves L2 or L3 acquisition. There is an assumption regarding the affiliation of the Japanese language, that is, alleging it is a mixed language made up of the Northern Tungstic and Southern Austronesian languages (Matsumoto 2007). Sasakiyama (2012) departed from the sound system and lexicons and discovered that in Old Japanese, the Ka(h) particle is used both in Indonesian (*kah*) and Japanese (*ka*) as a question marker, 八 (eight) behaves as a prefix, meaning 'many' instead of a quantifier, and 親 *oya* refers to an ancestor, which resembles Malaysian usage. The similarities in lexicons have inspired linguists to deem Indonesian and Japanese are close typologically. This immediately raises the question if Indonesian Japanese learners will perform well when learning Japanese.

The main purpose of the present study is to explore the measures of three new metrics of lexical sophistication, namely, MAMR, MAMSP, and mean word length (MWL). To this end, Indonesian Japanese learners' writing proficiency is selected as data. The following three

questions are addressed.

Question 1: Can MAMR, MAMSP, and MWL be used to measure the lexical sophistication of Indonesian Japanese learners' Japanese essays? Can these three metrics discriminate learning levels?

Question 2: Does the probability distribution of the mean word length of Indonesian Japanese learners follow a specific distribution?

Question 3: If a specific distribution is followed, what parameters in this distribution can discriminate language proficiency?

Hypothesis:

Suppose MAMR, MAMSP, and MWL can reliably index the lexical sophistication in writing proficiency. In that case, it is hoped that the outcome of this study may provide support for developing an automatic evaluation of writing proficiency for morphologically agglutinative languages, the Altaic language family, and the Austronesian language family.

In this article, Section 2 briefly touches upon the Indonesian language, i.e. the background of Japanese learners. Section 3 outlines the methodology (including the corpora, syntactic parser, and MWL calculation), Section 4 addresses results and discussions, and Section 5 presents the conclusion.

The Indonesian language

The Indonesian language is an Austronesian language. Phonologically, there are six vowels (i, u, é, è, o, a); four diphthongs (/ai/, /au̯/, /oi̯/, /ei̯/) and 24 consonants. It is a stress language, differing from Japanese (moraic language). Indonesian is agglutinative, with tense, aspect, voice, derivation, the case system, etc. rendered by affixes. The case system is conveyed by prefixes, while particles mark the case in Japanese. Syntactically, in contrast with Japanese, whose word order is subject-object-verb, Indonesian's basic word order is subject-verb-object.

METHODOLOGY

Data

The Indonesian writing data were drawn from the International Cross-Sectional Corpus of Japanese as a second language: 72 compositions written in Japanese, with essays titled "Our

Eating Life: Fast Food and Home-Made Food" and emails titled: "asking for a recommendation" were extracted. The data covers all Japanese writing/speaking forms, i.e., plain, honorific and humble forms. The number of tokens in these compositions totalled 34,175 words. Essays and emails written by native Japanese were also included as a comparison to Indonesian Japanese learners. Details of materials are provided in Table 1.

Table 1. The materials

Material	Total words	Title	Writing form
Essays by Indonesian	20728	Our Eating Life	Plain form
Emails by Indonesian	11769	Asking for recommendation Letter	Honorific and humble form

Procedures

The present study examines three lexical measures, MAMR, MAMSP, and MWL, to see whether they are linked to writing proficiency. The 'Learner Text Evaluation System' classification is made at three levels: primary, intermediate, and advanced. By looking into the writing proficiency of different levels, we may understand whether MAMR, MAMSP, and MWL may reliably indicate writing proficiency. The reason for tackling word length is that Japanese is phonologically moraic and morphologically agglutinative. The writing/speaking form includes three variations: plain, humble, and honorific. Crucially, humble and honorific forms are used in formal communication and bear longer word lengths than plain forms. When assessing writings that involve formal expressions, it is necessary to examine the word length of the texts. The following procedures were carried out:

