Introducing a Readability Evaluation System for Japanese Language Education

Yoichiro HASEBE (長谷部 陽一郎), Doshisha University
Jae-Ho LEE (李 在鎬), University of Tsukuba

2015 CASTEL/J

Introducing jreadaiblity.net

Japanese text readability evaluation system

http://jreadability.net

The system outputs six-level readability score

Lower-Elementary | Upper-Elementary | Lower-Intermediate | Upper-Intermediate | Lower-Advanced | Upper-Advanced

Yet another readability formula?

Aim of the present research

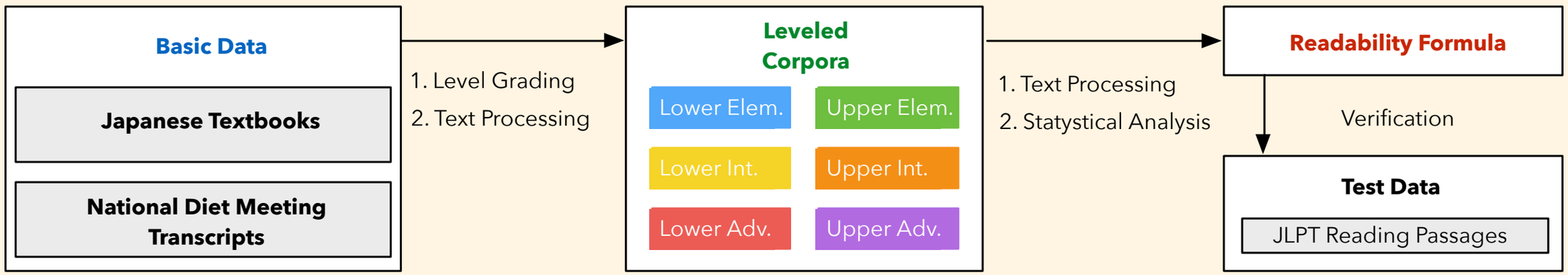
1. To produce a formula to measure the readability level of given text that is used in teaching Japanese
2. To create an online system based on the formula.

Difference from preceding studies

1. We use leveled corpora that consist of numerous textbooks for teaching Japanese as model data
2. The online system is equipped with functionalities that help both instructors and learners deal with text in their teaching and studying

Organization of the talk

- Preparation of original data
- Compiling leveled corpora
- Six-level readability scale
- Constructing a readability formula
- Basic design of the online system
- Additional features of the system
- Limitations and restrictions



About original data

Two types of text

Textbooks for teaching Japanese

- 83 textbooks of a variety of levels
- Traits: controlled vocabulary, idioms, structures, and types of logic

Lower-Elementary | Upper-Elementary | Lower-Intermediate | Upper-Intermediate | Lower-Advanced

National Diet meeting transcripts

- Distributed in *Balanced Corpus of Contemporary Written Japanese* (BCCWJ)
- Traits: actual utterances, variety of styles, and long sentences

Upper-Advanced

Compiling leveled corpora

Method

1. Split the whole dataset into files of about 1,000 characters, producing 995 text files in total
2. Preliminary grouping of the dataset by one of the authors (according to the general facts already known about the titles)

Compiling leveled corpora

Table 1. Basic statistics of the original data

	Lower-elem.	Upper-elem.	Lower-int.	Upper-int.	Lower-adv.	Upper-adv.
Files	133	117	148	286	117	194
Word types	3,178	2,858	5,156	10,291	6,833	4,712
Word tokens	72,691	68,746	87,433	174,953	69,268	122,269
Original data	83 textbooks for teaching Japanese					National Diet meeting transcripts

Total number of characters in total: 595,360

Compiling leveled corpora

Method (cont'd)

3. Three experienced teachers of Japanese examined all the files in each of the levels and picked out 30 files that well represented the level
4. Selection of 20 files that were chosen by multiple graders for each level
5. Discriminant analysis was conducted, using the resulting data as model, against the text files that had been “filtered out” in the previous process

