How is the relationship between vocabulary knowledge and reading comprehension?

A case of Japanese

Tatsuhiko Matsushita
University of Tokyo
matsushita@global.c.u-tokyo.ac.jp

AILA World Congress 2014
Brisbane, 15 August 2014

Contents

1. Introduction: Motives for the Research

When developing reading materials for learners of Japanese, I found:

- many English reading materials where the lexical level is systematically controlled for a certain purpose, e.g. graded readers, but few Japanese ones
- many studies for European languages which underpin the rationales for creating those reading materials, but few Japanese studies

How should we control the lexical level of reading texts in Japanese? What level of texts could be used for learners at what level?

2. Previous Studies

◆Importance of vocabulary in reading

- Many studies, but all in all, vocabulary accounts for approx. 30% in European languages（Bernhardt, 2005）
- Higher reliance on vocabulary in reading Japanese?

55%　calculated from Koda (1989)

47%　Komori, Mikuni & Kondoh (2004)

40%+　calculated from Noguchi (2008)

← Learning burden of Kanji (logographic characters)?
40% of texts (2/3 of content word tokens) are covered by Kanji words (Matsushita, 2012a)

◆ How high text coverage is needed for ‘adequate’ reading comprehension?

- 95％：Minimal ‘threshold’ for reading with some guidance and leading to independent reading
- 98％：Optimal ‘threshold’ for independent reading
  （Laufer & Ravenhorst-Kalovski, 2010, endorsing Hu & Nation, 2000）
- No vocabulary ‘threshold’ (Schmitt, Jiang & Grabe, 2011)

Lower ‘threshold’ in Japanese?

- Around 96% (Komori et al, 2004)
- Many highly-semantically-transparent Kanji compounds may get the threshold lower（Matsushita, 2012a）
- But, many more words are required to attain the same coverage level as English (Lexically more diverse in Japanese with very limited number of Kanji i.e. word components)

  e.g. 9500 lemmas for 95% text coverage

  more than 20,000 lemmas for 98% coverage

3. Research Questions (and Methods)

1) How much does vocabulary knowledge account for L2 reading comprehension in Japanese?

← Regression analysis

(r^2 : Coefficient of Determination)

2) Does the threshold level of vocabulary known in a text exist (in reading Japanese texts)?

← Polynominal approximation

(to see the curve fitting by checking r^2)

3) What level of vocabulary size will assure what level of reading comprehension?

← Regression formula

4) What level of vocabulary size will assure what level of text coverage by known words?

← Estimation of cumulative text coverage

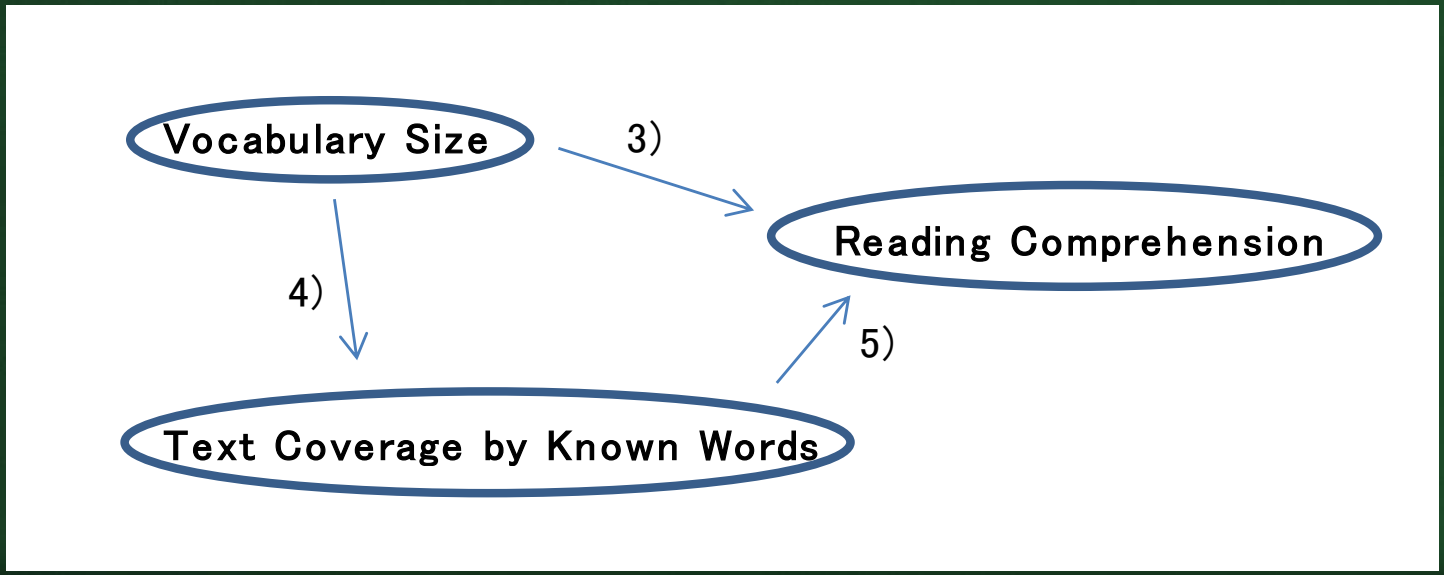
= Represented text coverage by each item

× the correct answer ratio for each item by learner level

5) What level of text coverage by known words will assure what level of reading comprehension?

