Estimating text difficulty with machine learning

Mark Brierley
mark2@Shinshu-u.ac.jp
August, 2023

AI: Three predictions

- Development will not stop
- It will be better than us
- It will change the world

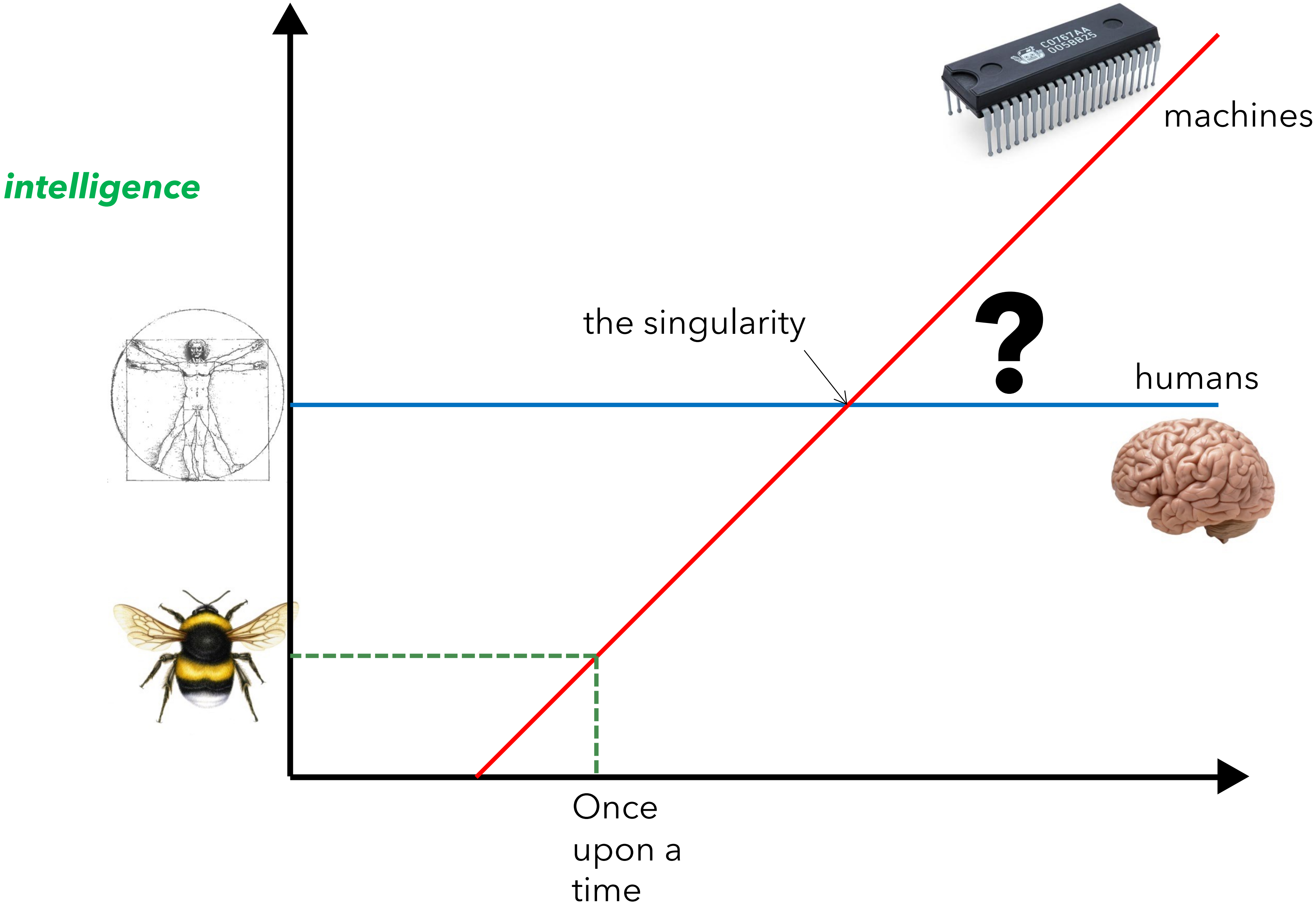
If you can measure it...

- AI will perform better

- “when you cannot measure it, when you cannot express it in numbers, your knowledge is of a meager and unsatisfactory kind”
  - Lord Kelvin

It will change the world

- The future is already here, it’s just not evenly distributed
  - William Gibson
- “Heavier than air flying machines are impossible”
  - Lord Kelvin (again), 1895



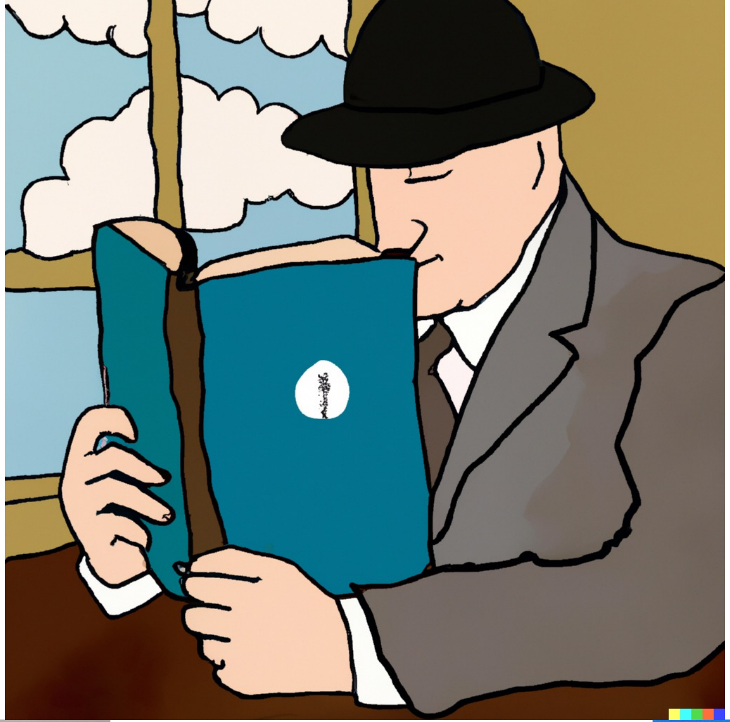
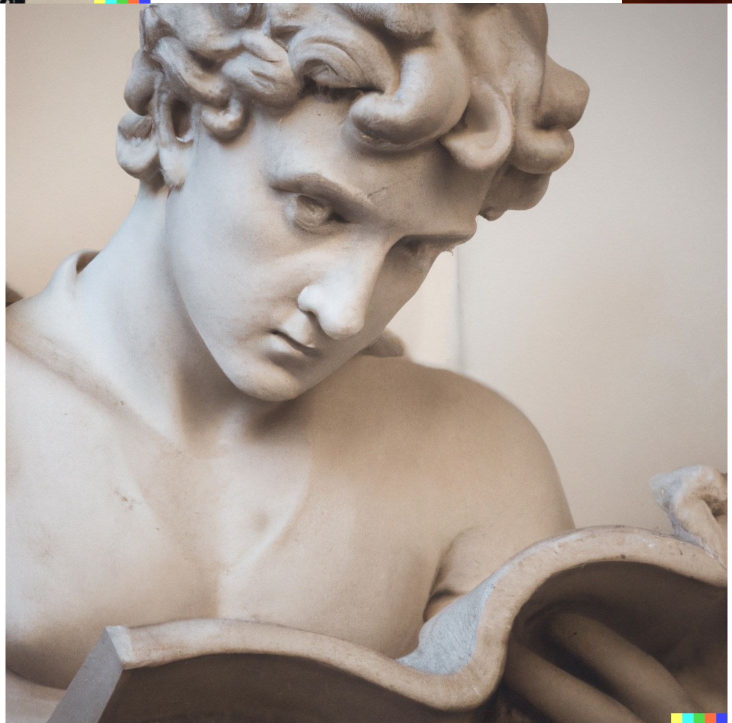
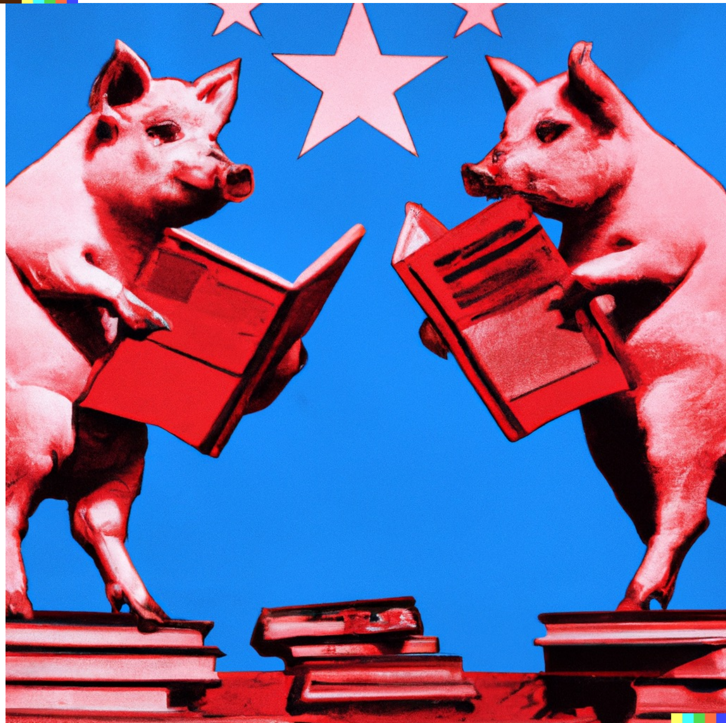
信州大学
SHINSHU UNIVERSITY

Threats and challenges for Language Teachers

- Short term:
  - How do we identify MT and ChatGPT?
  - How do we stop students using it?
- Medium term
  - How should students use it?
  - What lessons can we learn from AI?
- Long term
  - Where's the beach?

