Text Covering Efficiency and Word Tier Analysis for the Proposal of Vocabulary Learning Order and the Analysis of Text Genres

QUALICO 2023, at the University of Lausanne, 28-30 June 2023

Please leave your questions and/or comments here if any.

Or simply send an email to: tatsu.matsushita@ninjal.ac.jp

Tatsuhiko Matsushita (National Institute for Japanese Language and Linguistics, NINJAL)

Background

●L2 vocabulary learning is a significant burden for learners.

⇒**In what order** should learners learn vocabulary **to increase text coverage** most efficiently?

←The higher the text coverage, the better the comprehension will be. (e.g., Laufer & Ravenhorst-Kalovski, 2010)

⇒How can we **evaluate** the efficiency of **a word list** (e.g., Academic Word List, Coxhead, 2000) for learners?

*Simply checking the frequencies and text coverage of the grouped words (e.g., Hyland & Tse, 2007) is not enough to understand the efficiency (i.e., usefulness) of the words, especially when the total number of words are different between word lists (groups) (Matsushita, 2012). **Solution ⇒TCE**

- How can the characteristics of **text genres** be discerned based on the proportion of **grouped words with different domains** and **levels** (Basic/Intermediate/Advanced and so on) ?

Purpose

- To demonstrate how useful an index **'Text Covering Efficiency (TCE)'** (Matsushita, 2012) is for determining the vocabulary learning order for a particular text genre.
- To demonstrate the **'Word Tier Analysis (WTA)'** (Matsushita, 2012) applied to Japanese **medical texts** as an example to evaluate word groups (lists) and show the lexical characteristics of the target genre.

*The term "word tier" in Matsushita (2012) means a somewhat similar (especially in purpose) concept to the one used in Burch and Egbert (2022), but totally different in the method for identifying the tiers.

Methods

[# of lemmas, in one million lemmas*]

○Extraction of Domain-specific Words for Grouping Words

- **Target corpora:** Academic texts in the **4 domains** of Arts & Humanities (Arts)[0.47], Social Sciences[1.76] (Soc), Science & Technology (Tec)[0.41], Biology & Medicine (Bio)[0.26], (as well as Literary Works (Lit.)[8.25])
- **Reference corpus:** Non-academic texts from BCCWJ 2009 monitor version [29.9] *BCCWJ: Balanced Corpus of Contemporary Written Japanese
- **Index for keyness:** Log-likelihood Ratio

Words extracted from X domains are called AcaXD words, e.g., Aca4D words are the words shared in the domain-specific word lists extracted from the 4 academic domains respectively.

○Text Covering Efficiency (TCE)

・TCE (*E*) (Matsushita, 2012)

$$E = \frac{F_t}{L_{tw}} \times \frac{1{,}000{,}000}{N_t} = \frac{F_t \times 1{,}000{,}000}{L_{tw} \times N_t}$$

Mean coverage by a word In the tested word group

Standardization

E: **Text Covering Efficiency (TCE)** = Expected text coverage (=number of tokens) of a word in the tested word group in a one-million-token text in the target domain

Ft: **Number of tokens** (= **text coverage**) of the tested word group in the target text

Ltw: **Number of lemmas** of the tested word group

Nt: Number of tokens in the target text (**text length**)

TCE is relatively **easy to calculate,** and the results can be easily applied with **relatively little distortion due to different corpus sizes.** If word lists such as basic/academic/technical vocabulary for various genres are analyzed by a vocabulary frequency profiler such as AntWordProfiler (Anthony, 2022), for example, the word tiers can be shown by calculating the TCE easily.

○Word Tier Analysis (WTA):

An analysis to see the difference of genres based on TCE figures of various grouped words (e.g., Academic (4D-3D) /Technical (2D-1D) /Literary words) by frequency level (Rank 1-1,291:**Basic**; 1,292-5,000:**Inter.**; 5001-10000: **H-Adv.**; 10001-15000: **H-Adv.**; 15001-20000: **S-Adv.**) Rankings are by Matsushita (2012).

