A Critical Survey of JACET English Word Lists: Reconsideration of the Validity of the Frequency Integration Method

Ishikawa, Shin'ichiro

(Citation)
Journal of Corpus-based Lexicology Studies,1:53-80

(Issue Date)
2018-12

(Resource Type)
departmental bulletin paper

(Version)
Version of Record

(JaLCDOI)
https://doi.org/10.24546/81010588

(URL)
https://hdl.handle.net/20.500.14094/81010588

Journal of Corpus-based Lexicology Studies, Vol. 1 (2018)
Japan Association for English Corpus Studies, Lexicology SIG
ISSN: 2434-169X

A Critical Survey of JACET English Word Lists: Reconsideration of the Validity of the Frequency Integration Method

ISHIKAWA Shin'ichiro (Kobe University)
iskwshin@gmail.com

JACET 英語語彙表の批判的概観
―頻度統合手法の妥当性の再考―

石川　慎一郎（神戸大学）

Abstract

The current study surveyed the history of JACET English wordlists. Their recent edition integrate frequencies obtained from ten kinds of genre subcorpora with means and chose the important vocabulary for Japanese learners. This is a sophisticated approach, but there seems to remain some room for discussion about the appropriateness of using means. Therefore, we tested three alternative indices for integrated frequency: (a) trimmed means, (b) weighted means, and (c) principal components. Our analysis revealed that adoption of an alternative index led to a marked change in the word ranks, suggesting the need to reconsider the appropriateness of using means and to seek for a more reliable method to integrate genre frequencies in a balanced way.

Keywords

Wordlists, Text Genres, Frequency Integration

1. Introduction

Choosing the words to teach is of paramount importance in second language instructions. Thus, the Japan Association of College English Teachers (JACET) has struggled with the task of choosing "a pedagogical vocabulary for Japanese learners of English" since the 1970s (JACET, 2013, p. 2).

Ishikawa (in press) compared five editions of JACET wordlists and two types of recently compiled corpus-based wordlists to clarify the construction of a pedagogical

vocabulary for Japanese learners of English. The statistical analysis showed that the JACET teams have prioritized vocabulary concerning (a) family and people, (b) home and the daily life, (c) food and cooking, (d) clothes and fashion, (e) sports, (f) social meetings, (g) transportation, (h) emotions and mental states, and (i) spoken English. Moreover, they regard vocabulary concerning business and higher-order mental activities less appropriate for their target learners.

Then, how has the JACET chosen the vocabulary for four decades? Which problems remain, and what improvements should be made to make its word selection more reliable? The current study critically surveys the word selection protocol adopted in each of the five editions of JACET wordlists, and then scrutinizes the appropriateness of the frequency integration method adopted in their recent release, which integrates frequencies obtained from ten types of subcorpora as mean values. We examine the extent to which genres can be exclusive and independent and test three alternative methods of frequency integration: trimmed means, weighted means, and principal components. This preparatory analysis is expected to contribute to the preparation of the next edition of the JACET wordlist.

2. History of the JACET Wordlists

2.1 Outline

The JACET released the first edition (J1) in 1981, which was followed by the second edition (J2) in 1983, the third edition (J3) in 1993, the fourth edition (J4) in 2003, and the fifth edition (J5) in 2016. The author has been engaged in compilation of J4 and J5.

The first three editions, which chose approximately 4,000 words, were based on other existing wordlists. The JACET researchers compared several renowned wordlists and chose the vocabulary that were commonly included in many of them. Meanwhile, the 4th and 5th editions, which chose approximately 8,000 words, have become corpus-based. The researchers obtained the word frequencies directly from corpora. Especially, the current version (J5) payed attention to the difference between British English and American English as well as that between different genres to conduct a more reliable word selection. In the following sections, we like to see how words are selected in each of the five editions.