Step 1: Draw raw data from the corpora

Step 2: Save the csv data into visual studio code

Step 3: Parse each sentence via the GiNZA v4 Parser

Step 4: Produce a computer programme to calculate MAMR, MAMSP and MWL from the parsed outputs

Analysis

Moving-average morphological richness (MAMR) and moving-average mean size of paradigm (MAMSP) are calculated for measuring lexicon sophistication. Mean word length (MWL) is employed for testing syntactic diversities. The MAMR, MAMSP and MWL are computed using self-written computer programme scripts. Studies in Cech and Kubat (2018), Covington and McFall (2010), Yan and Liu (2021), Li, Liu and Li (2022) have confirmed that moving window can obtain a better average type-token ratio (TTR). In light of previous work, this study obtains the moving window of TTR in terms of word form via the following formula:

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

And obtain the moving window of TTR in terms of lemma in the following formula:

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

Building on this, we can have lexical sophistication via $\frac{\sum_{I=1}^{N-W+1} F_i}{\mathrm{W}\ \ (\mathrm{N}-\mathrm{W}+1)} - \frac{\sum_{I=1}^{N-W+1} F_i}{\mathrm{W}\ \ (\mathrm{N}-\mathrm{W}+1)}$. The higher of MAMR and MAMSP, the richer of lexicon.

RESULTS AND DISCUSSION

With the methodology highlighted above, this section proceeds to an assessment of the writing proficiency by Indonesian Japanese learners, assessing lexical sophistication in the essays and emails. Section 4.1 presents the MAMR and MAMSP that was calculated using a computer programme. Section 4.2 examines whether writing quality measured by mean word length fits a certain distribution regularity, and if so, the parameters that may suggest a trend in the probability distribution of Japanese writing proficiency.

Writing proficiency measured by MAMR and MAMSP

Writing proficiency in Japanese (plain form)

Figures 1 and 2 show the MAMR and MAMSP of essays by Indonesian Japanese learners and native Japanese speakers.

Figure 1. MAMR and MAMSP of essays by Indonesian Japanese learners

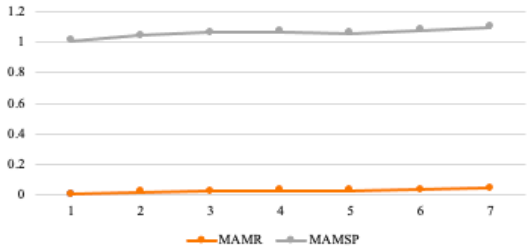
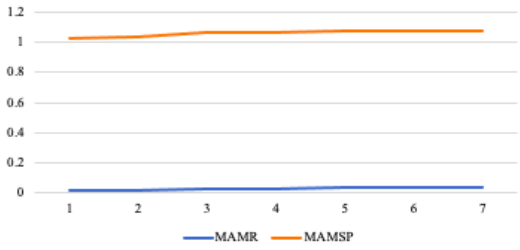


Figure 2. MAMR and MAMSP of essays by native Japanese speakers



It seems that in terms of writings in plain form (essays), along with the increase in learning level, the MAMR, and MAMSP increase. This inspires us to contend that lexical sophistication measured by MAMR and MAMSP is reliable. Moreover, Table 2 shows that Indonesian-written Japanese texts display a closer MAMR and MAMSP than native Japanese data. The similarity in lexical complexity lies in that both Indonesian and Japanese are morphologically agglutinative.

Table 2. Japanese writing proficiency (plain form) by Indonesian Japanese learners, a comparison with native Japanese

Indonesian Japanese leaners	**Average MAMR**	**Average MAMSP**	**Native Japanese**	**Average MAMR**	**Average MAMSP**
Essays	0.0271	1.0609	Essays	0.0299	1.0595

Writing proficiency in Japanese (honorific and humble form)

In email text, honorific and humble expressions are involved. Figure 3 and Figure 4 provide MAMR and MAMSP of emails by Indonesian Japanese learners. A comparison with native Japanese is presented in Table 3.