Compiling leveled corpora

Table 2. Discriminatory analysis results

		Predicted Group Membership						
		Upper-adv.	Lower-adv.	Upper-int.	Lower-int.	Upper-elem.	Lower-elem.	Num of files
Original	Upper-adv.	**152**	14	8	0	0	0	174
	Lower-adv.	6	**60**	24	7	0	0	97
	Upper-int.	8	70	**102**	61	22	3	266
	Lower-int.	0	4	39	**58**	21	6	128
	Upper-elem.	0	1	14	28	**37**	17	97
	Lower-elem.	0	0	0	7	28	**78**	113
	Num of files	166	149	187	161	108	104	875

Total number of files included in the leveled corpora

20 × 6 = 120 core data files

152 + 60 + 102 + 58 + 37 + 78 = 487 additional data files

Constructing a readability formula

NLP Tools

- Morphological Analyzer　MeCab (0.996)
- Morphological Dictionary　UniDic (2.2.0)

Candidate variables

- Mean length of sentence
- proportion of *kango* (words of Chinese-origin)
- proportion of *wago* (words of Japanese-origin)
- proportion of verbs
- proportion of auxiliary verbs
- etc.

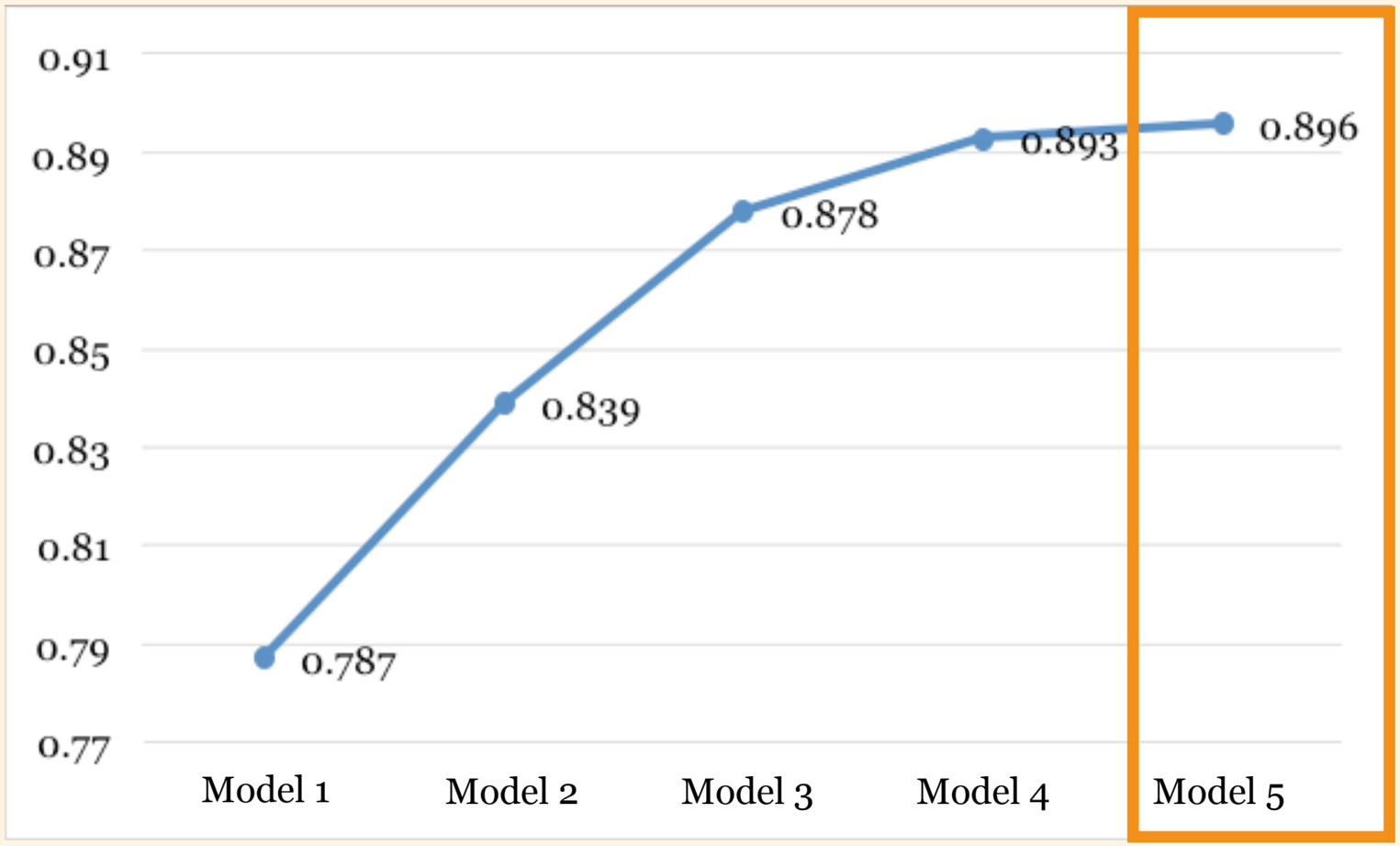
Cf. Shibasaki and Hara (2010)