← combining 3) + 4)

Similar to the viewpoints of Laufer & Ravenhorst-Kalovski (2010), being applied to Japanese



4. Tests

- Tests
  - Vocabulary Size Test for Reading Japanese (VSTRJ) (Matsushita, 2012b)

    Similar to English VST (Nation & Beglar, 2007)
  - Japanese Reading Comprehension Test (JRCT)

    6 questions × 4 passages = 24 questions

    Each passage is controlled at 98% coverage by top 2K, 6K, 10K 15K words

    Each passage consists of the same number (2) of the 3 different types of questions i.e. Replication, Analysis, Inference (Yano, Long & Ross, 1994)

* Both VSTRJ and JRCT are developed based on the same vocabulary database (VDRJ) (Matsushita, 2011)

Participants

- *n* = 213 (M:F=89:124; Age: M=22y , SD=4y)
- Learners at universities and language schools in Australia, New Zealand and Japan
- Chinese-background learners (CBLs): non-CBLs = 118:83 (14 types of L1)

5. Results

Descriptive Statistics for VSTRJ & JRCT (Full ver.)

	Max. point	*M*	*SD*	*n*
VSTRJ full version (Vocabulary)	150	84.4	39.7	213
JRCT full version (Reading)	24	12.0	5.45	213

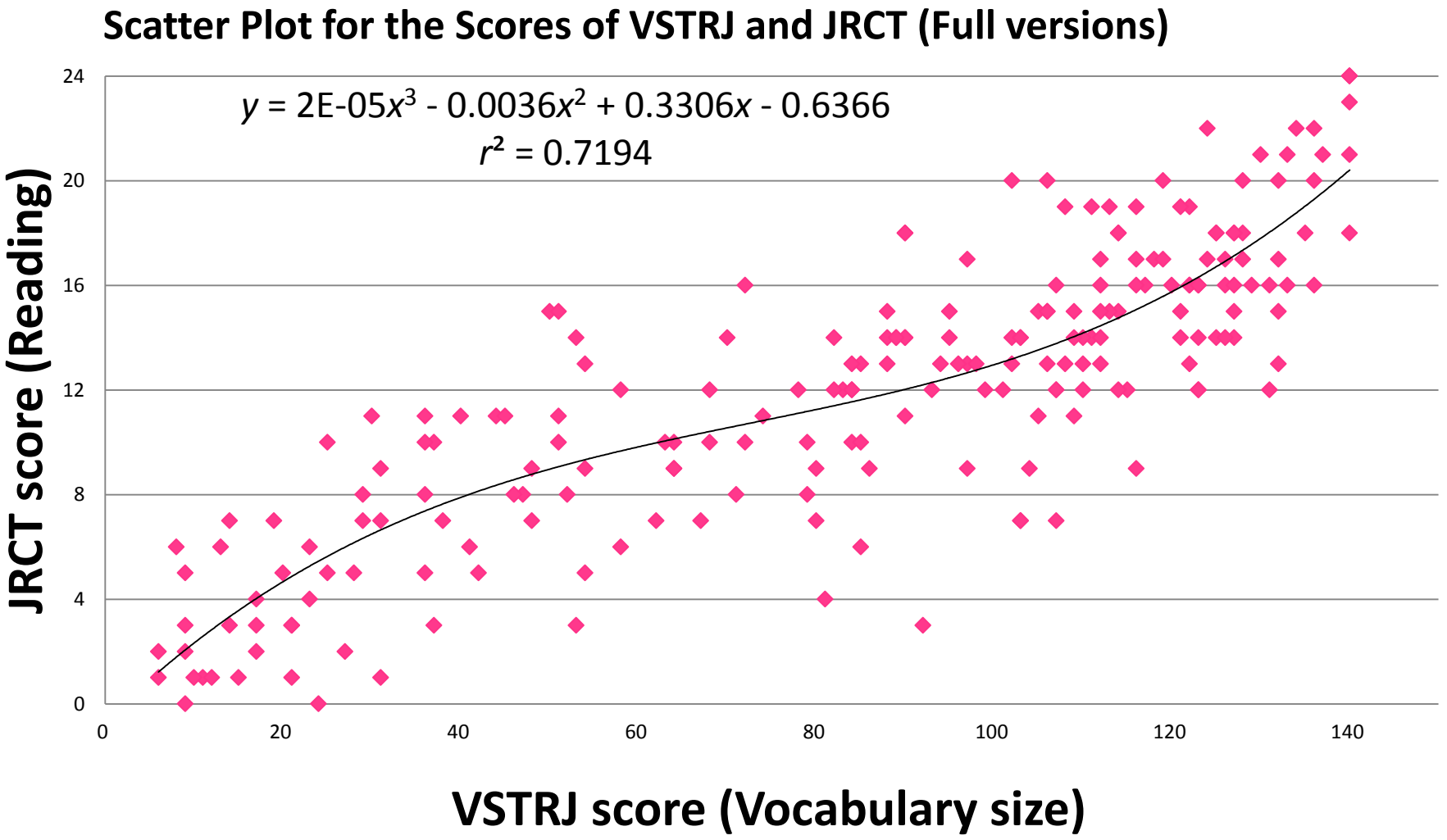
- Correlation (Pearson) between VSTRJ and JRCT