Figure 3. MAMR and MAMSP of emails by Indonesian Japanese learners

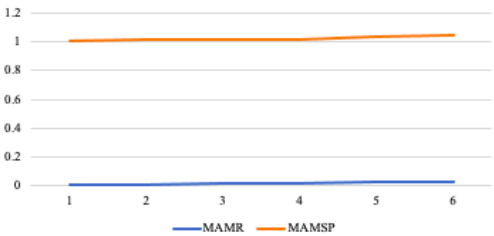


Figure 4. MAMR and MAMSP of emails by native Japanese speakers

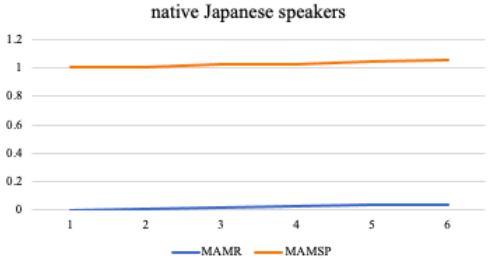


Table 3. Writing proficiency in Japanese (honorific and humble form) by Indonesian Japanese learners, compared with native Japanese

Indonesian Japanese leaners	**Average MAMR**	**Average MAMSP**	**Native Japanese**	**Average MAMR**	**Average MAMSP**
Emails	0.0190	1.0255	Emails	0.0222	1.0299

Along with the increase in learning level, MAMR and MAMSP increase. Although, the Indonesian-written texts show less MAMR and MAMSP than native Japanese texts. Perhaps, honorification and humble form, the uniquely Japanese form of writing and speaking, is hard to acquire.

Writing proficiency measure by word length

This section covers word length in terms of lexical diversity. The reason for paying particular attention to word length is that Japanese is agglutinative, with tense, aspect, voice, and honorification indicated by affixes (mainly suffixes) attached to the verb stem. Essentially, formal writing or speaking in Japanese involves honorification and humble expressions and thus lengthens the word rather than the causal writing or conversation rendered by the plain form. The vocabulary length may efficiently index the formal writing proficiency through incorporation. Emails addressed to a professor asking for a favour are selected as the data.

The findings suggest that the MWL of essays by native Japanese learners ranges from one to seven. The length of most tokens is one, and the average word length is 1.97. The MWL of emails by native Japanese ranges from one to nine, with most tokens being one in length. The average word length is 2.05, longer than essays, due to the writing style, such as emails asking for a favour involving honorification and humble expressions. Regarding Indonesian-written data, the MWL of essays by Indonesian Japanese ranges from one to six, indicating less freedom than native Japanese-written data. Most tokens go to the length of one. The average word length is 1.89, shorter than native Japanese-written data. The MWL of emails by native Japanese ranges from 1 to 9, with most tokens at the length of one. The average word length is 2.17. Tables 4 and Table 5 provide seven sample texts.

Table 4. MWL, MAMR, and MAMSP of essays by Indonesian Japanese learners

Indonesian-written essay	MWL	MAMR	MAMSP	Learning level
IID-e-01	1.89	0.0181	1.0432	Intermediate
IID-e-02	1.90	0.0278	1.0651	Intermediate
IID-e-03	1.91	0.0299	1.0708	Intermediate
IID-e-04	1.92	0.0303	1.0611	Intermediate
IID-e-05	1.94	0.0339	1.0768	Intermediate
IID-e-06	2.01	0.0444	1.0984	advanced
IID-e-07	2.03	0.0495	1.0991	advanced

Table 5. MWL, MAMR, and MAMSP of emails by Indonesian Japanese learners

Indonesian-written essay	MWL	MAMR	MAMSP	Learning level
IID-m-01	2.12	0.0018	1.0023	Intermediate
IID-m-02	2.15	0.0083	1.0112	Intermediate
IID-m-03	2.17	0.0211	1.0274	Intermediate
IID-m-04	2.17	0.0237	1.0308	Intermediate
IID-m-05	2.20	0.0377	1.0507	Intermediate
IID-m-06	2.22	0.0407	1.0575	Advanced
IID-m-07	2.24	0.0411	1.0610	Advanced