MT

- Freak cases
  - Entertaining!
- Epidemic
  - How do we detect it?
- Pandemic
  - How do we prevent it?
- Endemic
  - How do we live with it?

Lessons from MT and chat

- We don’t know all the rules
- We need loads of data
  - Millions of mysterious unknowable interactions

Extensive Reading

- Reading a lot of easy enjoyable books.

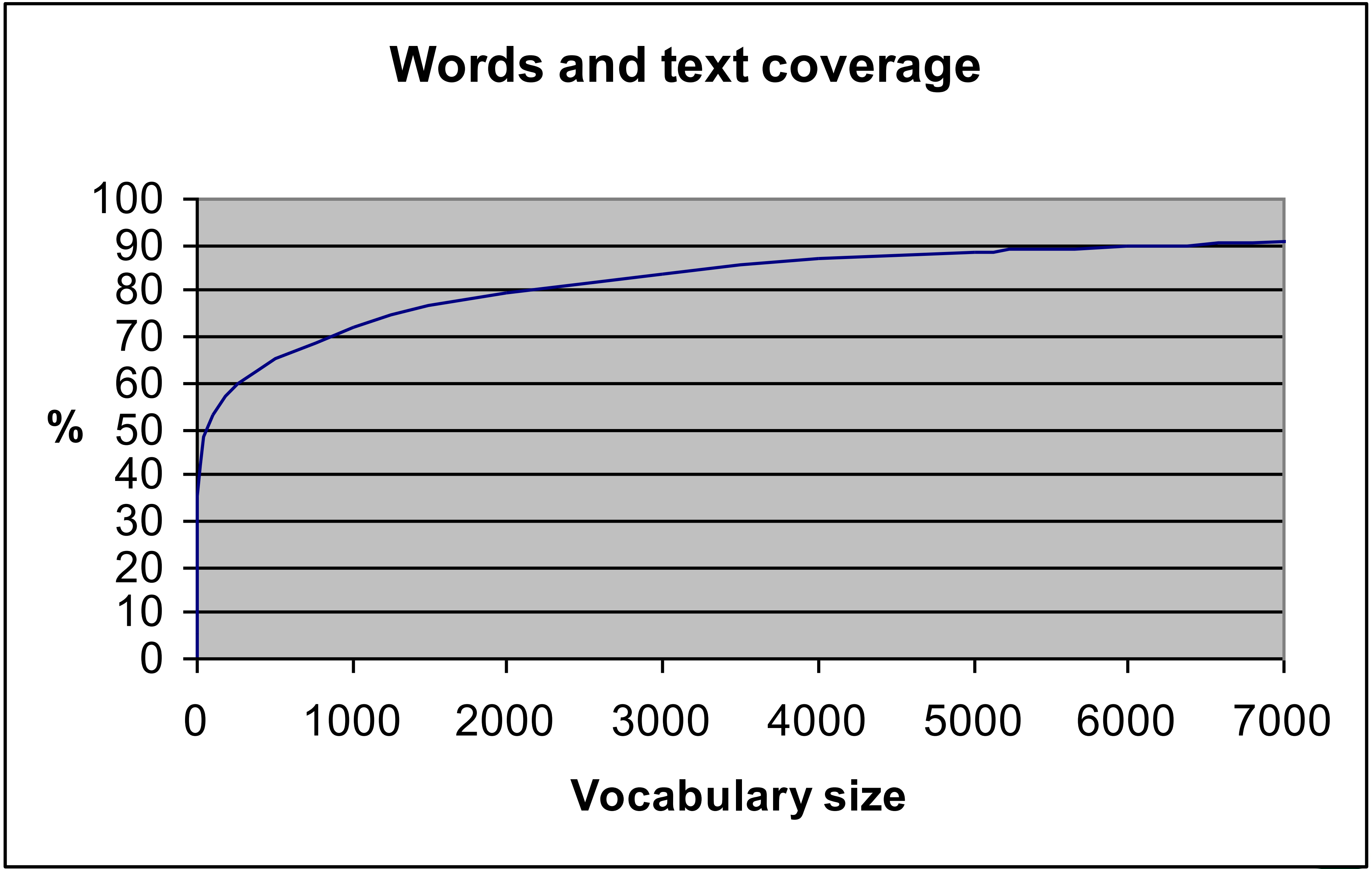
A lot?

- Read quickly
- >98% coverage
- Fluency practice

Requirements for ER

- Time
- Permission to enjoy reading
- Books at levels
  - Never enough reading material!

信州大学
SHINSHU UNIVERSITY

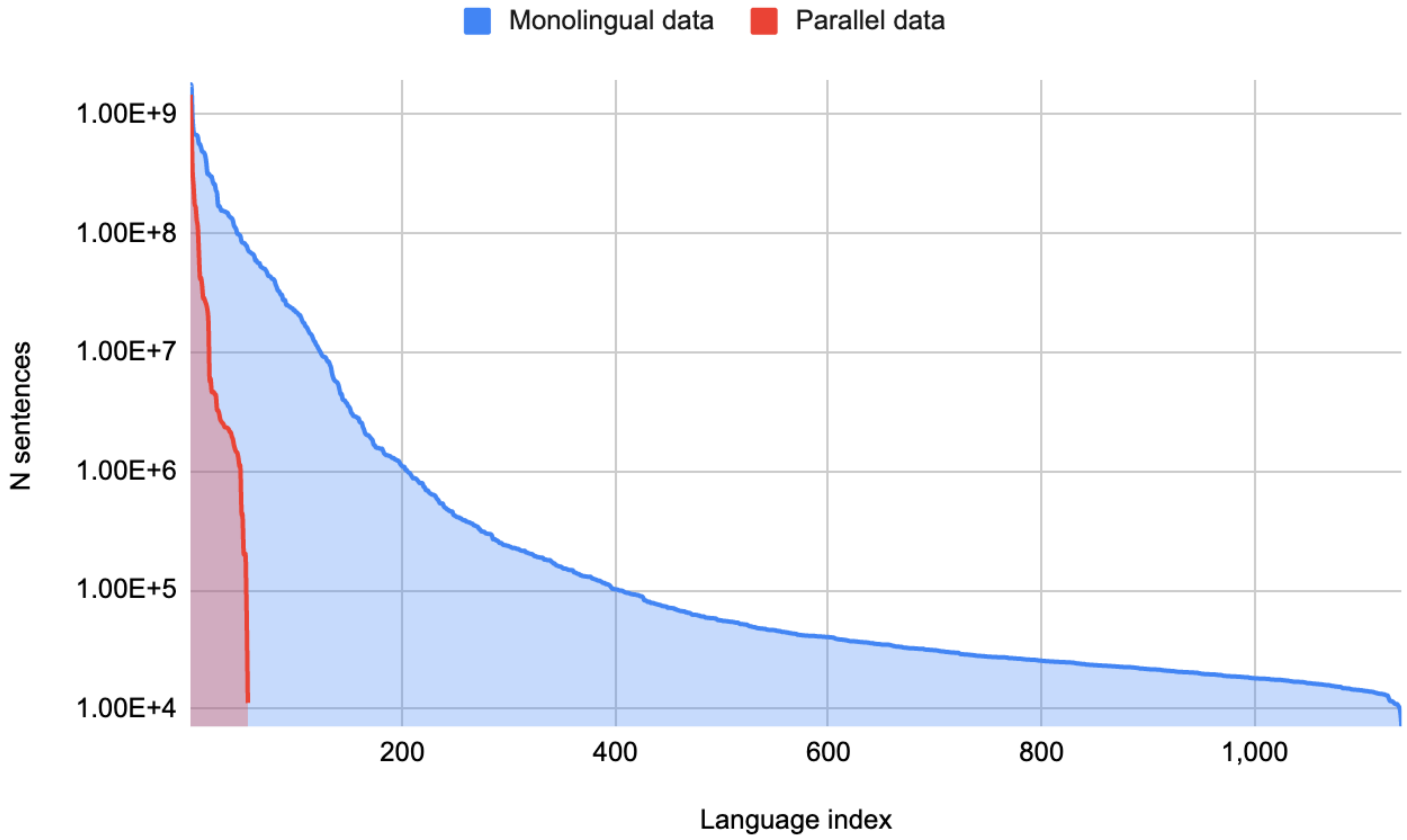


信州大学
SHINSHU UNIVERSITY

Medium-sized language models?

- Google translate: The next 1000 languages
  - Bapna et al. (2022)

信州大学
SHINSHU UNIVERSITY


Monolingual data
Parallel data
N sentences
1.00E+9
1.00E+8
1.00E+7
1.00E+6
1.00E+5
1.00E+4
200
400
600
800
1,000
Language index

UNIVERSITY

Two implications

- Nobody <u>needs</u> to learn a foreign language
  - Technology can perform all the functions

- Nobody <u>needs</u> to learn English
  - We can defeat the linguistic empire

信州大学
SHINSHU UNIVERSITY

MT 4 ER?