  $r = .836$ ($p < .001$)

- Regression analysis for polynomial approximation (curve fitting) to predict the JRCT score (y) from the VSTRJ score (x)

  $y = 2E\text{-}05x^3 - 0.0036x^2 + 0.3306x - 0.6366$

  $r^2 = 0.7194$


Scatter Plot for the Scores of VSTRJ and JRCT (Full versions)
y = 2E-05x3 - 0.0036x2 + 0.3306x - 0.6366
r² = 0.7194
JRCT score (Reading)
VSTRJ score (Vocabulary size)
0
4
8
12
16
20
24
0
20
40
60
80
100
120
140

Means, standard deviations and correlations for JRCT questions with texts at different lexical levels and differently sized VSTRJ (2K, 6K, 10K, 15K)

	Variables	Max. points	1	2	3	4	5	6	7	8	9
1	JRCT (2K)	6	--								
2	JRCT (6K)	6	.495***	--							
3	JRCT (10K)	6	.424***	.602***	--						
4	JRCT (15K)	6	.452***	.567***	.515***	--					
5	JRCT Total	24	.745***	.841***	.802***	.791***	--				
6	VSTRJ (2K)	20	.680***	.666***	.597***	.568***	.790***	--			
7	VSTRJ (6K)	60	.625***	.703***	.678***	.622***	.827***	.937***	--		
8	VSTRJ (10K)	100	.612***	.714***	.699***	.647***	.841***	.906***	.989***	--	
9	VSTRJ (15K)	150	.605***	.715***	.691***	.646***	.836***	.881***	.972***	.990***	--
	M		4.18	2.97	2.55	2.27	11.97	16.20	42.77	63.50	84.43
	SD		1.71	1.75	1.74	1.65	5.45	4.32	16.35	27.72	39.66

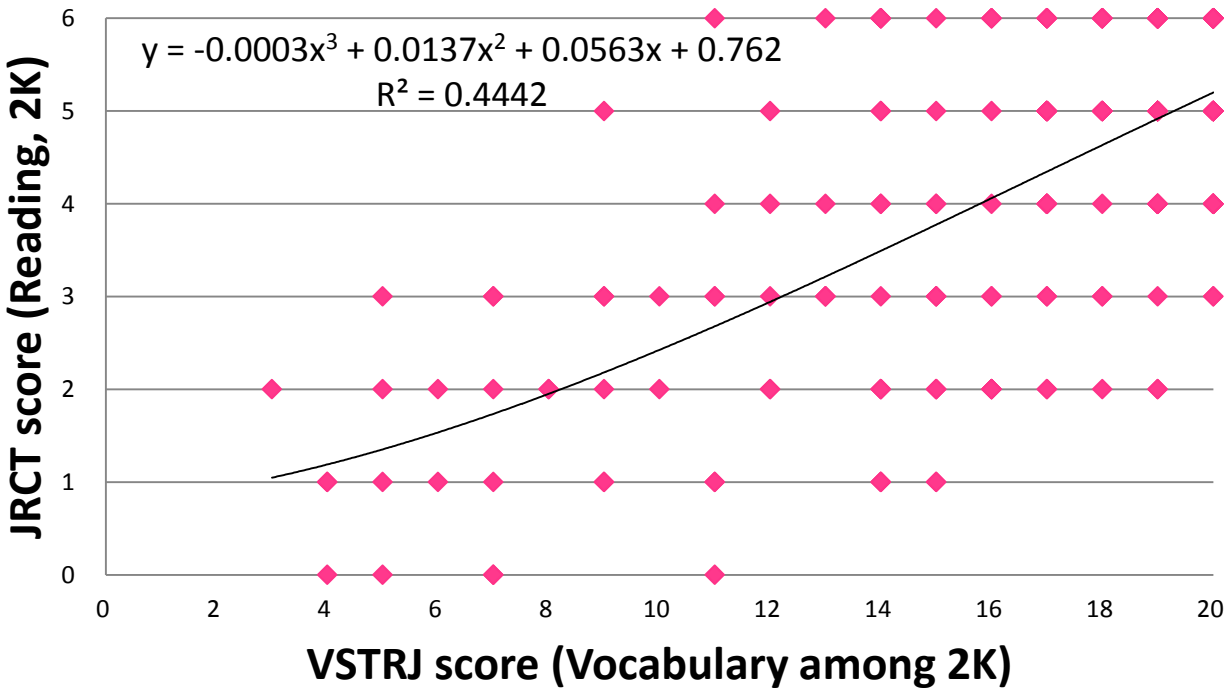
Note: N=213 All correlations are statistically significant (*** p<.001)

Descriptive Statistics for VSTRJ (2K) & JRCT (2K)