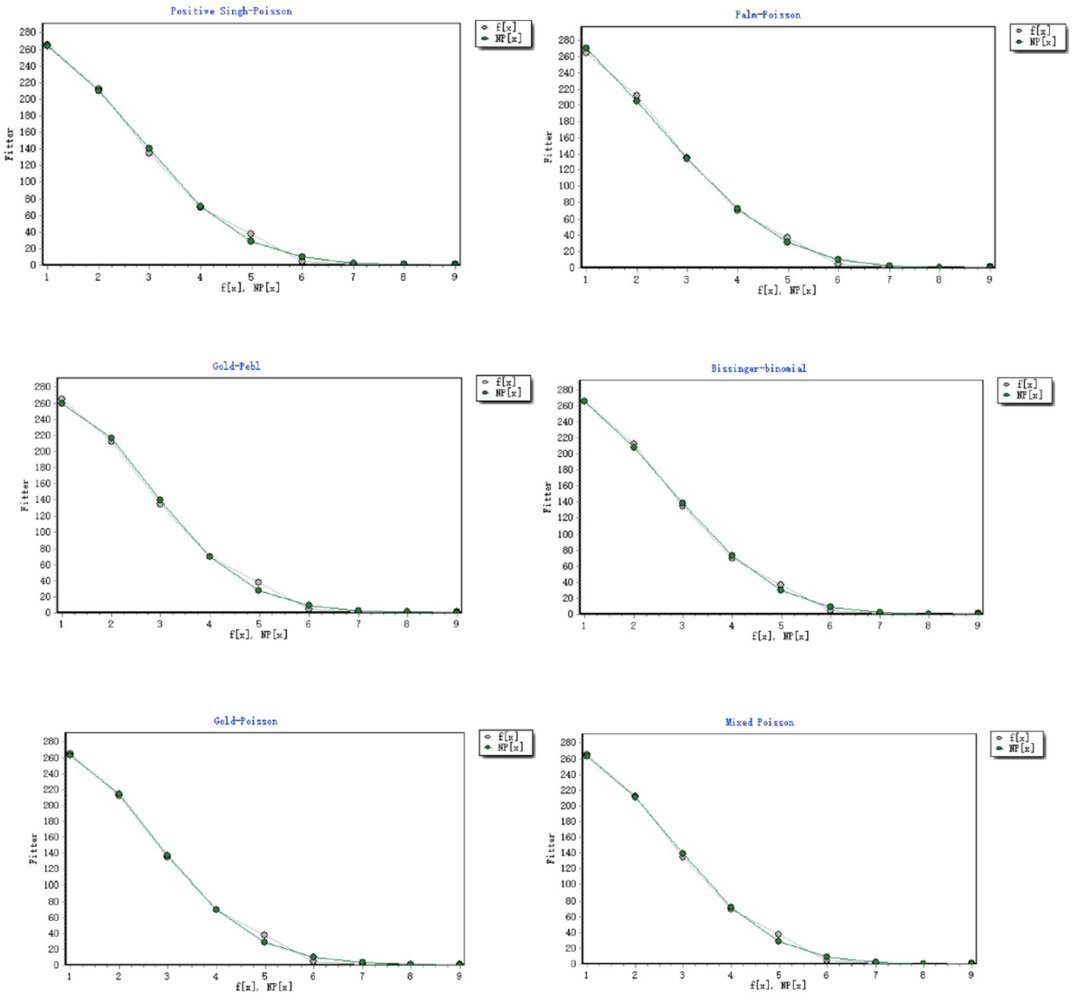
Tables 4 and 5 suggest that word length, MAMR, and MAMSP present a synchronisation trend. Building on these, we may argue that lexical diversity can be indexed via MAMR and MAMSP; writing style can be checked via word length. The three metrics together well