- Translation software
- Treat Levelled English as a language

Plan

- Identify "bilingual" texts
  - Texts at defined levels
- Build mono-lingual corpus
- Train translation software
- Trial with learners

Shinshu University ER research

- Language education and IT department

- ERS (online word counting system)
- ERF Placement test
- ER Cloud
- Machine learning to estimate text difficulty
- Machine translation to create levelled texts

信州大学
SHINSHU UNIVERSITY

Japan Grants-in-Aid for Scientific Research (Kaken)

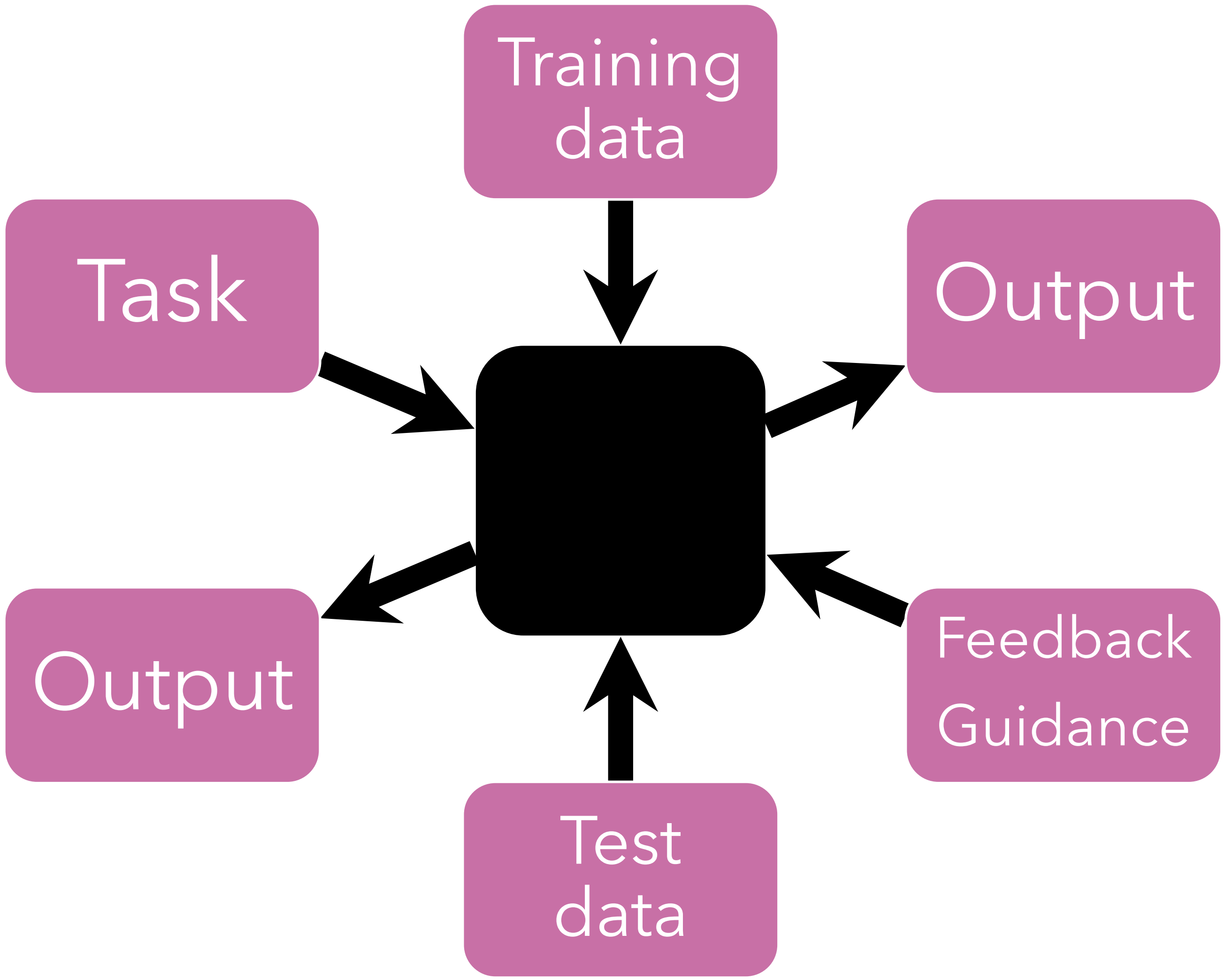
- 2009 Development of an Extensive Reading Support System based on information share between learners
- 2012 Development of a System for Recommending Graded Readers based on Estimates of Degree of Difficulty
- 2017 Online Systems to Support Extensive Reading
- 2020 Estimating Extensive Reading Text Difficulty Using Machine Learning
- 2023 Machine Learning to Simplify English for Extensive Reading

信州大学
SHINSHU UNIVERSITY

What is AI?

- What does “artificial” mean?
- What is “intelligence”?

What is machine learning?



信州大学 SHINSHU UNIVERSITY

Hara (2020) A machine learning method for estimating the difficulty of graded readers

- Focus on syntax
- Order of parts of speech
- Eg
- <noun> <conj> <noun> <conj> <noun>

Sakaguchi (2023) Proposal of a Difficulty Estimation Method for Extensive Reading of General Books in English

- Coh-metrics
  - Cohesion
- 106 parameters

Category	Parameters
Descriptive	Number of paragraphs, sentences, words, average length of paragraphs and sentences, average number of syllables per word, etc.
Text easability	Narrativity, syntactic simplicity, cohesion, z-scores for specific word types, z-scores for paradoxes, additions, comparative conjunctions, etc.
Referential cohesion	percentage of overlapping nouns, percentage of overlapping arguments, percentage of overlapping content words, etc.
Latent Semantic Analysis (LSA)	Mean of Cosine Similarity, Standard Deviation of Cosine Similarity, etc.
Lexical diversity	content word type token ratio, measure of textual lexical diversity (MTLD) of all words, etc.
Connectives	incidence of all conjunctions, incidence of causal conjunctions, etc.
Situation Models	Occurrences of causative verbs, occurrences of causative verbs and particles, etc.
Syntactic complexity	average of modifiers per noun, minimum edit distance of headwords, syntactic similarity, etc.
Density of syntactic patterns	incidence of noun phrases, verb phrases, adverb phrases, etc.
Word information	average age of acquisition of content words, average of familiar content words, etc.
Readability index	FRE, FKG, etc.

大学
NIVERSITY

Method

- Select books
- Idenfity parameters
- Linear regression

Graded readers and Project Gutenburg

- 164 books

Lasso Regression analysis

- Avoid over learning

- Explanatory variables
  - The obtained parameters
- Response variable
  - The YL

Indentify Parameters

- No significant difference in corellation with YL between Graded Readers and Gutenberg texts
- Parameter from each group with the strongest correlation with YL
  - Ignored Readability index
  - Ignored Lexical diversity

Best correlating parameters to YL

Group name	The parameters that had the strongest correlation with YL
Descriptive	Word count
Text Easability Principal Component Scores	Z score of adversative, additive, and comparative connectives
Referential Cohesion	The proportion of explicit content words that overlap between adjacent sentences
LSA	LSA overlap between adjacent sentences
Connectives	Causal connectives incidence
Situation Model	Causal verb incidence
Syntactic Complexity	Minimum edit distance score between adjacent sentences from lemmas
Syntactic Pattern Density	Verb phrase incidence
Word Information	Mean of familiarity for content words