	Max. point	*M*	*SD*	*n*
VSTRJ (2K) (Vocabulary)	20	15.8	4.1	194
JRCT (2K) (Reading)	6	4.3	1.7	194

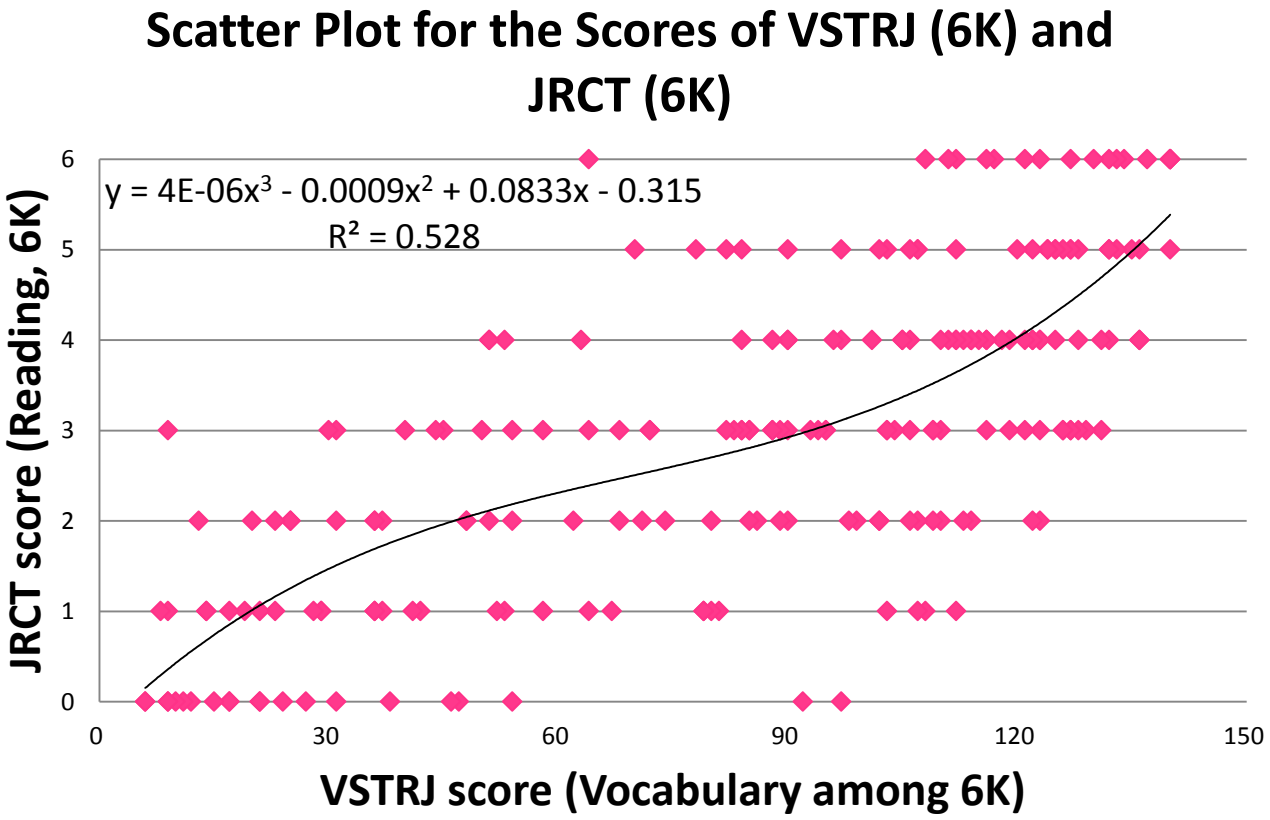
Note: In order to avoid the ceiling effect, participants who obtained 130 or more in full VSTRJ are excluded from the data.

Figure 2. Scatter Plot for the Scores of VSTRJ (2K) and JRCT (2K)