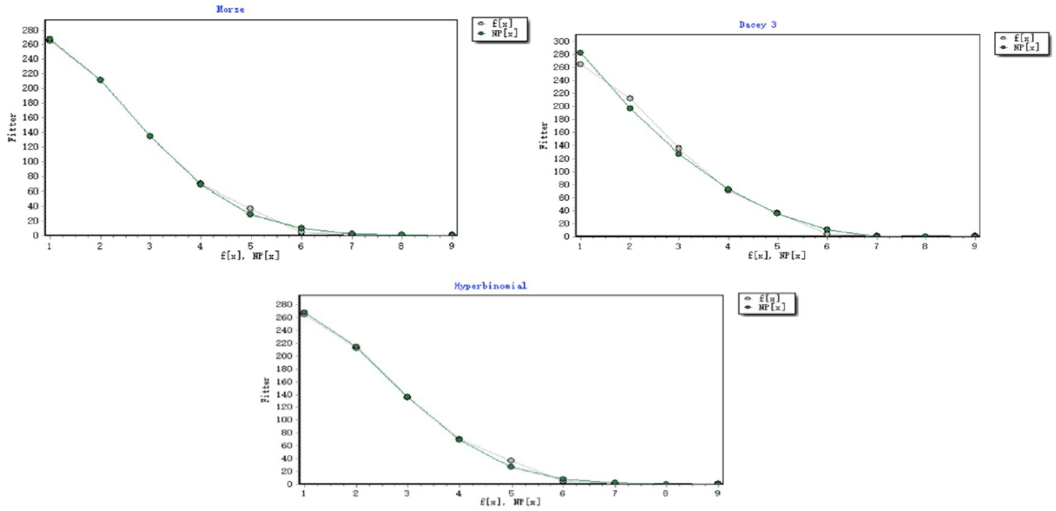
indicate lexical sophistication in writing proficiency.

The probability distribution of writing proficiency

A statistical analysis of the probability distribution of mean word length in Indonesian-written text is carried out to further look at the metrics. Figure 5 shows the relationships between the mean word length and their frequency regarding essay (plain form) concave downwards.

Figure 5. Fitting outcome of the relationship of the distribution of MWL and their frequencies by Indonesian Japanese learners (essays)





As shown in Table 6, nine distribution models are fitted, including the Poisson families, namely, Positive Singh-Poisson, Palm-Poisson, Gold-Pebl, Bissinger-binomial, Gold-Poisson, Mixed Poisson, Morse, Hyperbinomial, Dacey, with 0.9918 as the lowest and 0.9987 as the highest determination coefficient R^2 (R^2 > 0.90, very good; R^2 > 0.80, good; R^2 > 0.75, acceptable; R^2 < 0.75, unacceptable)

Table 6. Fitting the distribution of MWL and their frequencies to models (essays)

Models	**Parameters**									
	a	*b*	*p*	*q*	*m*	*n*	*α*	*R*	X^2	R^2
Positive Singh-Poisson	2.0094						0.9237		6.7824	0.9982
Palm-Poisson	0.1085							9.0000	5.2936	0.9981
Gold-Pebl	1.7933								6.9249	0.9978
Bissinger-binomial	9.0000		0.3752						4.6833	0.9986
Dacey					4.1877	5.0000			5.3524	0.9918
Gold-Pebl	1.7933								6.9249	0.9978
Gold-Poisson	1.9438			0.9510					6.8105	0.9985
Hyperbinomial				0.7616	8.5684	9.0000			5.8217	0.9984
Mixed Poisson	1.5884	0.2675					0.7176		6.0396	0.9984
Morse	0.7912	0.8983							6.1575	0.9987
Palm-Poisson	0.1085							9.0000	5.2936	0.9981
Positive Singh-Poisson	2.0094						0.9237		6.7824	0.9982

A similar analysis is carried out on the relationships between the mean word length and their frequencies regarding emails (honorific and humble form). A concave-down trend and eight fitting distribution models are confirmed, with 0.9917 as the lowest and 0.9985 as the highest determination coefficient R^2.