信州大学 SHINSHU UNIVERSITY

Training data and test data

Text data: 164 books

80%

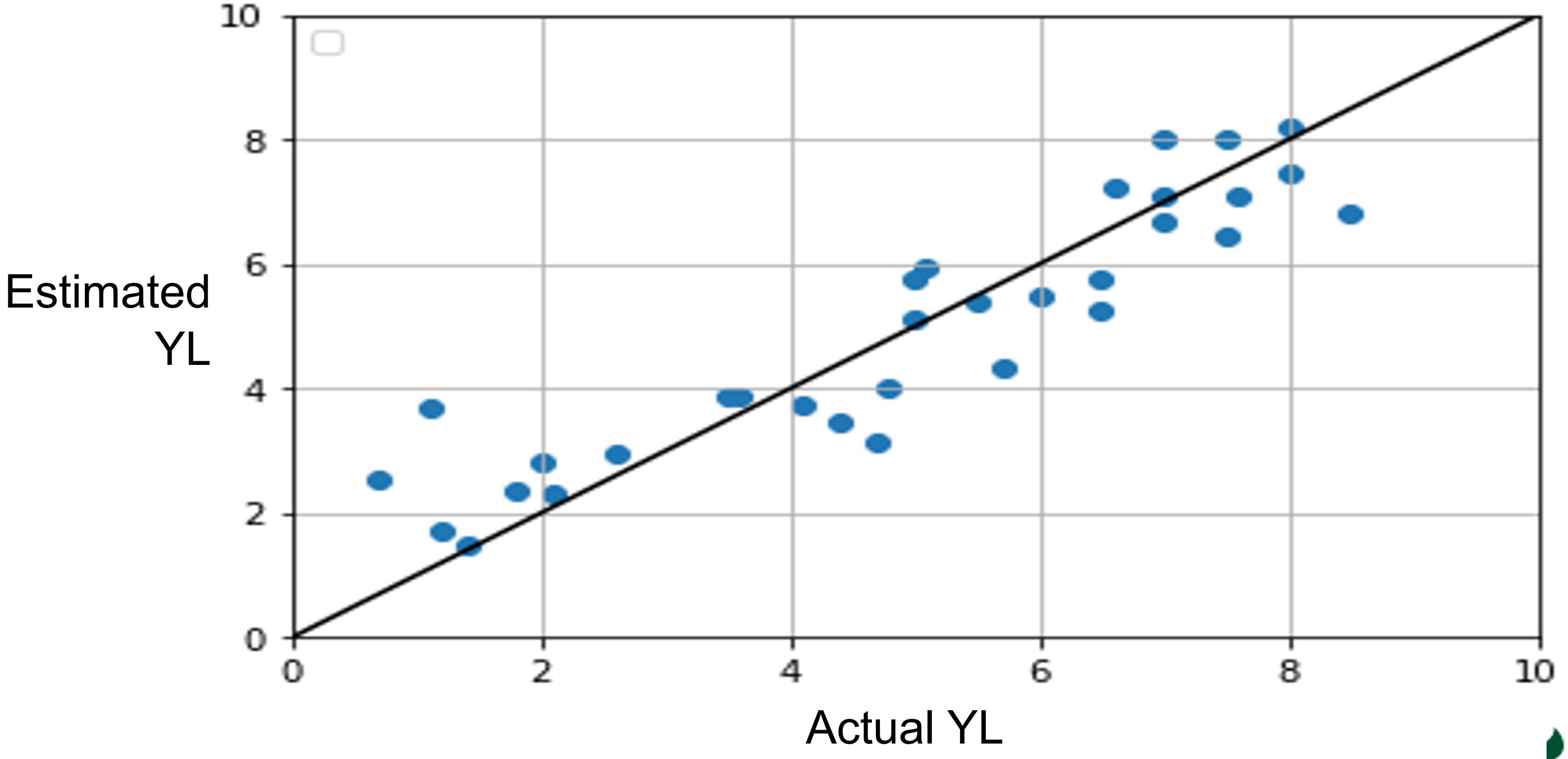
20%

Training data

Test data

信州大学
SHINSHU UNIVERSITY

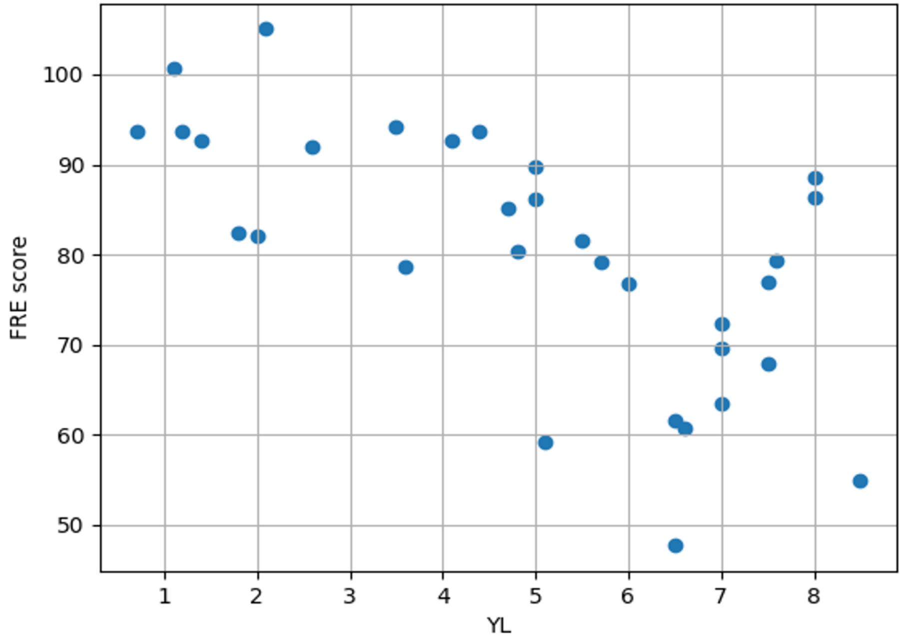
Result



Comparison with Flesch Reading Ease (FRE)

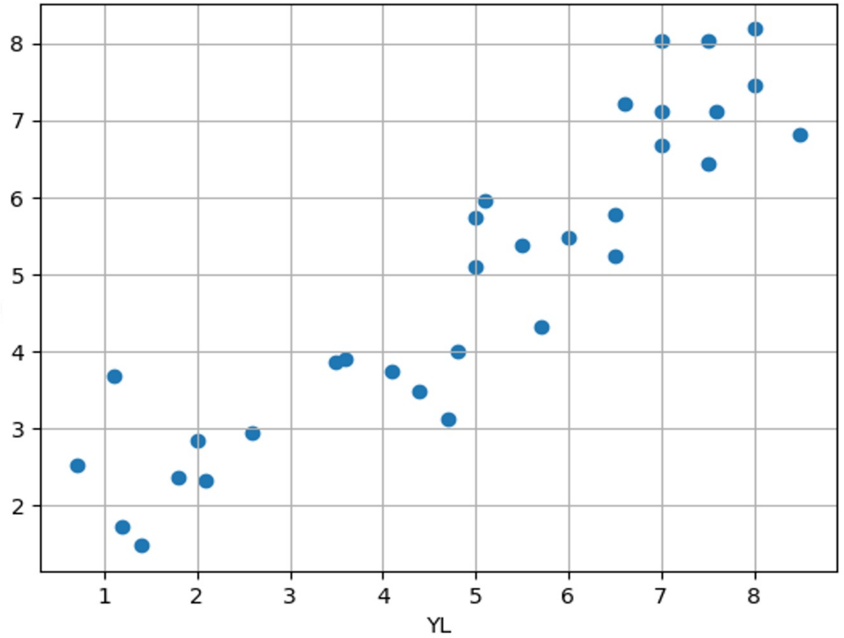
- The average number of words in a sentence
- The average number of syllables in a word

Comparison with existing difficulty estimation methods



FRE

Correlation coefficient：**-0.650**



Our estimated YL

Correlation coefficient：**0.917**

信州大学
SHINSHU UNIVERSITY

Parameters and correlation coefficient with YL

Parameters with strongest correlation with YL	Correlation coefficient
Word count	0.800
Z score of adversative, additive, and comparative connectives	0.672
proportion of explicit content words that overlap between adjacent sentences	-0.571
LSA overlap between adjacent sentences	0.094
Causal connectives incidence	0.584
Causal verb incidence	0.590
Minimum edit distance score between adjacent sentences from lemmas	0.591
Verb phrase incidence	-0.660
Mean of familiarity for content words	-0.848

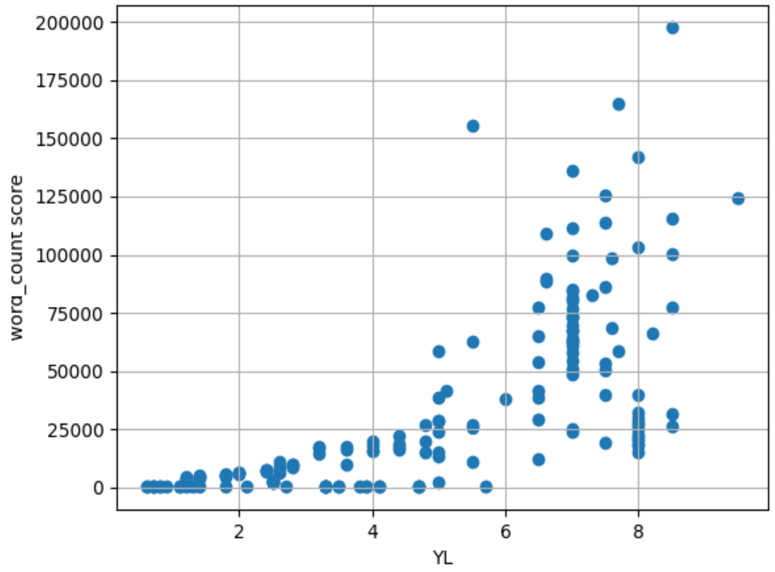
信州大学
SHINSHU UNIVERSITY

Further research

- Improvement of parameters selection method
- Consideration of parameters that were not used

Total number of words

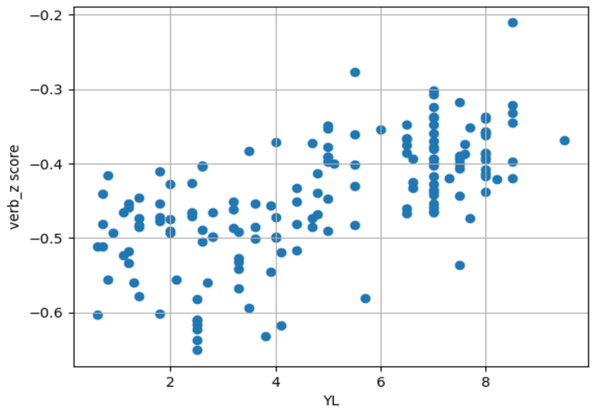
- Strongest correlation
- See Holster, Lake and Pellowe (2017)



信州大学
SHINSHU UNIVERSITY

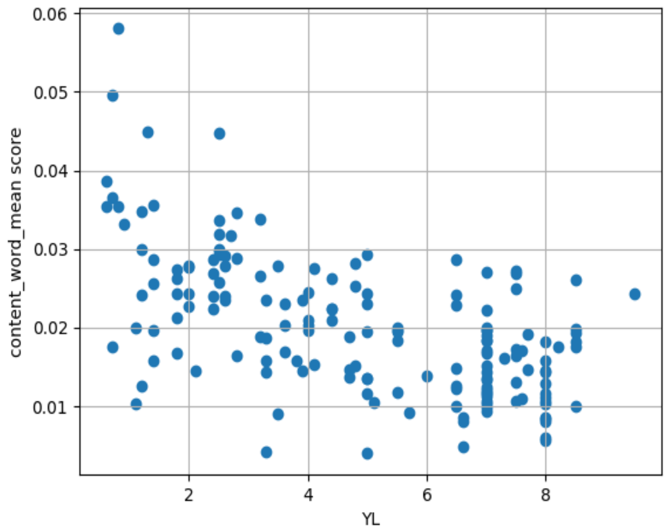
Z-scores for paradoxical, appositional, and comparative conjunctions