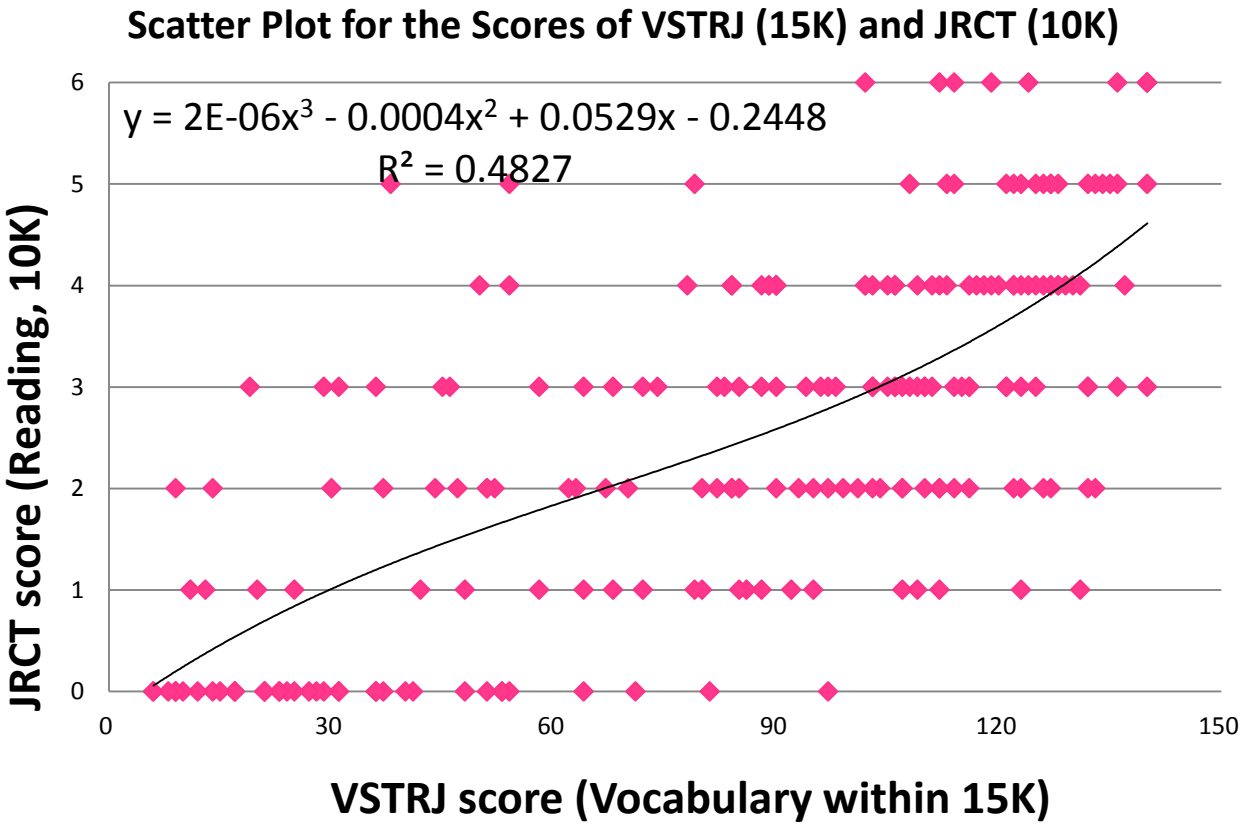
Descriptive Statistics for VSTRJ (15K) & JRCT (2K)

	Max. point	*M*	*SD*	*n*
VSTRJ (15K) (Vocabulary)	150	84.4	39.6	213
JRCT (6K) (Reading)	6	3.0	1.7	213



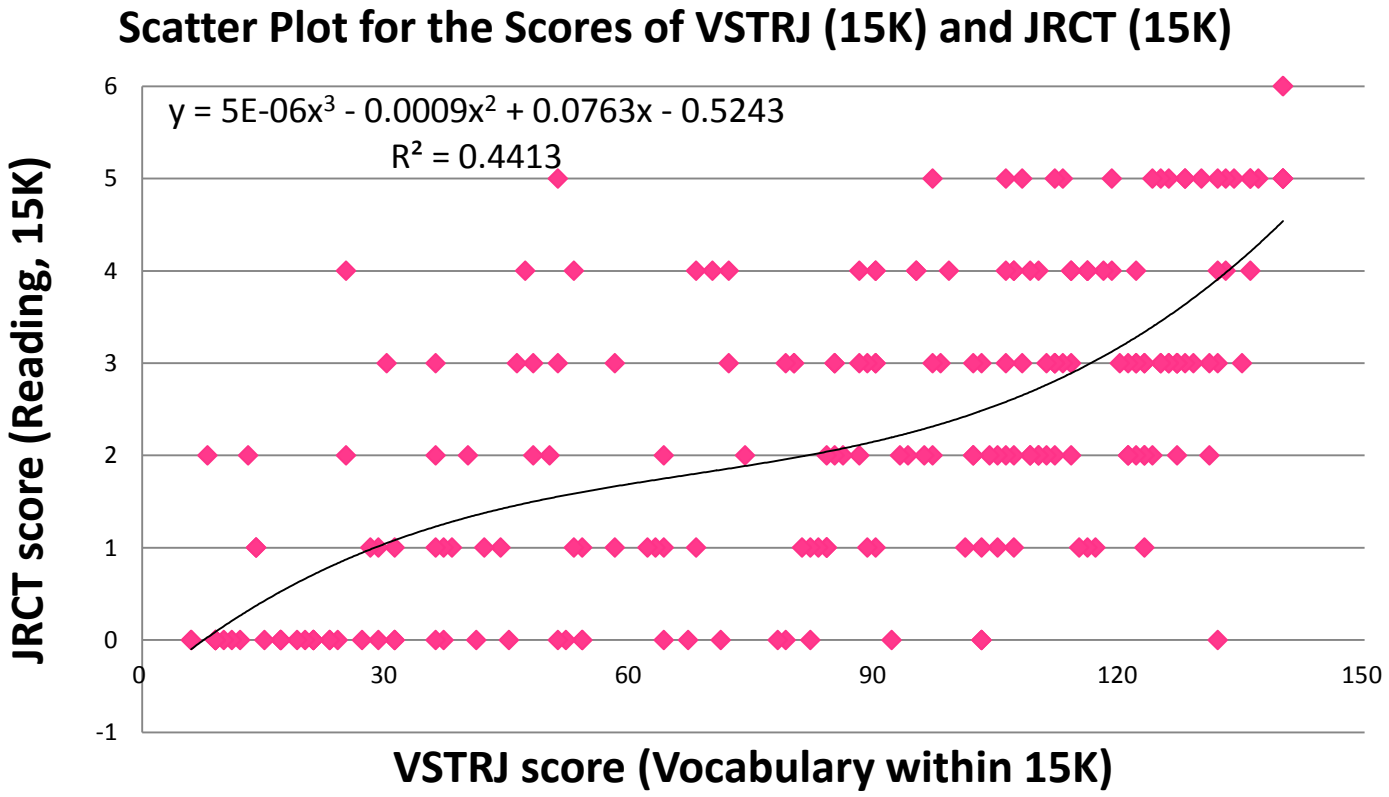
Descriptive Statistics for VSTRJ (15K) & JRCT (10K)