Constructing a readability formula

Table 3. Results of multiple linear regression analysis

Model	Variables	Coefficients	R^2
Model 1	Constant Mean length of sentence	5.938 -.099	0.787
Mode 2	Constant Mean length of sentence Proportion of *kango*	6.691 -.082 -.073	0.839
Model 3	Constant Mean length of sentence Proportion of *kango* Proportion of *wago*	13.195 -.063 -.153 -.086	0.878
Model 4	Constant Mean length of sentence Proportion of *kango* Proportion of *wago* Proportion of verbs	12.128 -.057 -.142 -.061 -.159	0.893
Model 5	Constant Mean length of sentence Proportion of *kango* Proportion of *wago* Proportion of verbs Proportion of auxiliary verbs	11.724 -.056 -.126 -.042 -.145 -.044	0.896

Constructing a readability formula



Transition of the coefficient of determination (R^2)

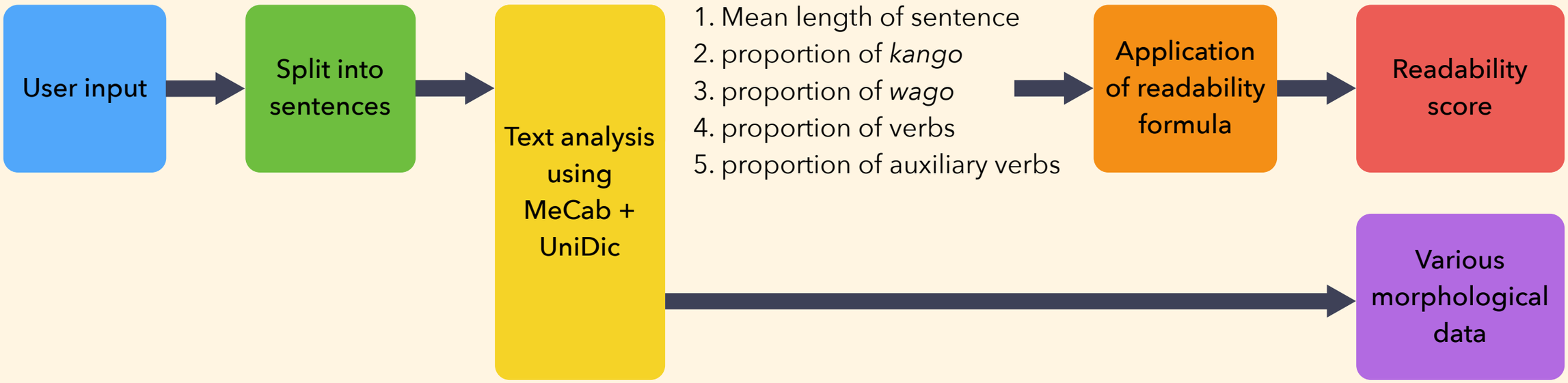
The readability formula

X = (mean length of sentence * -0.056)
+ (proportion of *kango* * -0.126)
+ (proportion of *wago* * -0.042)
+ (proportion of number of verbs among all words * -0.145)
+ (proportion of number of auxiliary verbs * -0.044) + 11.724 (R^2 = .896)