- the extent to which paradoxical (but, however, etc.), additional (and, moreover, etc.), and comparative (although, whereas, etc.) conjunctions are used in a text compared to the mean for other parts of speech.
- increases as YL increases (Figure 2)



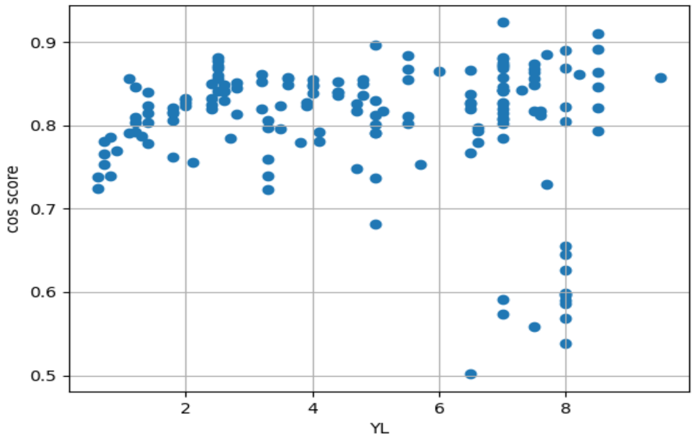
Content word overlap

- the extent to which the same content words are used in adjacent sentences
- decreases as YL increases



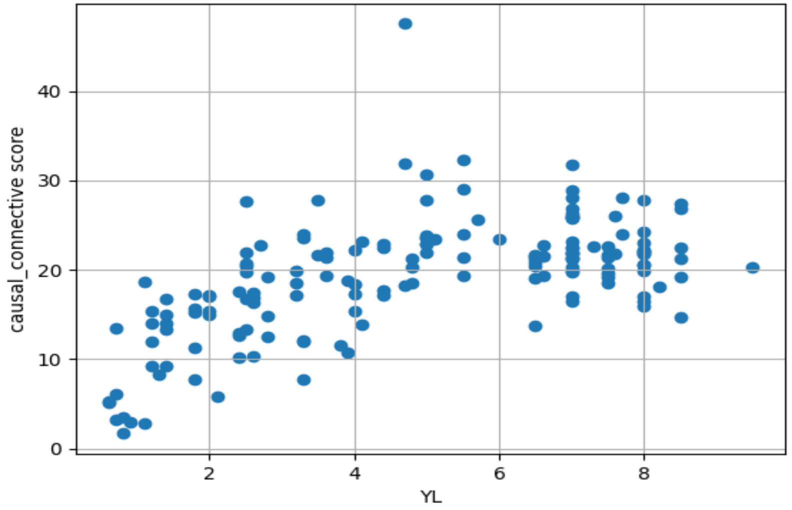
Cosine Similarity in Adjacent Sentences

- how conceptually similar each sentence is to the next sentence.
- The correlation with YL is very weak.



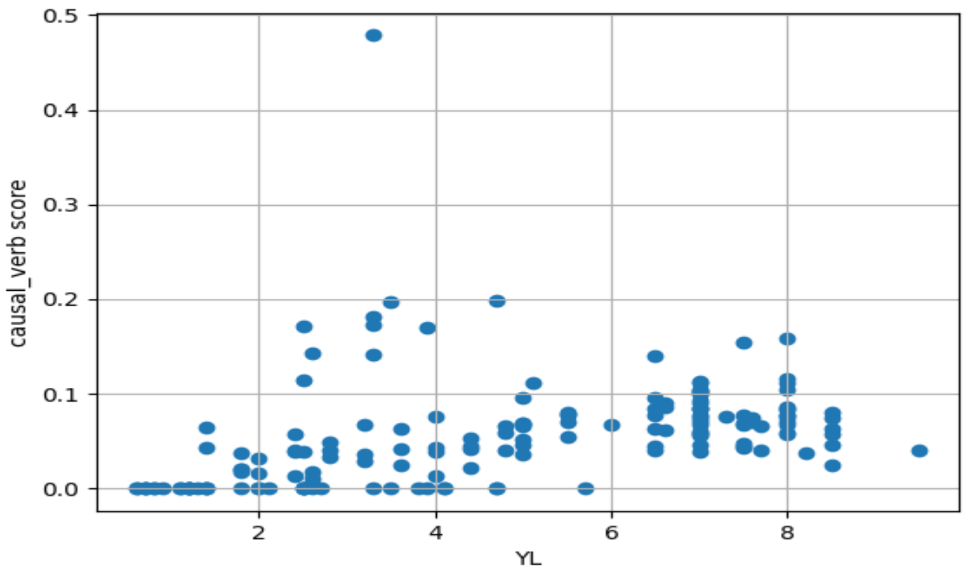
Occurrence of causal connectives

- percentage of causal connectives (e.g., because, since) among all parts of speech.
- Increases as YL increases (Figure 5).



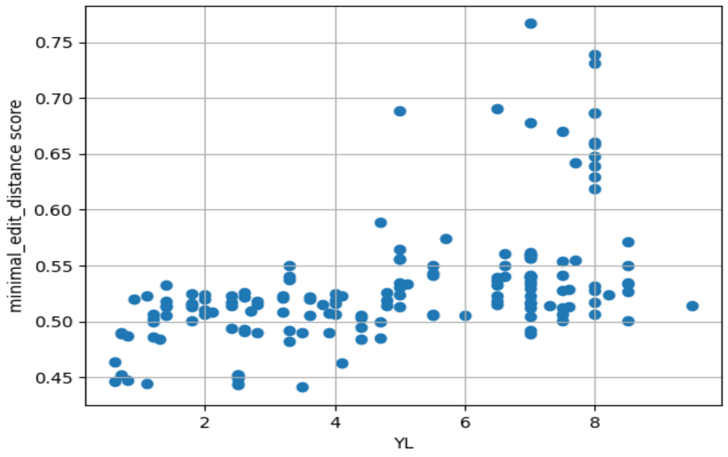
Percentage of causative verbs

- percentage of causative verbs (result, lead, etc.) among all parts of speech.
- increases as YL increases



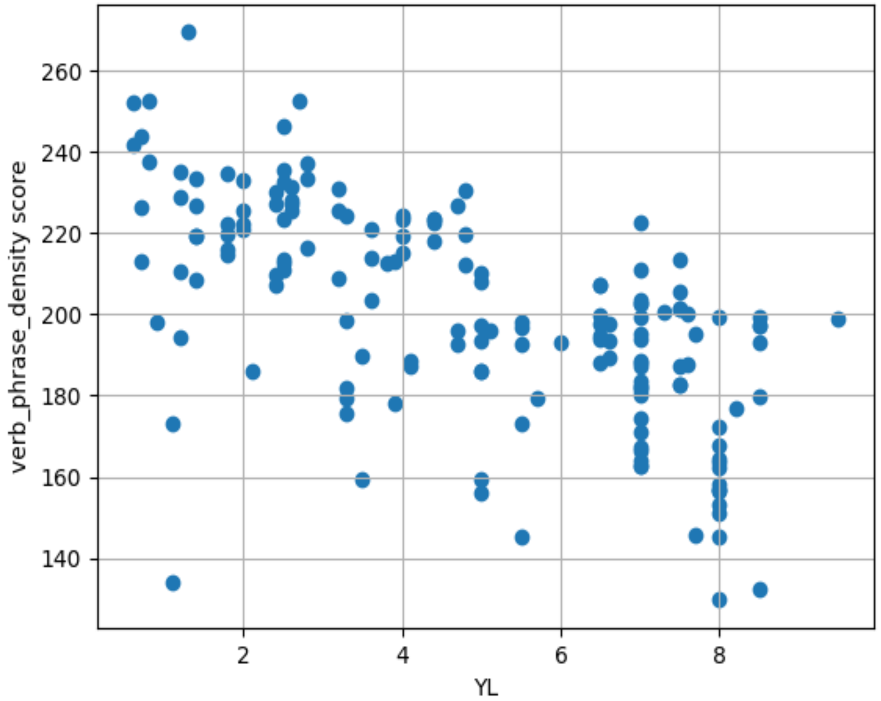
Minimum edit distance of adjacent sentences by Lemma

- The edit distance:
  - how different two strings of words are,
  - number of edits that must be made to convert one string into another.
- Lemma
  - dictionary form of a word.
- Increases with with YL



