

# 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 jreadability.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

The screenshot displays the jreadability.net interface. On the left, a sidebar shows navigation options like 'システム利用' and '利用規約'. The main content area is divided into several sections:

- システム概要:** A table showing system statistics: 28, 16, 15, 2, 1, 1.
- システム概要 (Pie Chart):** A pie chart showing the distribution of readability scores across different levels.
- システム概要 (Table):** A table showing readability scores for various levels: 52, 39, 25, 10, 9, 3, 1, 1, 1, 39.
- システム概要 (Text):** A text box containing a paragraph of Japanese text about Spain's economy and tourism.
- システム概要 (Table):** A table showing readability scores for 21 different sentences. The scores range from 5 to 21. The table includes columns for '文章ID', '文章', '読者', 'レベル', and 'スコア'.

文章ID	文章	読者	レベル	スコア
1	スペインの経済が回復している。観光業の回復、という観点、スペインの観光業の回復が待っています。	5	スペイン語	5
2	この文章の読者がこの記事を再読するのを望むたいと書いていますが、スペインの経済は許可していません。	11	中級者	11
3	この文章はスペインに由来するものではありませんが、7月31日、フランスの観光業の回復が待っています。	1	初級者	1
4	観光業の回復が待っています。観光業の回復が待っています。	5	初級者	5
5	観光業の回復が待っています。観光業の回復が待っています。	5	初級者	5
6	観光業の回復が待っています。観光業の回復が待っています。	1	初級者	1
7	観光業の回復が待っています。観光業の回復が待っています。	5	初級者	5
8	観光業の回復が待っています。観光業の回復が待っています。	1	初級者	1
9	観光業の回復が待っています。観光業の回復が待っています。	1	初級者	1
10	観光業の回復が待っています。観光業の回復が待っています。	1	初級者	1
11	観光業の回復が待っています。観光業の回復が待っています。	1	初級者	1
12	観光業の回復が待っています。観光業の回復が待っています。	1	初級者	1
13	観光業の回復が待っています。観光業の回復が待っています。	5	初級者	5
14	観光業の回復が待っています。観光業の回復が待っています。	1	初級者	1
15	観光業の回復が待っています。観光業の回復が待っています。	5	初級者	5
16	観光業の回復が待っています。観光業の回復が待っています。	5	初級者	5
17	観光業の回復が待っています。観光業の回復が待っています。	5	初級者	5
18	観光業の回復が待っています。観光業の回復が待っています。	1	初級者	1
19	観光業の回復が待っています。観光業の回復が待っています。	1	初級者	1
20	観光業の回復が待っています。観光業の回復が待っています。	3	初級者	3
21	観光業の回復が待っています。観光業の回復が待っています。	1	初級者	1

# Yet another readability formula?

## Aim of the present research

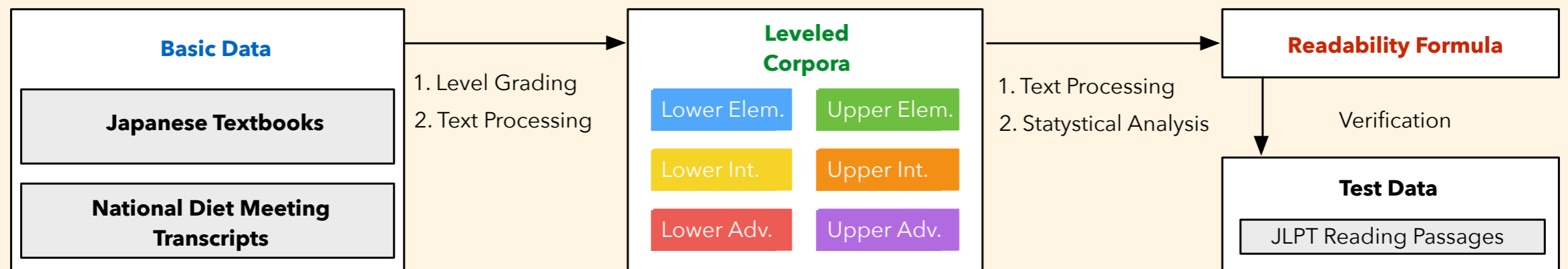
1. To produce **a formula** to measure the readability level of given text that is used in teaching Japanese
2. To create **an online system** based on the formula.

## 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						Num of files
		Upper-adv.	Lower-adv.	Upper-int.	Lower-int.	Upper-elem.	Lower-elem.	
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

$20 \times 6 = 120$  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.