Table 4. Levels and readability score ranges

Level	Readability score range
Upper-advanced	0.5 - 1.4
Lower-advanced	1.5 - 2.4
Upper-intermediate	2.5 - 3.4
Lower-intermediate	3.5 - 4.4
Upper-elementary	4.5 - 5.4
Lower-elementary	5.5 - 6.4

Basic design of the online system



Word	Part-of-Speech	Pronunciation	Word Category
建設	名詞-普通名詞-サ変可能	ケンセツ	漢（Chinese-origin）
中	接尾辞-名詞的-副詞可能	チュー	漢（Chinese-origin）
の	助詞-格助詞	ノ	和（Japanese-origin）
美しい	形容詞-一般形容詞-連体形	ウツクシイ	和（Japanese-origin）
ビルディング	名詞-普通名詞-一般	ビルディング	外（Foreign-origin）

The web interface

Paste text to the input box and press 実行 ("run") button

Once a readability evaluation has been complete, Read-aloud functionality will be available on Google Chrome.

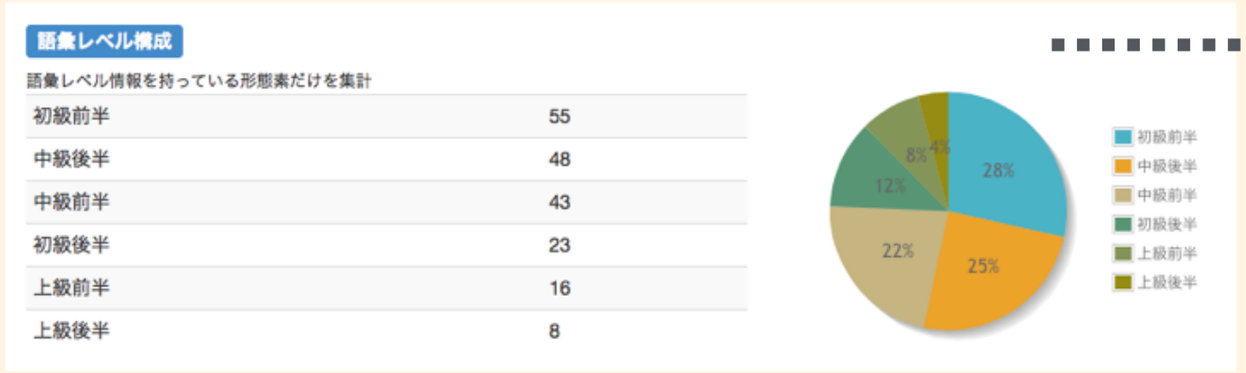
Main results panel

テキストの概要

総形態素数（異なり）を表示するには「語彙リストを出力」をオンに

文章難易度	上級前半	Readability Level 上級後半: Upper-advanced 上級前半: Lower-advanced 中級後半: Upper-intermediate 中級前半: Lower-intermediate 初級後半: Upper-elementary 初級前半: Lower-elementary
リーダビリティ・スコア	2.15	Readability Score
総文数	12	Total number of sentences
総形態素数（延べ）	458	Total number of word tokens
総形態素数（異なり）	184	Total number of word types
総文字数（記号・空白を含む）	741	Total number of characters
一文の平均語数	38.17	Mean number of words per sentence