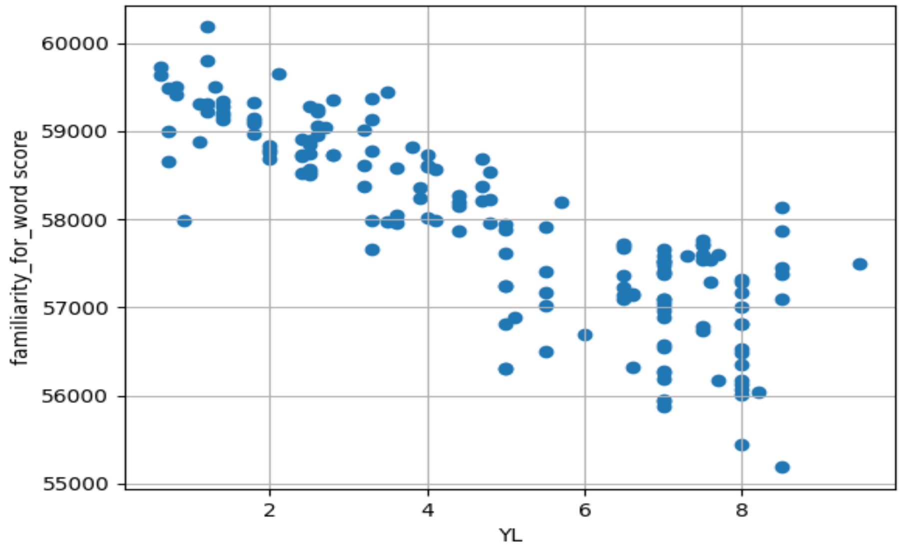
Verb Phrase Occurrence Rate

- Strongly correlated with YL
- decreasing as YL increases



Average number of familiar content words

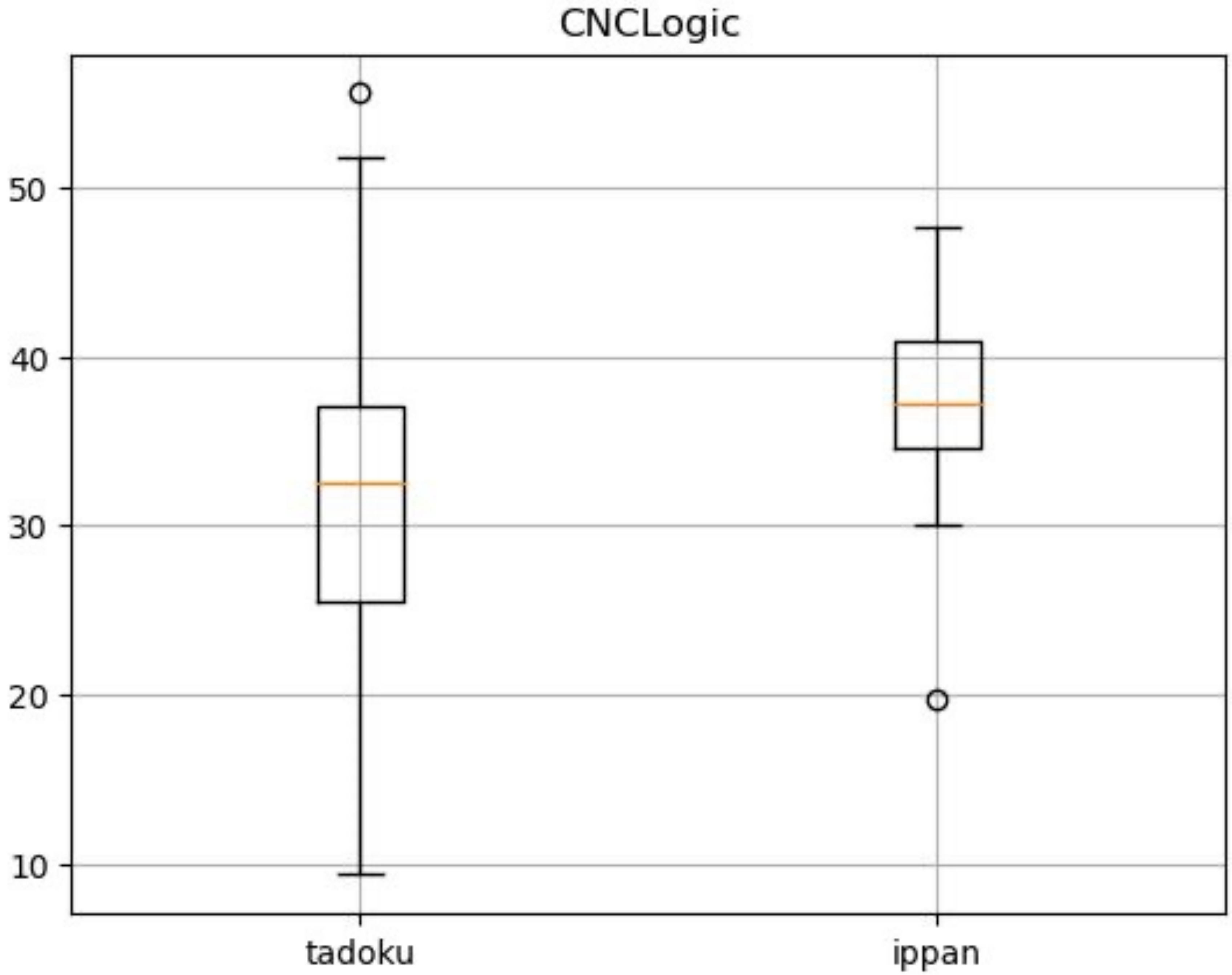
- MRC Psycholinguistic Database
- lower values for unfamiliar words and higher values for frequently seen words.
- The correlation very strong
- Decreases with YL

