READABILITY OF EXAMPLE SENTENCES IN WRITING ASSISTANCE TOOL NATSUME

Bor Hodošček, Tokyo Institute of Technology

Abekawa Takeshi, National Institute of Informatics

Masao Murota, Tokyo Institute of Technology

Kikuko Nishina, Tokyo Institute of Technology

Abstract: In the online writing assistance tool Natsume, learners of Japanese as a Second Language learn how to use collocations by viewing them in authentic example sentences obtained from a varied collection of corpora. The present research examines the effectiveness of different predictors of readability on sentences, paragraphs, and samples contained in Natsume to help users more easily understand and make better use of collocations. Preliminary results from a model based on 14 predictors indicate that while prediction accuracy is high at the level of whole-text samples, most predictors do not work reliably at the sentence level.

Keywords: Readability, L2 writing assistance, Example sentences, GLM

1: INTRODUCTION

Natsume (http://hinoki.ryu.titech.ac.jp/natsume/) is an online writing assistance tool that enables learners of Japanese as a foreign language to find and make correct use of collocations appropriate to their writing context. Users searching for a specific word (noun, verb or adjective) with Natsume are presented with, in order of importance and grouped by case particle, lists of collocating words. Additionally, searching for a specific pattern (noun, particle, and verb/adjective) enables the user to both visually compare the pattern's usage in different genres, and see authentic example sentences taken from various corpora that contain the pattern.

Currently, all example sentences in Natsume are shown in random order with no consideration for readability, which is often defined as the relative difficulty of reading or understanding a text. The present research aims to improve the readability of example sentences in Natsume by eventually developing a readability measure for L2 Japanese learners that can be used to sort example sentences based on the users query.

2: PREVIOUS RESEARCH

Readability research in English has a long and varied history beginning as early as the late 19th century (Benjamin, 2012). In contrast, there is considerably less published research on Japanese readability formulas. One exception is Tateishi et al.'s (1988) readability formula, which is a linear combination of the average number of characters, average roman, hiragana, katakana, and kanji characters per sentence, and the ratio of commas to periods. More recently, two new readability assessment systems have been developed: one using a multivariable polynomial model based on surface features, but including predicate counts (Shibasaki & Hara, 2010), and another using a bigram language model (Sato et al., 2008).

3: PREDICTORS OF READABILITY IN NATSUME

The aforementioned Japanese readability formulas and language-model based classifiers view readability in the context of the Japanese education system, and as such their applicability to L2 learners is not clear. However, as we do not have access to a corpus graded for L2 Japanese learners, we use the Textbook sub-corpus from the *Balanced Corpus of Contemporary Written Japanese* version 1.0 (Maekawa 2007). The Textbook sub-corpus consists of samples from nationally approved K-12 textbooks, with every sample containing information on its grade level, which is used in the present research as a proxy for readability. However, in contrast to previous studies use of more traditional surface features, we include several vocabulary and syntactic-level features that may be more applicable to L2 learners (Heilman et al. 2007). Overall, we consider 14 different linguistic features for use in predicting readability, which can be grouped according to complexity and the linguistic level they operate on:

- Syntactic features: Using the CaboCha Japanese dependency structure analyzer (http://code.google.com/p/cabocha/), we calculate the average chunk depth, as well as the average distance between a chunks and its dependent chunk. Additionally, while not strictly a syntactic feature, we measure the number of predicates per sentence (see Shibasaki & Hara 2010).
- Surface features: In addition to the number of characters in a sentence, we use the dependency structure of sentences to further measure the amount of tokens (morphemes) and chunks.
- Writing system features: The Japanese writing system employs several different scripts in writing: romaji, katakana, hiragana, kanji, and symbols.
- Vocabulary features: As an alternative to the other features that focus on surface or structural features of language, we include two vocabulary level features. The first is the average (pre-2010) Japanese Language Proficiency Test (JLPT) word level, which is in line with our goals of measuring L2 readability. The second is the average log-scaled probability of tokens (matched on POS and lemma) found in the BCCWJ Library Books (LB) sub-corpus, which is used as a measure of general vocabulary usage in the context of printed materials for adult native speakers—contrasting with vocabulary usage encountered in the Textbook sub-corpus, which is meant for adolescents.

4: RESULTS

In order to identify correlates of readability at the sentence, paragraph, and sample level, we compute the Pearson correlation coefficient against the grade level for every feature (Table 1). At all levels of analysis, hiragana and kanji consistently showed the highest correlations with grade level. When comparing between levels and groups, hiragana and kanji from the writing system group, and surface and syntactic groups in general correlate higher as the level of analysis increased from sentence to sample, while vocabulary group features remain mostly unaffected. Finally, the group of surface features (characters, tokens, and chunks), as well as link distance and chunk depth from the syntactic group were all found to be highly correlated ($r>0.9$, $p<0.001$) at all levels of analysis.