2.2 Wordlists in the Pre-Corpus Age

2.2.1 The 1st Edition (JACET, 1981)

In 1972, a special committee to research and develop English teaching materials was established in JACET. The committee published several textbooks and conducted a

large-scale teacher survey on teaching materials. Based on the survey, they published a research report in 1981 (JACET, 1981), where they discussed the level of constructions, grammar, usage, collocations, and vocabulary that were needed by Japanese college students. As a part of this, they proposed the "JACET List of Basic Words" (J1) to illustrate the content of vocabulary that college students should learn. J1 included approximately 4,000 words because the committee thought that college students should learn 1,000 words in college general education in addition to the approximately 3,000 words that they had learned at secondary schools.

2.2.1.1 Making the Base-List

The J1 editors thought that they could select important vocabulary by relying on large data-based wordbooks. Thus, they decided on the following criteria for their vocabulary selection (Table 1).

Table 1 *Criteria for the vocabulary selection in J1*

Steps	Criterion
1	The words included in the top 6,000 words of both *Computational Analysis of Present-day American English* (CAPAE) (Kučera & Francis, 1967) and *The American Heritage Word Frequency Book* (AHWB) (Carroll, Davies, & Richman, 1971)
2	The words included in the defining vocabulary (2,000 words) of *Longman Dictionary of Contemporary English* (LDOCE) (Procter, 1978).

CAPAE and AHWB were the most reliable data-based wordlists available at that time. The former is based on an analysis of the one-million-word Brown Corpus, the latter on an analysis of approximately five million words of a text corpus, which includes varied written samples encountered by American students in the 3rd to 9th grades.

The defining vocabulary of *LDOCE*, which is based on the General Service List (West, 1953), was added as data of British English (JACET, 1981, p. 50) to supplement two kinds of American English wordbooks. However, it also represented a different principle in word selection. CAPAE and AHWB are based on corpus frequency, but the defining vocabulary is based on the subjective judgments of lexicographers as language experts. Though they may not have been conscious of this, the J1 editors tried to conduct a better-balanced vocabulary selection by paying attention to both frequency-based wordlists and a wordlist based on experts' subjective judgments.

2.2.1.2 Adjustments

The J1 editors recognized that their list, made from three existing lists, was not necessarily suitable for Japanese learners. Therefore, based on discussions of the committee members, they decided to add several high-frequency conjugated forms of irregular verbs (e.g., *bought*), and delete those words whose frequency ranks in CAPAE and AHWB were below 5000 and which were included only in one of the two lists, as well as the words that were not necessarily important in the *LDOCE* definition vocabulary (e.g., *mosque* and *archway*).

Meanwhile, in spite of this adjustment, they did not delete the words that were (1) related to everyday life, (2) used widely in Japanese as imported words, (3) related to English or Japanese cultures, and (4) indispensable in English teaching (e.g., *comma* and *subject*).

The series of procedures for the vocabulary selection in J1 is summarized in Figure 1 below:

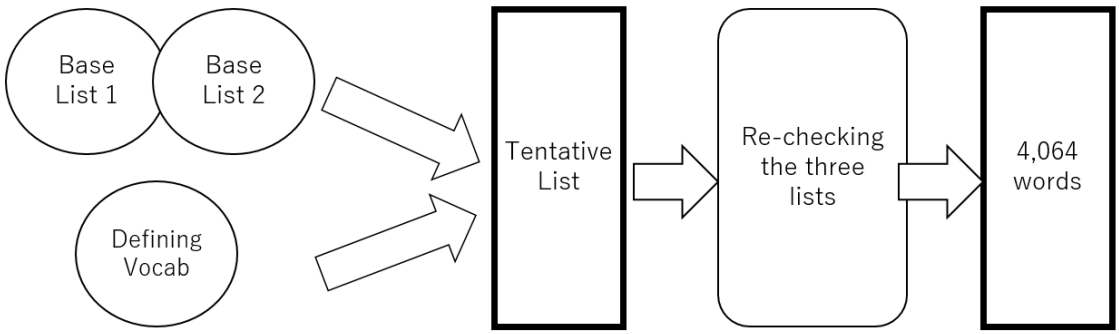

Fig. 1 The word selection process in J1

2.2.2 The 2nd Edition (JACET, 1983)

As J1 was regarded as a tentative proposal, the committee members soon began to revise the list. First, they conducted a teacher survey on J1 and obtained feedback and comments from 206 experts (114 college teachers, 49 high school teachers, 7 junior high school teachers, and 36 experts from local boards of education), and based on the results of the survey, they decided a new procedure to choose the vocabulary for Japanese learners of English. The J2 editors aimed to examine CAPAE and AHWB more carefully and take more wordlists into consideration.