Results

(WTA applied to med. and other corpora)

				Corpus Genre	Medical Books		Technical ← Formal ←				→ Informal
					For Students	For Doctors	Bio Articles	Tec Articles	News-paper	Whole BCCWJ 2009	Essays, Novels etc.
				Corpus Size	0.59	1.27	0.72	2.71	5.68	32.82	2.10
Word Tier	# of Lemmas	Cum. # of Lemmas	Example Words (Japanese)	Example Words (Translation)	TCE values in Medical Books Order						
Basic_Aca4D	31	31	場合　行う	case; act	875	1073	1099	1069	667	525	382
Inter_Aca1D_Bio	52	83	症状　治療	symptom; medical treatment	612	628	89	9	50	42	25
Basic_Aca1D_Bio	9	92	子　鏡	element; mirror	394	322	78	53	71	103	78
Basic	1,027	1,119	する　ある	do; be	384	415	496	481	531	640	673
H_Adv_Aca1D_Bio	54	1,173	動脈　投与	artery; administration(of drug)	278	246	22	3	6	5	2
Basic_Aca3D	39	1,212	必要　試験	need; test	263	341	425	384	475	354	356
Adv_Aca1D_Bio	68	1,280	肝　血管	liver; blood vessel	200	178	20	2	12	11	6
Inter_Aca4D	559	1,839	-性/性　-的	[suffix] -ty (quality); -ive (adjectival)	182	189	241	271	156	82	56
Inter_Aca3D	53	1,892	腎　リンパ	kidney; lymph	164	121	18	2	3	3	1
Basic_Aca2D	45	1,937	原因　-やすい	cause; easily	108	89	113	119	274	96	59
Inter_Aca3D	542	2,479	障害　パーセント	impairment; percent	91	94	95	84	121	48	27
Inter_Aca2D	391	2,870	細胞　療法	cell; method of medical treatment	85	86	59	48	113	47	25
Adv_Aca2D	429	3,299	肺　来す	lung; bring about	44	36	22	26	20	9	5
Basic_Aca1D_Tec	5	3,304	-用　- 室	for the use of; -room	39	69	92	97	97	92	61
Basic_Lit	142	3,446	血　壁	blood; wall	38	34	44	46	55	149	248
Inter_Aca1D_Tec	46	4,452	鉄　-波	iron; -wave	20	21	39	77	40	43	23
Inter_Aca1D_Soc	111	774	不良　小-/-小	mal-; small/little	12	17	30	15	168	57	26

- Medical books are extremely biased toward technical words in biology and medicine.
- WTA allows you to say things like, "Learning the intermediate words in biology and medicine (Inter_Aca1D_Bio) is 12+ times more efficient in covering medical books for students than in covering newspaper texts, and 3.1 times more efficient than learning advanced words in biology and medicine (Adv_Aca1D_Bio) .

Conclusion

- ◆ WTA using TCE can clarify 1) differences in the relative importance of different groups of words according to the purpose of learning, and 2) lexical differences among various text genres.
- ◆ The higher the TCE figure is, the more useful the words will be. The most efficient learning order can be determined by the TCE figures.
- ◆ TCE is a simple and robust index for comparing the usefulness of word groups (lists) as well as for discerning the lexical characteristics of different text genres.

Limitations & Future Research

- ◆The quality of the words in the text is not considered.
- ◆by incorporating these methods into a word frequency profiler such as J-LEX (Suganaga and Matsushita, 2013), such analysis can be facilitated.

References (excerpt) *For the full reference list, please check it out here.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Hyland, K., & Tse, P. (2007). Is there an "Academic Vocabulary"? *TESOL Quarterly, 41*(2), 235-253.

Matsushita, T. (2012). In what order should learners learn Japanese vocabulary? A corpus-based approach (PhD thesis). Victoria University of Wellington. Downloaded from https://researcharchive.vuw.ac.nz/xmlui/handle/10063/4476

Suganaga, Y. and Matsushita, T. (2013). J-LEX: An Online Lexical Analyzer of Japanese Texts. Available from http://www17408ui.sakura.ne.jp/index.html

This work was supported by JSPS KAKENHI Grant Number 18H00679, 20KK0005 and 23H00072