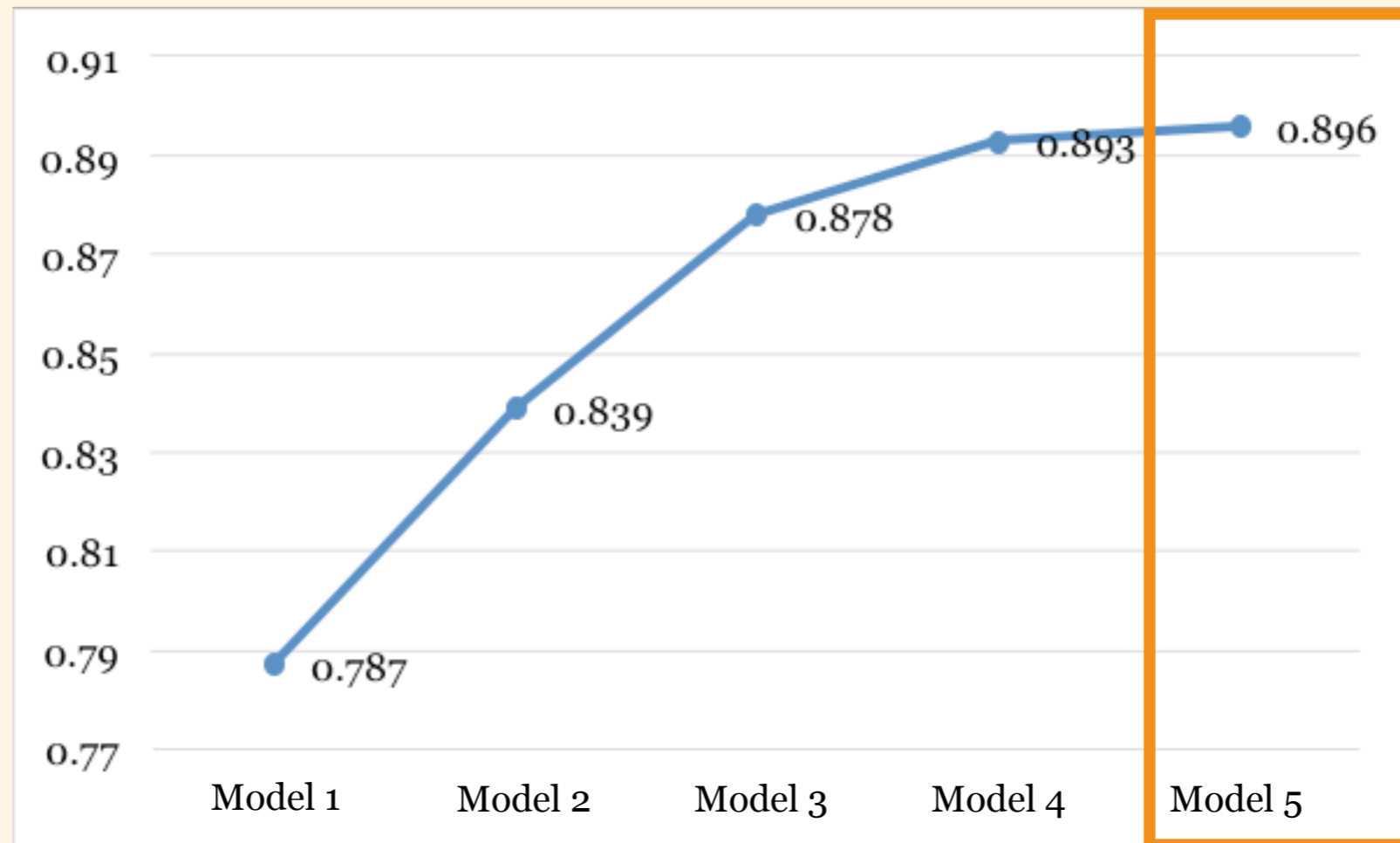
Cf. Shibasaki and Hara (2010)

# Constructing a readability formula

Table 3. Results of multiple linear regression analysis

Model	Variables	Coefficients	R <sup>2</sup>
Model 1	Constant Mean length of sentence	5.938 -.099	0.787
Mode 2	Constant Mean length of sentence Proportion of <i>kango</i>	6.691 -.082 -.073	0.839
Model 3	Constant Mean length of sentence Proportion of <i>kango</i> Proportion of <i>wago</i>	13.195 -.063 -.153 -.086	0.878
Model 4	Constant Mean length of sentence Proportion of <i>kango</i> Proportion of <i>wago</i> Proportion of verbs	12.128 -.057 -.142 -.061 -.159	0.893
Model 5	Constant Mean length of sentence Proportion of <i>kango</i> Proportion of <i>wago</i> 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 ( $R^2$ )