Table 1: Correlations of linguistic features with Textbook grade levels

Feature group	Feature	Correlation at level		
		Sentence	Paragraph	Sample
Writing system	Hiragana	**-0.334**	**-0.470**	**-0.720**
	Katakana	0.091	0.142	0.168
	Kanji	**0.331**	**0.443**	**0.730**
	Romaji	0.067	0.145	0.153
	Symbols	-0.071	-0.056	-0.373
	Commas	-0.047	-0.114	-0.028
Surface	Characters	0.206	0.147	0.334
	Tokens	**0.209**	0.303	0.589
	Chunks	0.201	0.290	0.597
Syntactic	Link distance	0.174	0.302	**0.631**
	Chunk depth	0.182	**0.311**	0.606
	Predicates	0.126	0.190	0.544
Vocabulary	JLPT word level	-0.206	-0.169	-0.283
	BCCWJ-LB word level	-0.154	-0.185	-0.301

Having examined individual predictor's correlation to grade level, we now examine the effect of combining predictors in order to improve readability prediction performance. Using the glmnet package for R, we constructed a generalized linear model of readability in the Textbook sub-corpus (lambda chosen using 10-fold cross validation with the caret package for R). Because of the small amount of data available for grades 1-9, we chose to group grades into three categories: elementary school (G1-6), middle school (G7-9), and high school (G10-12). In addition, the Textbook sub-corpus was partitioned into training and test sets (75%:25%) containing equivalent proportions of G1-6, G7-9 and G10-12 observations for model evaluation purposes (Table 2).

The confusion matrix and performance statistics for each model in Table 3 show a general increase in misclassification from samples to sentences, and in G7-9, in particular. When comparing grade levels, we see that G7-9 is consistently misclassified compared to the other grades, with the sentence model only classifying sentences as either G1-6 or G10-12.

Table 2: Textbook sub-corpus training and test sets

	Sentences	Paragraphs	Samples
Training set	40136	7314	311
Testing set	13377	2436	101
Total:	53513	9750	412

Table3: Confusion matrix and model statistics for G1-6, G7-9 and G10-12 on the test set

Reference / Prediction	Sentences			Paragraphs			Samples		
	G 1-6	G 7-9	G 10-12	G 1-6	G 7-9	G 10-12	G 1-6	G 7-9	G 10-12
G1-6	665	257	255	264	83	51	21	1	2
G7-9	0	0	0	32	68	25	0	5	0
G10-12	1671	2543	7986	182	444	1287	2	10	60
Accuracy:	0.647			0.665			0.851		

Although not shown here because of space constraints, an examination of importance of predictor variables using ROC curve analysis for each model indicated that features that do not necessarily appear in most sentences, like romaji, katakana, symbols and commas, are less important predictors in the sentence model. This suggests that sentence-level predictors should be general enough to be measurable in most sentences or at the very least, help with discrimination where other more general features fail. In cases where predictors are dependent on each other (such as the inverse relationship between hiragana and kanji in a sentence), inclusion of all dependent predictors in the model may be detrimental to classifier performance.

5: CONCLUSIONS AND FUTURE WORK

We have investigated which linguistic features correlate best with Japanese K-12 Textbooks' grade levels, constructed a model using a combination of features, and measured the effectiveness of each predictor in the model at each level of analysis (sentence, paragraph, and sample). Our preliminary findings suggest that for certain features to be effective, they need to be used at the paragraph or greater level, though some predictors, like JLPT word level, were found to be less affected by level of analysis.

While the present model and predictors were effective on the scale of samples, future work should focus on developing new predictors that also perform effectively at the sentence level. In addition, a prerequisite for the development of an L2 readability measure is to construct an L2 readability-graded Japanese language corpus.

ACKNOWLEDGMENTS

This work was supported by a Challenging Exploratory Research Grant-in-Aid (No. 22652048) from the Ministry of Education, Science, Sports and Culture of Japan.

REFERENCES

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educ Psychol Rev 24*, 63–88.

Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. *In Proceedings of the NAACL Human Language Technology Conference*, 460–467.

Maekawa, K. (2007). Design of a balanced corpus of contemporary written Japanese. *In Proceedings of the Symposium on Large-Scale Knowledge Resources (LKR2007)*, 55–58.

Sato, S., Matsuyoshi, S., & Kondoh Y. (2008). Automatic assessment of Japanese text readability based on a textbook corpus. *LREC-08*, 654–660.

Shibasaki, H. & Hara S. (2010). 12 gakunen wo nan'isyakudou to suru nihongo riidabiritii hanteisiki [Japanese K-12 readability formula]. *Mathematical Linguistics, 27*(6), 215–232.

Tateisi Y., Ono Y., & Yamada H. (1988). A computer readability formula of Japanese texts for machine scoring. *In Proceedings of the 12th Conference on Computational Linguistics, 2*, 649–654.