CONCLUSION

This study has tested three measures for lexical sophistic in writing proficiency: moving-average morphological richness, moving-average mean size of paradigm for testing the lexical diversity, and mean word length for testing writing form. Indonesian Japanese learners writing data were selected. Specifically, 72 compositions written in Japanese, with essays titled, 'Our Eating Life: Fast Food and Home-Made Food' and emails titled: 'Asking for a recommendation' were extracted. It was decided to focus on Indonesian for two reasons. First, both languages are morphologically agglutinative. Second, there is an assumption that the Indonesian and Japanese languages are genetically associated. The findings are summarised as follows.

Regarding lexical diversity, MAMR and MAMSP of Indonesian Japanese learners-written texts are close to native Japanese-written data. This confirms that language typology between mother tongue and L3 language acquisition affects L3 learning quality. Lexical complexity measured by word length by Indonesian Japanese learners is characterised by less richness than native Japanese data but remains very close. With these in place, we conclude that moving-average morphological richness and moving-average mean size of paradigms are good indexes for testing lexical diversity, and mean word length can discriminate writing form (style).

Furthermore, the word length-frequency relationship in the Indonesian-written data presents outstanding fitting results to nine models, including the Poisson Model families and Binomial Model families, with 0.9918 as the lowest and 0.9987 as the highest determination coefficient R^2. It is hoped that this study's outcome may be of help to developing automatic evaluation in writing proficiency of agglutinative language with diverse writing forms.

References

Bachman, L. F. 1989. Language testing–SLA research interfaces. Annual Review of

Applied Linguistics 9, 193–209.

Bachman, L., and Cohen, A. 1998. Language testing–SLA interfaces: An update. In L. Bachmann and A.Cohen (Eds.), Interfaces between second language acquisition and language testing research (pp.29–31). Cambridge, England: Cambridge University Press.

Brindley, G. 1998. Describing language development? Rating scales and SLA. In L. Bachmann and A. Cohen (Eds.), Interfaces between second language acquisition and language testing research (pp. 112–140). Cambridge, England: Cambridge University Press.

Cartier, F. A. 1980. Alternative methods of oral proficiency assessment. In J. R. Frith (Ed.), Measuring spoken language proficiency (pp. 7–14). Washington, DC: Georgetown University Press.

Čech, Radek and Kubát, M. 2018. Morphological richness of text. *Taming the corpus: From inflection and lexis to interpretation*. Fidler M., Cvrček V. (eds.). Cham: Springer, 2018, 63–77. DOI: 10.1007/978-3-319-98017-1_4. Champaign, IL, USA: NCTE.

Falhive, D., and. Snow, B. G. 1980. The use of objective measures of syntactic complexity in the evaluation of compositions by EFL students. In J. Oller & K. Perkins (Eds.), Research in language testing (pp. 171–176). Rowley, MA: Newbury House.

Foster, P., Tonkyn, A., &Wigglesworth, G. 2000. Measuring spoken language: A unit for all reasons. Applied Linguistics, 21, 354–374.

Hunt, K. 1965. Grammatical structures written at three grade levels. *NCTE Research Report* No. 3. Champaign, IL, USA: NCTE.

Hunt, K. 1970. Recent measures in syntactic development. In M. Lester (Ed.), *Reading in applied transformation grammar*. New York.

Ishikawa, S. 1995. Objective measurement of low-proficiency EFL narrative writing. *Journal of Second Language Writing* 4, 51–69.

Jiang, J., and Liu, H. 2015. The effects of sentence length on dependency distance, dependency direction and the implications-based on a parallel English–Chinese dependency treebank. *Language Sciences, 50*, 93–104.

Jiang, J., Bi, P., and 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.

Komori, S., Sugiura, M., Li, W. 2019. Examining MDD and MHD as syntactic complexity measures with intermediate Japanese learner corpus data. *Proceedings of the Fifth International Conference on Dependency Linguistics (Depling, Syntax Fest)*:130–135.

Li, W. and Yan. J. 2021. Probability Distribution of Dependency Distance Based on a Treebank of Japanese EFL Learners' Interlanguage, *Journal of Quantitative Linguistics*, 28:2, 172-186

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

Skehan, P., and Foster, P. 1999. The influence of task structure and processing conditions on narrative retellings. Language Learning, 49, 93–120.