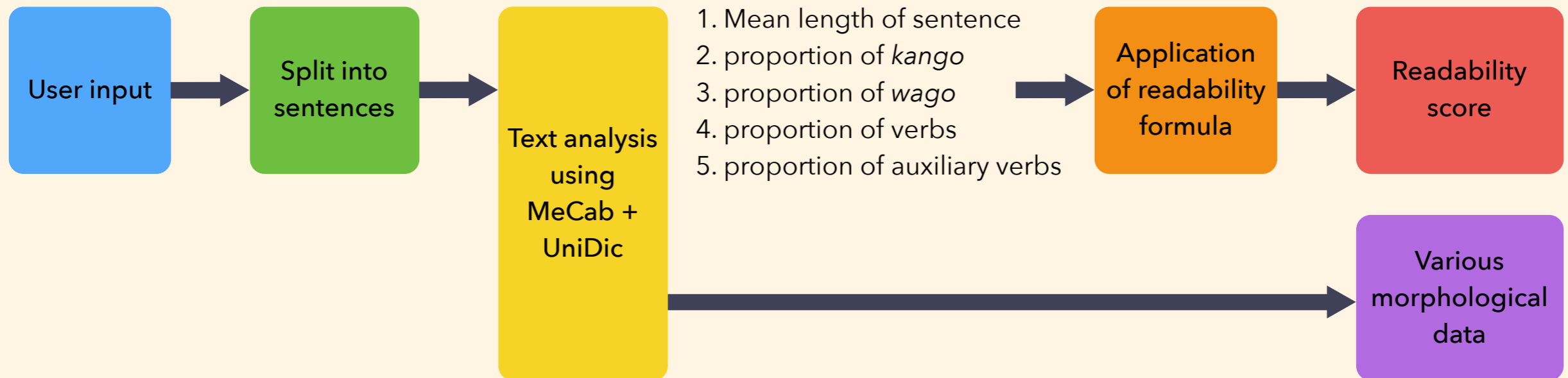
# The readability formula

$$\begin{aligned} X = & \text{ ( mean length of sentence * -0.056 )} \\ & + \text{ ( proportion of } kango \text{ * -0.126 )} \\ & + \text{ ( proportion of } wago \text{ * -0.042 )} \\ & + \text{ ( proportion of number of verbs among all words * -0.145 )} \\ & + \text{ ( proportion of number of auxiliary verbs * -0.044 )} + 11.724 \quad (R^2 = .896) \end{aligned}$$

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

The screenshot shows the web interface for the Japanese text readability evaluation system. At the top left is the logo 'JR' and the title '日本語文章難易度判別システム alpha版'. On the top right are navigation links: 'システム利用', '利用規約', 'よくある質問', '掲示板', and '日本語教育語彙表'. Below the header is a blue button labeled '日本語テキストを入力して実行'. The main content area contains a text input box with a sample article about the artificial breeding of Japanese quail. Below the input box are several checkboxes for output options: 'テキスト詳細情報を出力' (checked), '丸括弧とその内側を除去' (unchecked), '母語話者文章評価' (checked), '学習者文章評価 (試験中)' (unchecked), '語彙リストを出力' (checked), and '青空文庫のルビを除去' (unchecked). At the bottom right of the form are three buttons: '実行' (Run), 'クリア' (Clear), and 'リセット' (Reset).

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** .....

リーダビリティ・スコア 2.15 .....

総文数 12 .....

総形態素数（延べ） 458 .....

総形態素数（異なり） 184 .....

総文字数（記号・空白を含む） 741 .....

一文の平均語数 38.17 .....

## 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

## Readability Score

Total number of sentences

Total number of word tokens

Total number of word types

Total number of characters

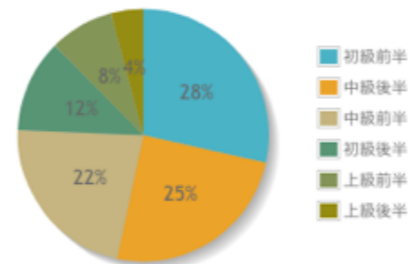
Mean number of words per sentence

# 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

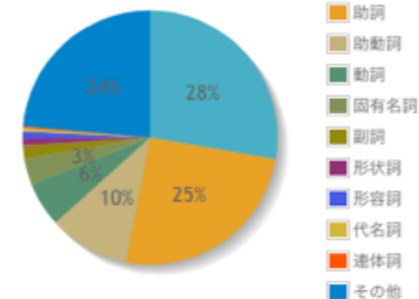


Distribution of words of different vocabulary levels

## 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

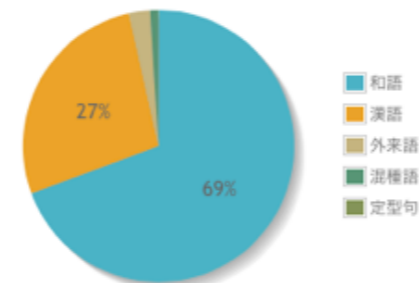


Distribution of words of different parts-of-speech

## 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

本システムについて テキスト情報 テキスト詳細 語彙リスト

テキスト詳細

結果保存 (CSV : Shift-JIS) 結果保存 (CSV : UTF-8)