	Max. point	*M*	*SD*	*n*
VSTRJ (15K) (Vocabulary)	150	84.4	39.6	213
JRCT (10K) (Reading)	6	2.5	1.7	213

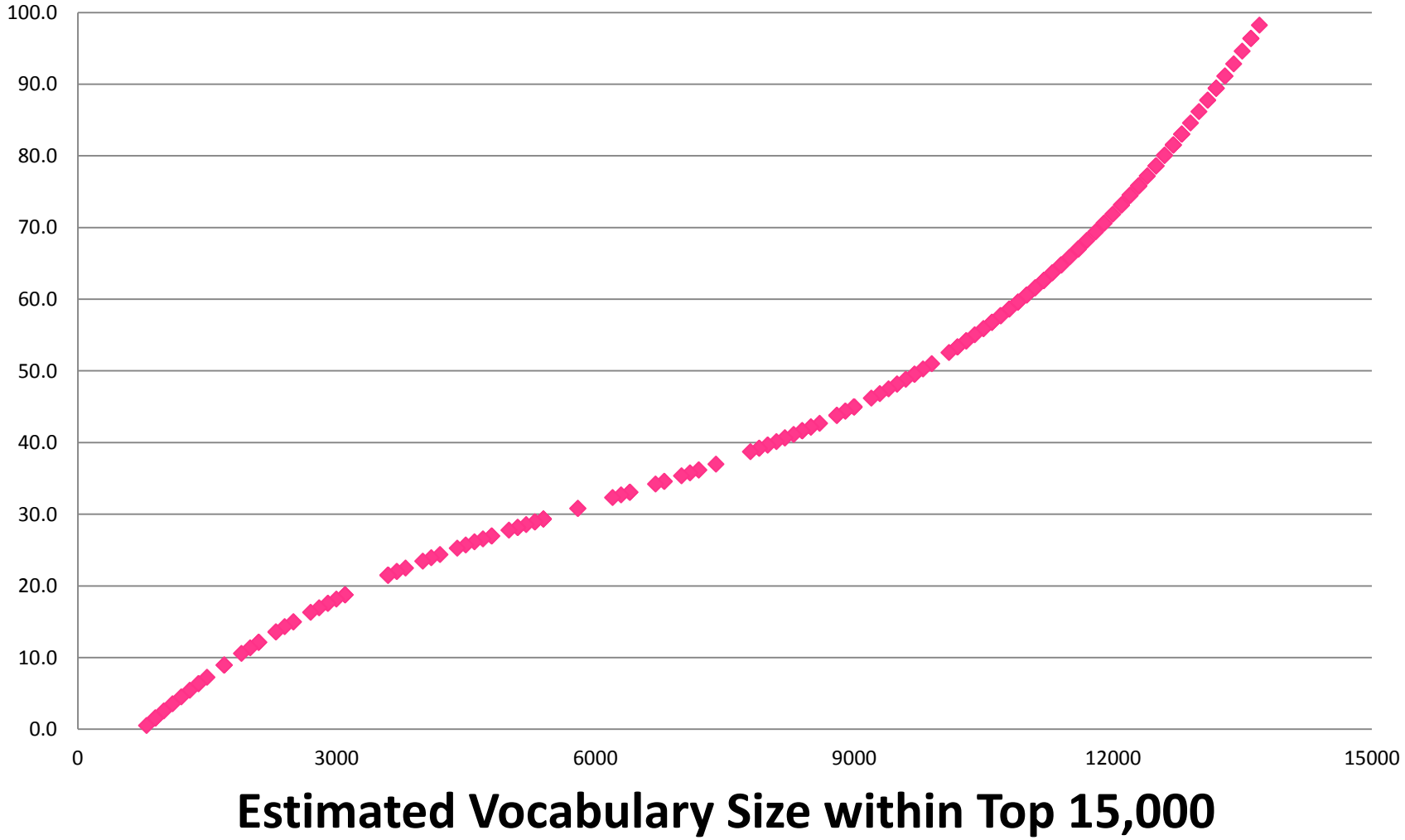


Descriptive Statistics for VSTRJ (15K) & JRCT (15K)