Main results panel



語彙レベル構成

語彙レベル情報を持っている形態素だけを集計

初級前半	55
中級後半	48
中級前半	43
初級後半	23
上級前半	16
上級後半	8

Distribution of words of different vocabulary levels

品詞構成

記号類は除外

普通名詞	127
助詞	116
助動詞	47
動詞	26
固有名詞	15
副詞	8
形状詞	4
形容詞	4
代名詞	2
連体詞	1
その他	108

Distribution of words of different parts-of-speech

語種構成

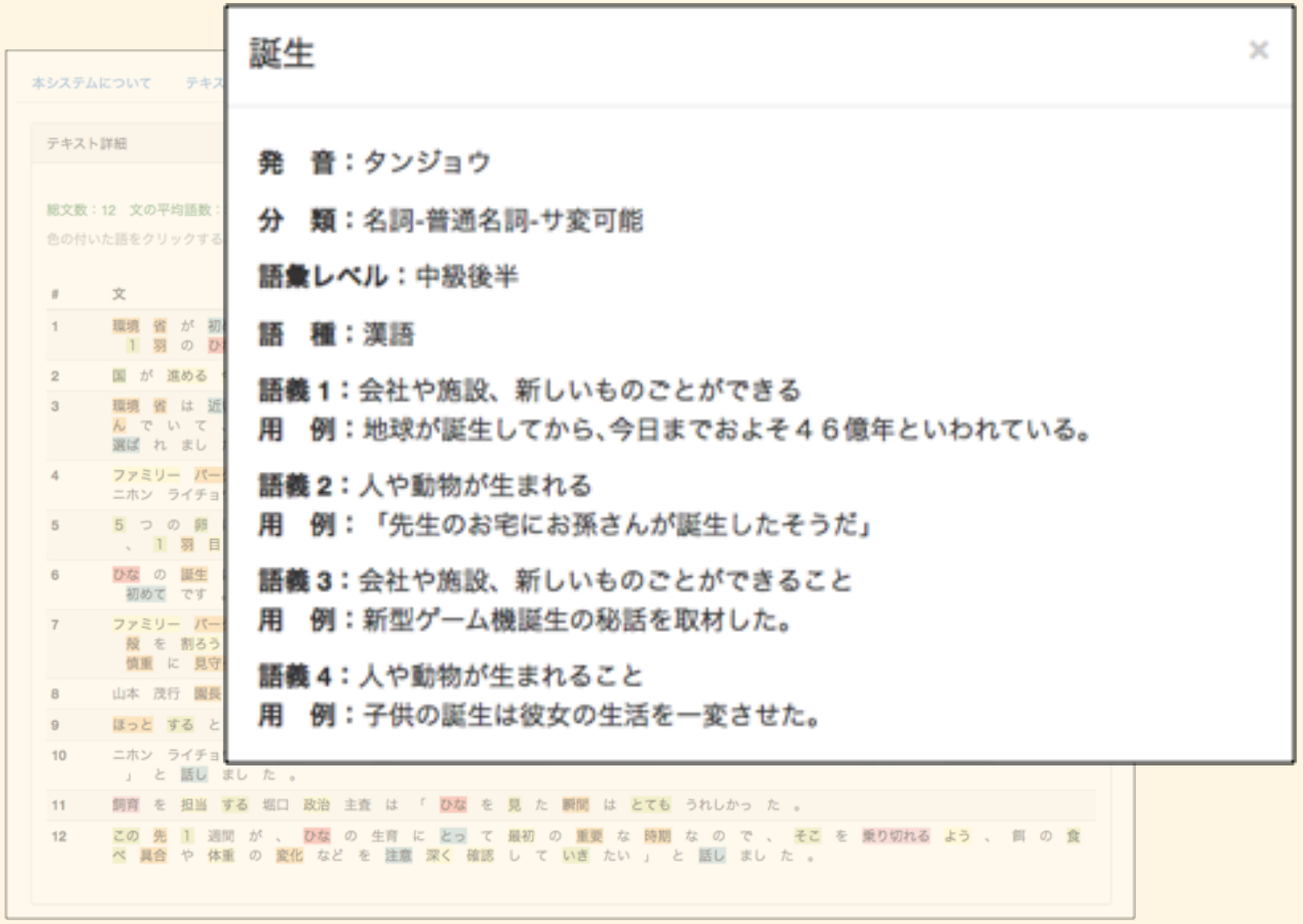
定型句は「ありがとう」などを指す

和語	276
漢語	108
外来語	10
混種語	4
定型句	0

Distribution of words of different word categories (*kango*, *wago*, *gairaigo*, etc.)