総文数 : 12 文の平均語数 : 38.17

色の付いた語をクリックすると辞書引きを行います。

初級前半  初級後半  中級前半  中級後半  上級前半  上級後半

#	文
1	環境省が初めて取り組んでいる国の特別天然記念物、ニホンライチョウの人工飼育事業で27日未明、1羽のひなが富山市ファミリーパークで誕生しました。
2	国が進める今回の飼育事業で、ひなが生まれたのは初めてです。
3	環境省は近い将来、絶滅するおそれが高いと指摘されているニホンライチョウの人工飼育に初めて取り組んでいて、富山市の動物園、富山市ファミリーパークが、東京の上野動物園とともに飼育施設に選ばれました。
4	ファミリーパークには今月23日、長野県と岐阜県にまたがる乗鞍岳で3つの巣から採取されたニホンライチョウの卵5個が運び込まれ、人工ふ化させる取り組みが行われてきました。
5	5つの卵は専用の部屋に設置した「人工ふ卵器」で温められ、採取から4日後の27日午前、1羽目のひなが誕生したのをファミリーパークの担当者が確認しました。
6	ひなの誕生は午前1時半ごろとみられ、環境省が進める今回の飼育事業で、ひなが生まれたのは初めてです。
7	ファミリーパークによりますと、午前7時半の時点で、別の卵1個でも、中のひながくちばしで殻を割ろうとする「はし打ち」という行為が確認されているということで、残りの卵のふ化を慎重に見守っています。
8	山本茂行園長は「「こんにちは赤ちゃん」という感じですよ。
9	ほっとすると同時に、これからが大変なので、さらに気を引き締めなくてはという気持ちです。
10	ニホンライチョウの安定した飼育につなげていくための出発点であり、まず第一歩を踏み出しました」と話しました。
11	飼育を担当する堀口政治主査は「ひなを見た瞬間はとてもうれしかった。
12	この先1週間が、ひなの生育にとって最初の重要な時期なので、そこを乗り切れるよう、餌の食べ具合や体重の変化などを注意深く確認していきたい」と話しました。

# Dictionary lookup

The image shows a screenshot of a Japanese dictionary application. On the left, there is a list of text snippets with highlighted words. On the right, a detailed dictionary entry for the word '誕生' (tanjyō) is displayed.

**誕生**

発音：タンジョウ

分類：名詞-普通名詞-サ変可能

語彙レベル：中級後半

語種：漢語

語義1：会社や施設、新しいものごとができる  
用例：地球が誕生してから、今日までおよそ46億年といわれている。

語義2：人や動物が生まれる  
用例：「先生のお宅にお孫さんが誕生したそうだ」

語義3：会社や施設、新しいものごとができること  
用例：新型ゲーム機誕生の秘話を取材した。

語義4：人や動物が生まれること  
用例：子供の誕生は彼女の生活を一変させた。

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

# Vocabulary list

本システムについて テキスト情報 テキスト詳細 語彙リスト

語彙リスト

結果保存 (CSV : Shift-JIS) 結果保存 (CSV : UTF-8)

形態素数 (延べ) : 458 形態素数 (異なり) : 184

リンクをクリックすると辞書引きを行います。

出現順	基本形	読み	分類	出現順で並べ替え	読みで並べ替え	分類で並べ替え	頻度で並べ替え	語彙レベルで並べ替え
1	環境	カンキョウ	名詞-普通名詞-一般	3	0.66	環境 (3)		中級後半
2	省	ショウ	接尾辞-名詞的-一般	3	0.66	省 (3)		中級後半
3	が	ガ	助詞-格助詞	16	3.49	が (16)		
4	初めて	ハジメテ	副詞	4	0.87	初めて (4)		初級後半
5	取り組む	トリクム	動詞-一般	2	0.44	取り組ん (2)		中級後半
6	で	テ	助詞-接続助詞	2	0.44	で (2)		
7	いる	イル						
8	国	クニ						
9	の	ノ						
10	特別	トクベツ						
11	天然	テンネン						
12	記念	キネン						
13	物	ブツ						
14	,							

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27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 Japanese]. *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](mailto:yohasebe@gmail.com)



Jae-Ho LEE (李 在鎬), University of Tsukuba  
[jhlee.n@gmail.com](mailto:jhlee.n@gmail.com)