	Max. point	*M*	*SD*	*n*
VSTRJ (15K) (Vocabulary)	150	84.4	39.6	213
JRCT (15K) (Reading)	6	2.3	1.6	213



Expected Value for Reading Comprehension from the Regression Formula（%）



Summarized Results of Vocabulary Size Test for Reading Japanese (VSTRJ)

Rounded Mean VSTRJ Scores for Each K by Learners at Different Estimated Vocabulary Size

		Max. points	150	10	10	10	10	10	10	10	10	10	10	10	10	10	10
Score Range	n	Estimated Vocabulary Size (by Lemma)	Total	01K	2K	3K	4K	5K	6K	7K	8K	9K	10K	11K	12K	13K	14K	15K
131-140	19	13526	135.3	10	10	10	10	10	10	10	9	9	9	8	8	7	8	8
121-130	30	12483	124.8	10	10	9	10	10	10	9	8	8	8	7	7	6	7	7
111-120	23	11461	114.6	10	10	9	9	9	9	8	7	7	7	6	6	5	6	6
101-110	27	10585	105.9	10	9	8	8	9	9	7	7	6	7	5	6	5	5	6
91-100	11	9573	95.7	10	8	8	7	8	9	7	6	6	6	5	5	5	3	4
81-90	22	8623	86.2	10	8	7	7	8	8	6	5	4	5	4	4	3	3	4
71-80	9	7611	76.1	8	6	6	6	6	7	5	5	3	4	4	5	3	3	4
61-70	9	6556	65.6	9	6	5	5	7	7	4	3	3	3	3	4	3	2	3
51-60	11	5355	53.5	9	7	5	3	4	5	5	2	2	3	2	3	2	2	1
41-50	9	4567	45.7	8	5	4	3	3	4	3	2	2	2	2	2	2	2	2
31-40	11	3536	35.4	9	5	3	3	2	3	3	2	1	1	1	2	1	1	1
21-30	13	2508	25.1	8	4	2	2	2	2	2	0	1	0	1	1	1	1	0
11-20	11	1536	15.4	6	2	1	1	1	1	1	0	0	0	0	1	1	0	0
1-10	8	825	8.3	4	1	0	1	1	0	1	0	0	0	0	0	0	1	0
Total	213	8443	84.4	9.0	7.2	6.4	6.5	6.7	7.0	6.0	5.0	4.8	5.0	4.2	4.7	3.8	4.1	4.1

Mean Expected Text Coverage by Words Known to Learners at Different Estimated Vocabulary Size (VS) within the Top 15,000 lemmas

Mean Expected Text Coverage by Words Known by Learners at Different Estimated Vocabulary Size (VS) within the Top 15,000 lemmas

			01K	2K	3K	4K	5K	6K	7K	8K	9K	10K	11K	12K	13K	14K	15K
Learners' Score Range	Expected Text Coverage by the Range Represented by Each 10 Items %		71.1	10.2	4.1	2.3	1.5	1.2	0.9	0.7	0.6	0.5	0.4	0.4	0.3	0.3	0.3
131-140	VS: 13,526	Correct Ans. %	99.5	99.5	95.8	97.9	97.4	98.9	96.8	86.3	87.9	91.1	78.9	83.2	73.7	83.7	82.1
Expected Text Coverage by Known Words (%) n=19			70.8	10.1	3.9	2.3	1.5	1.1	0.9	0.6	0.5	0.5	0.3	0.3	0.2	0.2	0.2
Cummulative Expected Text Coverage (%)			70.8	80.9	84.8	87.0	88.5	89.7	90.6	91.2	91.7	92.2	92.6	92.9	93.1	93.4	93.6
91-100	VS: 9,576	Correct Ans. %	96.4	78.2	80.0	70.9	80.0	87.3	65.5	60.0	57.3	61.8	52.7	50.0	45.5	34.5	37.3
Cummulative Expected Text Coverage (%) n=11			68.5	76.5	79.8	81.4	82.6	83.6	84.2	84.7	85.0	85.3	85.6	85.8	85.9	86.0	86.1
51-60	VS: 5,355	Correct Ans. %	88.2	65.5	49.1	32.7	38.2	50.0	46.4	21.8	20.9	25.5	17.3	29.1	18.2	20.9	11.8
Cummulative Expected Text Coverage (%) n=11			62.7	69.4	71.4	72.1	72.7	73.3	73.7	73.9	74.0	74.2	74.2	74.3	74.4	74.5	74.5
11-20		Correct Ans. %	60.0	16.4	8.2	12.7	10.0	7.3	11.8	1.8	2.7	2.7	1.8	5.5	6.4	3.6	2.7
Cummulative Expected Text Coverage (%) n=11			42.7	44.3	44.7	45.0	45.1	45.2	45.3	45.3	45.3	45.4	45.4	45.4	45.4	45.4	45.4

6. Discussion

Learners will not necessarily learn all words by frequency order, which means that text coverage represented each test item should be multiplied by correct answer ratio by learners at different levels, so that we are able to capture how the text coverage (by words known to learners) develops.

1) How much does vocabulary knowledge account for L2 reading comprehension in Japanese?