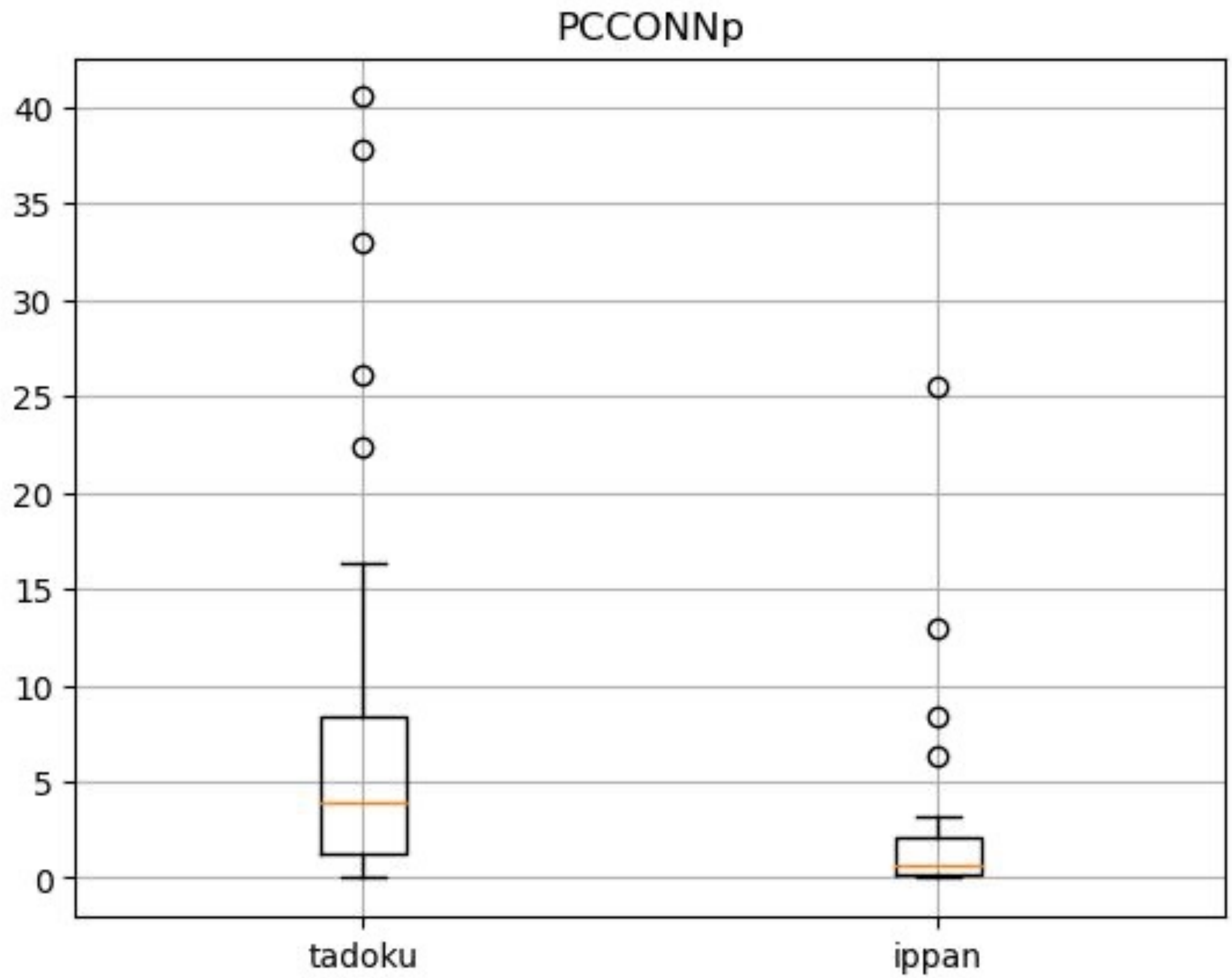


Current research

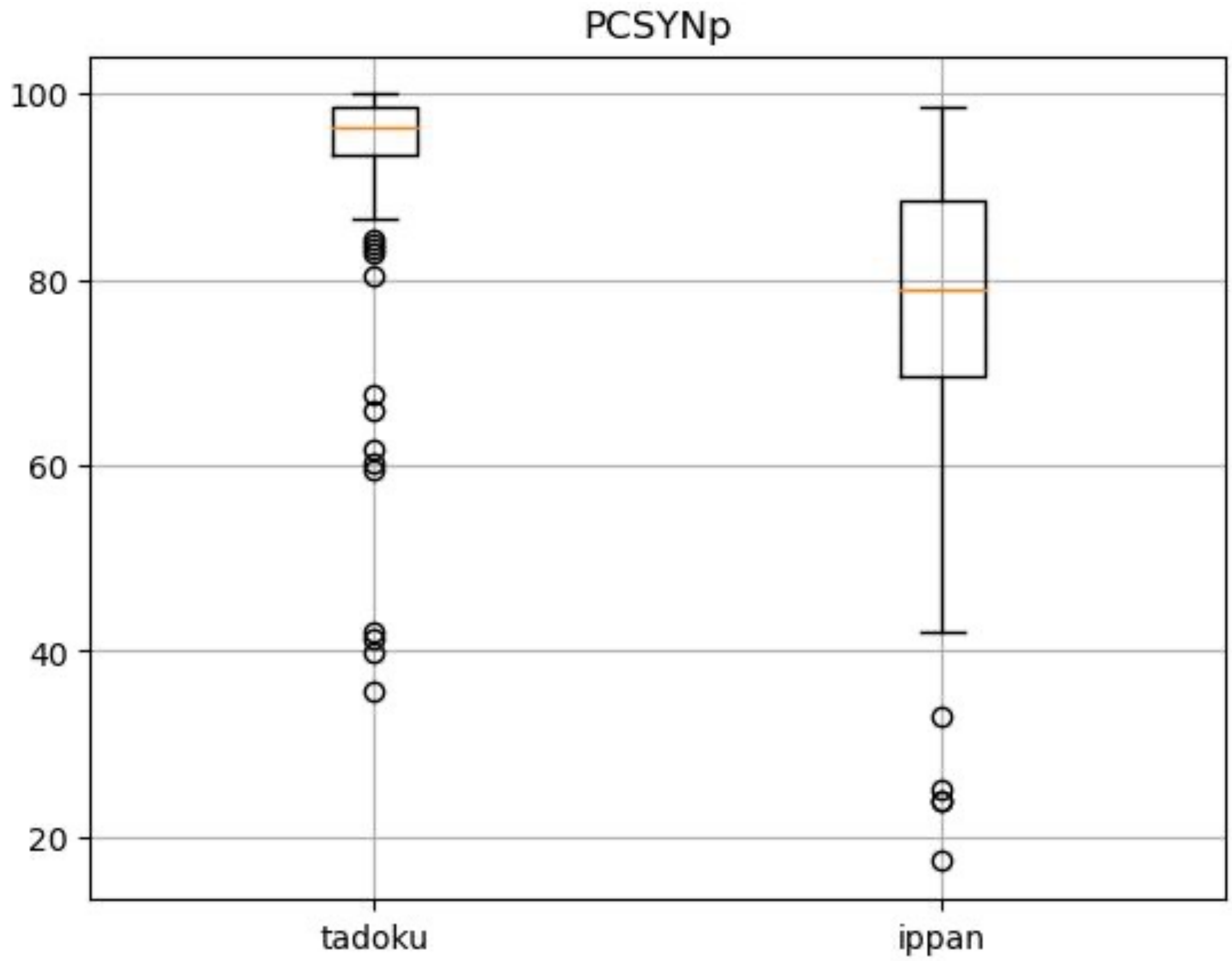
- What is different between graded readers and “authentic” texts?
- What parameters can best tell the difficulty of text?