Text details panel

Dictionary lookup



Dictionary data created and discussed in Sunakawa et al. (2012)

Vocabulary list

Basic form	**取り組む**	(*torikumu*)
Pronunciation	**トリクム**	(*torikumu*)
Grammatical category	**動詞-一般**	(verb–general)
Surface form(s)	**取り組ん**	(*torikun-*)
Frequency (%)	2	(0.44%)
Vocabulary level	**中級後半**	(upper-intermediate)

Limitations and restrictions

Morphological parsing is not 100% perfect

たまご	の	きみ	を	こぼし	た
名詞	助詞	代名詞	助詞	動詞	助動詞
Noun	Particle	Pronoun	Particle	Verb	Auxiliary Verb

Short-unit words (SUW) instead of Long-unit words (LUW)

環境 省 の 発表 に よる と 、 ニホン ライチョウ の 人工 飼育 事業 で 2 7 日 、 1 羽 の ひな が 富山 市 ファミリー パーク で 誕生 し まし た 。

A ptarmigan chick was hatched artificially at Toyama Municipal Family Park Zoo on June 27, part of an effort to boost the numbers of the protected game bird in Japan, the Environment Ministry announced.

Limitations and restrictions

Text that is too short may get inaccurate evaluation

→ All model data consists of files of about 1,000 words

Text with too many *kango* words are problematic

→ The readability score easily goes out of range

Summary and conclusion

Data

Text data extracted from textbooks for teaching Japanese and National Diet meeting transcripts

Method

A multiple regression analysis on the results of semi-manual/semi-statistical grouping of text to construct a readability formula

Application

A web-based system to evaluate readability of input text as well as to provide functionalities that benefit both instructors and learners of the Japanese language

References

Den, Yasuharu. (2009) A Multi-Purpose Electronic Dictionary for Morphological Analyzers [in Japanese]. *Journal of the Japanese Society for Artificial Intelligence* 24(5), 640-646.

Flesch, Rudolph. (1948) A new readability yardstick. *Journal of Applied Psychology* 32(3), 221-233.

Lee, J.-H. (2011) The Utility of Corpora for Composing Reading Comprehension Questions for Large-Scale Tests. *Nihongo Kyouiku* [Journal of Japanese Language Teaching] 148, 84–98.

Lee, J.-H. and H. Shibasaki. 2012. Relationship between Readability and Vocabulary: What Lexical Properties Determine the Characteristics of Text for Different School Grades [in Japnaese]. *Corpus and Text Mining*. Tokyo: Kyoritsu Shuppan, 181–192.

Sakai Yukiko. (2011) Improvement and evaluation of readability of Japanese health information texts: An experiment on the ease of reading and understanding written texts on disease [in Japanese]. *Library and Information Science* 65, 1-35.

Sakamoto, Ichiro. (1967) Assessing the weight of sentence length: An attempt to approach the readability [in Japanese]. *Dokusho Kagaku* [Science of Reading] 7, 1-6.

Sato, Satoshi. (2011) Measuring text readability based on balanced corpus [in Japanese]. *IPSJ Journal* 52(4), 1777-1789.

Shibasaki, Hideko and Shin-ichiro Hara. (2010) The readability formula to predict school grades 1-12 based on Japanese language school textbooks [in Japanese]. *Keiryoo Kokugogaku* [Mathematical Linguistics] 26(6), 215-232.

Smith, Edgar A. and J. Peter Kincaid. (1970) Derivation and validation of the automated readability index for use with technical materials. *Human Factors* 12, 457-464.

Sunakawa, Yuriko, Lee, Jae-ho, and Takahara, Mari. (2012) The Construction of a Database to Support the Compilation of Japanese Learners Dictionaries. *Acta Linguistica Asiatica* 2(2), 97-115.

Tateishi, Yuka, Yoshihiko Ono, and Hisao Yamada. (1988) A computer readability formula of Japanese texts for machine scoring. *Proceedings of the 12th Conference on Computational Linguistics* 2, 649-654.

Zhang, Yujie and Kazuhiko Ozeki. (1998) Automatic bunsetsu segmentation of Japanese sentences using a classification tree. *Language, Information and Computation (Proceedings of PACLIC12)* 18-20, 230-235.

Thank you!

Yoichiro HASEBE (長谷部 陽一郎), Doshisha University

yohasebe@gmail.com

Jae-Ho LEE (李 在鎬), University of Tsukuba

jhlee.n@gmail.com