- At least more than 40% in reading Japanese
- Considerably much, probably more than English and many other languages
- Learning burden of Kanji, which are used for 40% lemma types, is heavy

*Note: this study does not compare vocabulary knowledge with other variables

2) Does the threshold level of vocabulary known in a text exist (in reading Japanese texts)?

- It depends on the definition of ‘threshold’.
- “A level, rate, or amount at which something comes into effect” (Oxford Dictionary)

  → Once ‘adequate reading comprehension’ is defined, threshold will always exist
- A long S-curve?: the slope is not very steep, but it does not seem to be linear, either

2) Does the threshold level of vocabulary known in a text exist (in reading Japanese texts)?

- In Japanese, 93% coverage with vocabulary size of 11,000-12,000 lemmas seems to be a critical stage for around 70% comprehension and independent reading with a little help from dictionary etc.
- This could be a criteria for providing students with semi-authentic materials for reading courses.

3) What level of vocabulary size (*x*) will assure what level of reading comprehension (*y*)?

$y = 2E\text{-}05x^3 - 0.0036x^2 + 0.3306x - 0.6366$

- 14,000 lemmas → 75-100% (?) comprehension
- 12,000 → 65-91%
- 10,000 → 55-68%
- 8,000 → Approx. 54%
- 6,000 → Approx. 44%
- 4,000 → Approx. 34%
- 2,000 → Approx. 20%

Still difficult to make a precise prediction

4) What level of vocabulary size will assure what level of text coverage by known words?

← Estimation of cumulative text coverage

= Represented text coverage by each item

× the correct answer ratio for each item by learner level

Among the top 15,000 words, learners with *x*K vocabulary size are expected to be able to understand y % of the tokens of the text (*x*K → *y* %)

14K → 94%, 10K →86%,

6K →75%, 2K→ 45%

5) What level of text coverage by known words will assure what level of reading comprehension?

(Learner's estimated vocabulary size (by lemma):

Text coverage by tokens known to learners

→ Level of reading comprehension)

2K: 45% → 20%

6K: 75% → 44%

10K: 86% → 55-68%

14K: 94% → 75-100% (?)

(Learner's estimated vocabulary size (by lemma):
Text coverage by tokens known to learners
→ Level of reading comprehension)

2K: 45% → 20%

6K: 75% → 44%

10K: 86% → 55-68%

14K: 94% → 75-100% (?)

'Adequate reading comprehension' seems to require less amount of text coverage (words known to learners) than 95% claimed in previous studies.

7. Implications

Results will provide useful information for rewriting reading materials to adjust the lexical level of the text to the learner's lexical level, in order to assure that the learners will be able to understand the text to an appropriate extent with an appropriate amount of learning target.

- In Japanese, 93% coverage with vocabulary size of 11,000-12,000 lemmas seems to be a critical stage for around 70% comprehension and independent reading with a little help from dictionary etc.
- This could be a criteria for providing students with semi-authentic materials for reading courses.

8. Limitations & Future Research

- No comparison with factors other than vocabulary
- Kanji level was not controlled in both the vocabulary test and the reading texts
- Difficulty in measuring reading comprehension
  - Limited types and number of texts
  - Limited types and number of reading questions
- Test participants' L1 may need to be considered

→ Validate the tests further to predict the relationship between the factors more precisely

THANK YOU!

References

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, *25*, 133–150.

Hu, M. H., & Nation, P. (2000). Unknown vocabulary density and reading comprehension. Reading in a Foreign Language, 13(1), 403–430.

Koda, K. (1989). The Effects of Transferred Vocabulary Knowledge on the Development of L2 Reading Proficiency. *Foreign Language Annals*, *22*(6), 529–540.

Komori, K., Mikuni, J., & Kondo, A.（小森和子・三國純子・近藤安月子）. (2004). 文章理解を促進する語彙知識の量的側面　—既知語率の閾値探索の試み— (What percentage of known words in a text facilitates reading comprehension: a case study for exploration of the threshold of known words coverage). *日本語教育 (Journal of Japanese Language Teaching)*, *125*, 83–92.

Laufer, B., & Ravenhorst-Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*(1), 15–30.

Matsushita, T.（松下達彦）. (2011a). 日本語を読むための語彙データベース (The Vocabulary Database for Reading Japanese). Downloaded from http://www.geocities.jp/tatsum2003/

Matsushita, T. (2012a) In What Order Should Learners Learn Japanese Vocabulary? A Corpus-based Approach. Unpublished PhD Thesis. Victoria University of Wellington

Matsushita（松下達彦）(2012b). 日本語を読むための語彙量テスト」の開発」(Development of Vocabulary Size Test for Reading Japanese).『2012年日本語教育国際研究大会予稿集第一分冊』[Proceedings for ICJLE Naogya 2012], 310.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.

Schmitt, N., Jiang, X., & Grabe, W. (2011). The Percentage of Words Known in a Text and Reading Comprehension. *Modern Language Journal*, *95*(1), 26–43.

Yano, Y., Long, M. H., & Ross, S. (1994). The effects of simplified and elaborated texts on foreign language comprehension. Language Learning, 44(2), 189–219.