信州大学
SHINSHU UNIVERSITY

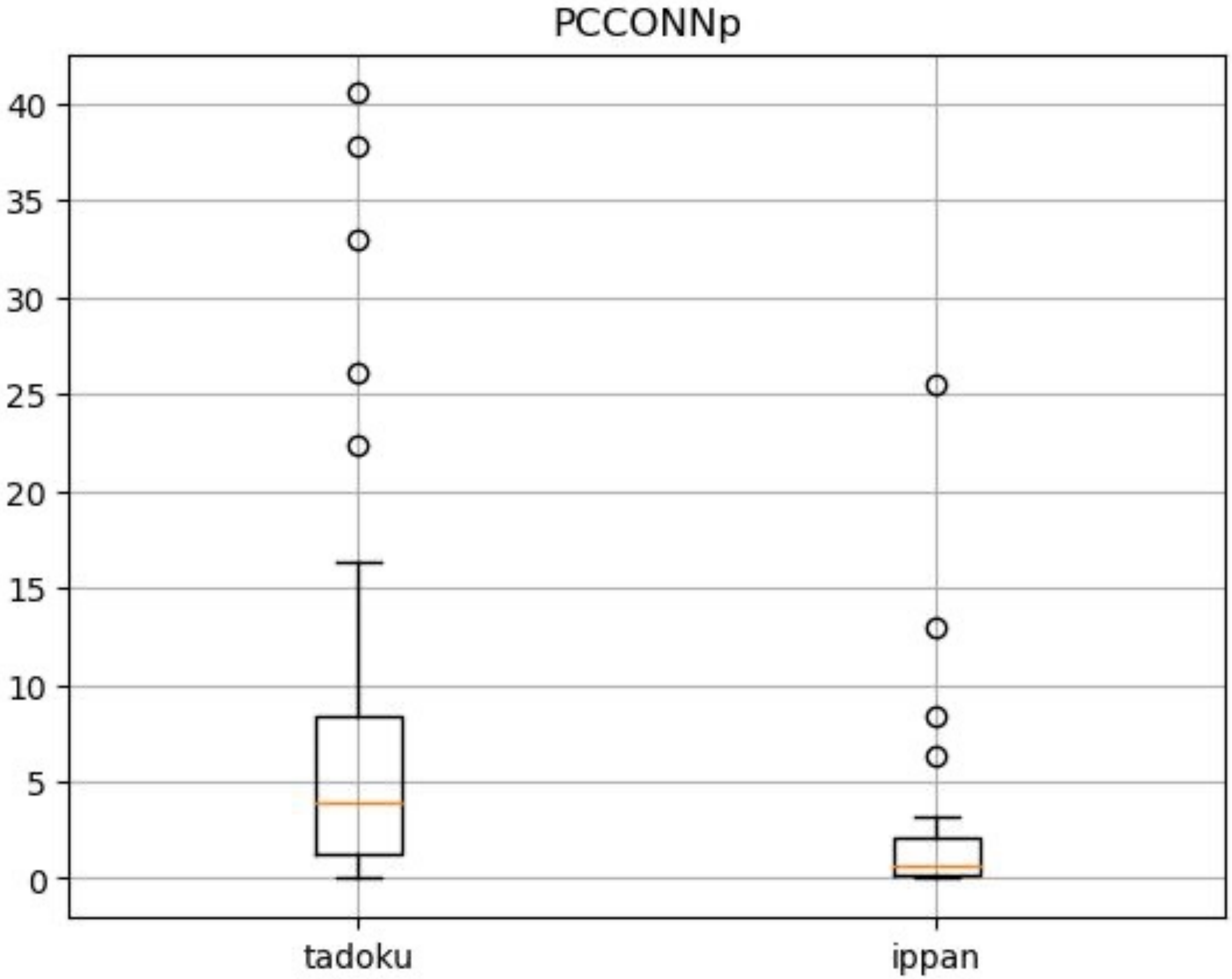

CNCLogic
50
40
30
20
10
tadoku
ippan



信州大学
SHINSHU UNIVERSITY



信州大学
SHINSHU UNIVERSITY

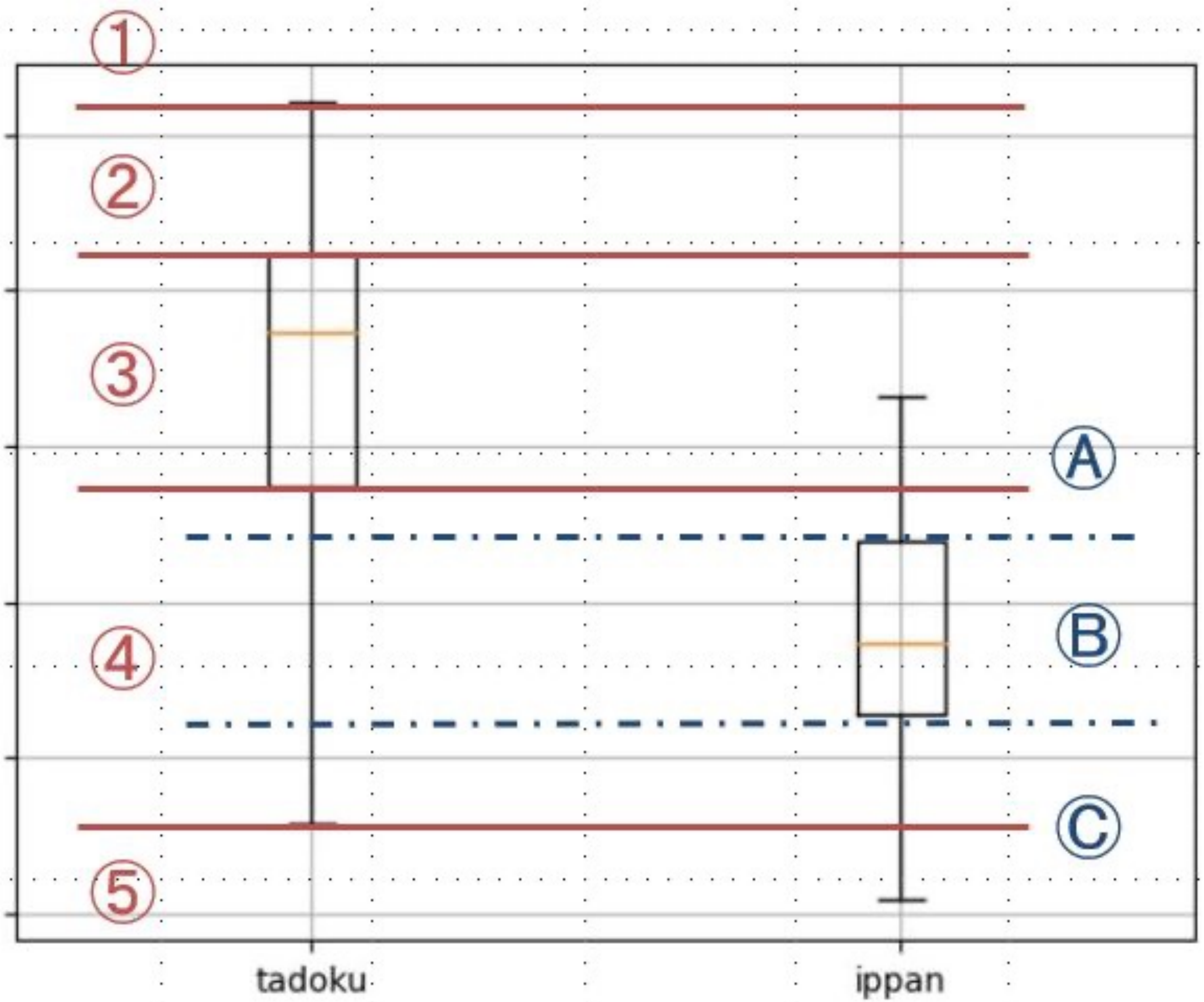


信州大学
SHINSHU UNIVERSITY

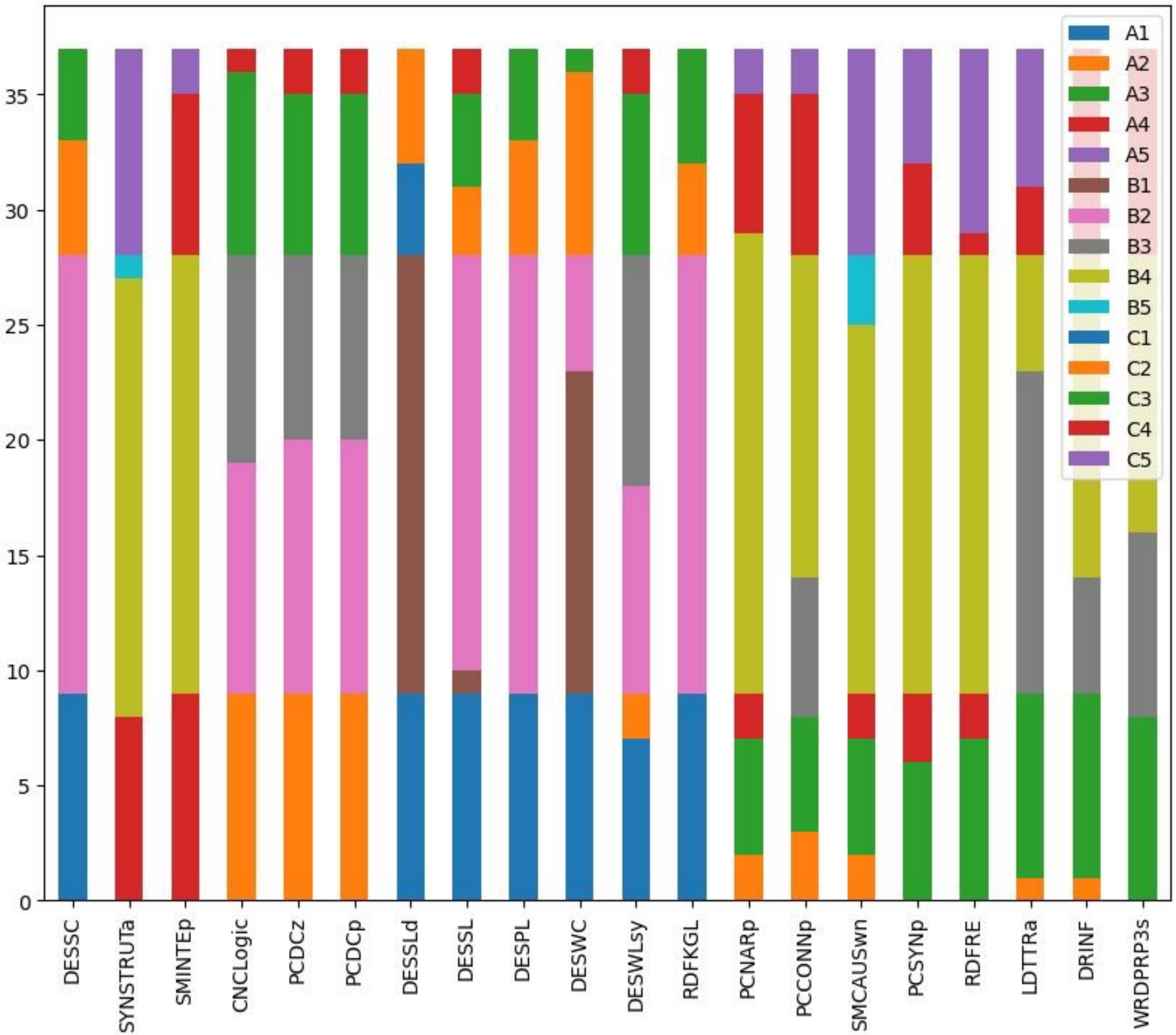
Parameters (104)

Similar average **Similar spread (13)**	**Different average** **Similar spread (16)**
Similar average Different spread (39)	Different average Different spread (36)

信州大学
SHINSHU UNIVERSITY



信州大学
SHINSHU UNIVERSITY



信州大学
SHINSHU UNIVERSITY

Conclusion

- AI is not our enemy
- It can solve our problems
- It can help provide learning texts
- It can promote "minor" languages

Detail of parameters in each groups

The result of selecting new parameters that contribute to YL

Group name	The parameters that had the strongest correlation with YL	Description
Descriptive	Word count	The total number of words in the text
Text Easability Principal Component Scores	Z score of adversative, additive, and comparative connectives	$\text{Z score} = \frac{(\text{Number of connectives - Population mean})}{\text{Population standard deviation}}$
Referential Cohesion	The proportion of explicit content words that overlap between adjacent sentences	The proportion of content words (Words that describe content, like nouns, verbs, adjectives, and adverbs)
LSA	LSA overlap between adjacent sentences	LSA is the degree of similarity words and words, words and documents, and documents and documents

信州大学
SHINSHU UNIVERSITY

Detail of parameters in each groups

The result of selecting new parameters that contribute to YL

Group name	The parameters that had the strongest correlation with YL	Description
Connectives	Causal connectives incidences	The proportion of causal connectives（ ”because", “since", “as” and so on ）
Situation Model	Causal verb incidences	The proportion of causal verb ("result", "lead", "bring" and so on)
Syntactic Complexity	Minimum edit distance score between adjacent sentences from lemmas	It is a way of quantifying how dissimilar two strings are to one another
Syntactic Pattern Density	Verb phrase incidence	The proportion of verb phrase
Word Information	Mean of familiarity for content words	This is a rating used MRC Psycholinguistic Database of how familiar a word seems to an adult

信州大学
SHINSHU UNIVERSITY

Detail of each groups

Group name	Description	example of Parameters
Descriptive	This group helps to confirm the output of Coh-Metrix and also to interpret patterns in the data.	Word count, Sentence count, Paragraph length, …
Text Easability Principal Component Scores	The group provides a more complete picture of the textual ease that results from the linguistic characteristics of the text.	Percentile of syntactic simplicity, Z score of syntactic simplicity, Z score of connectives, …
Referential Cohesion	This group refers to overlap in content words between local sentences, or co-reference.	Noun overlap, Argument overlap, Content word overlap, …

信州大学 SHINSHU UNIVERSITY

De

Group name	Description	example of parameters
LSA	This group provides measures of semantic overlap between sentences or between paragraphs.	LSA overlap between adjacent sentences, LSA overlap between all sentences, LSA overlap between adjacent paragraphs, ...
Lexical Diversity	This group refers to the variety of unique words (types) that occur in a text in relation to the total number of words (tokens).	Type token ratio for all words, MTLD lexical diversity measure, VOC lexical diversity measure, ...
Connectives	This group plays an important role in the creation of cohesive links between ideas and clauses and provide clues about text organization.	All connectives incidence, Causal connectives incidence, Logical connectives incidence, ...
Situation Model	The expression Situational Model is a cognitive science that refers to the level of mental representation for a text.	Causal verb incidence, Intentional verbs incidence, WordNet verb overlap, ...

信州大学 SHINSHU UNIVERSITY

Detail of each groups

Group name	Description	example of parameters
Syntactic Complexity	Theories of syntax assign words to part-of-speech categories, group words into phrases or constituents, and construct syntactic tree structures for sentences.	Number of modifiers per noun phrase, Minimum edit distance score between adjacent sentences from lemmas, Sentence syntax similarity, ...
Syntactic Pattern Density	This group provides information on the incidence of noun phrases, verb phrases, adverbial phrases, and prepositions.	Noun phrase incidence, Verb phrase incidence, Adverbial phrase density,...
Word Information	This group computes word frequency scores and psychological ratings.	Noun incidence, First person singular pronoun incidence, Mean of familiarity for content words, …
Readability	This group consists of existing difficulty estimation methods.	Flesch Reading Ease (FRE), Flesch-Kincaid Grade Level (FKG), Coh-Metrix L2 Readability, ...

信州大学
SHINSHU UNIVERSITY

Recent developments in AI chat are sending shockwaves through the language teaching community, both with short-term challenges of instructing students when and how to use this technology and as a longer-term existential threat to the teaching vocation. On the other hand, this same technology presents an opportunity for the automatic production of compelling input, not only in English but potentially for many other languages. Critical to providing suitable input is determining the level of readability, for example measured in YL (Yomiyasusa Level), which is based on impressions of difficulty by readers in Japan. This presentation reports on research into machine learning techniques used to estimate YL using the Coh-metrix analysis tool, Lasso linear regression and grid search cross-validation. The model predicted YL with a strong correlation of .91, significantly better than the Flesch Reading index. The results suggest that the developed model is a promising tool for predicting YL.

信州大学
SHINSHU UNIVERSITY