2.2.2.1 Making the Base-List

The J2 editors established the criteria for vocabulary selection in greater detail than

in the J1 project (Table 2).

Table 2 *Criteria for the vocabulary selection in J2*

Steps	Criterion
1	Top 2,000 words in CAPAE
2	Top 2,000 words in AHWB
3	Top 1,215 words (Level 1-2) in *Cambridge English Lexicon: A Graded Word List for Materials Writers and Course Designers* (Hindmarsh, 1980)
4	The words whose frequencies are higher than 100 in *Frequency Analysis of English Usage: Lexicon and Grammar* (Francis & Kučera, 1982)
5	The words included in the top 4,000 words in both of CAPAE and AHWB
6	*LDOCE* Defining vocabulary (2,000 words)

Two lists were newly added. *The Cambridge English Lexicon*, which includes 4,500 words with more than 8,000 semantic values, is the wordbook edited for those who will take the Cambridge First Certificate Examination. *Frequency Analysis of English Usage* is an updated version of CAPAE. In this new version, the word-forms having different lexical statuses (e.g., "will" as a modal, "will" as a verb, and "will" as a noun) and inflectional variants are treated separately. By adding these new lists, J2 editors aimed to conduct a better-balanced word selection, though using CAPAE and *Frequency Analysis of English Usage* in parallel seems technically somewhat inappropriate.

2.2.2.2 Preparation for Adjustments

The J2 editors created a new list of 4,000 words from five kinds of wordlists, but they realized that it was not fully suitable for Japanese college students. Therefore, the J2 editors decided to conduct a pedagogical adjustment. Unlike the J1 editors, who just rechecked the data, the J2 editors collected varied verification data and tried to reexamine the base-list with a new external dataset. It is of note that the J2 editors paid attention to domestic vocabulary studies in addition to overseas studies.

The J2 editors were fully conscious that an original list did not appropriately reflect possible English varieties (such as everyday English, spoken English, and business English) as well as the real state of English education in Japan. Thus, they paid attention to a picture dictionary including much of the everyday vocabulary, a colloquial vocabulary list covering spoken English, a practical English vocabulary list covering business English, and textbook-based wordlists reflecting English language teaching in the country, as well as a famous pedagogical wordlist edited for Japanese students

(Palmer, West, & Faucett, 1936). They also paid attention to the feedback comments from 206 English teachers and experts (Table 3).

Table 3 *Materials collected for adjustments to J2*

Data Type	Materials
Vocabulary lists reflecting varieties of English	
Everyday English	*The Oxford-Duden Pictorial English Dictionary* (Pheby, 1981)
Spoken English	Basic Wordlist of Colloquial English (Kiyokawa, 1976)
Business English	Practical English Wordlist (Arc, 1983)
English in education	Vocabulary List of High School English Textbooks (Yodonawa, 1983) *Basic English Vocabulary List for Japanese High School Students* (Zeneiren, 1981) *Interim Report on Vocabulary Selection for the Teaching of English As A Foreign Language* (Palmer, West, & Faucett, 1936)
Experts' Judgments	
Survey	The result of the teacher survey on J1 (JACET, 1983)

2.2.2.3 Adjustments

By comparing this verification dataset and the tentative base-list made from five kinds of wordlists, the J2 editors picked approximately 1,300 words to be reexamined, and based on the discussions of fifteen editors, they decided to add 239 words and delete 313 words from the tentative base-list for J2, which resulted in a list of 3,990 headwords.

The series of procedures for vocabulary selection in J2 is summarized in Figure 2 below:

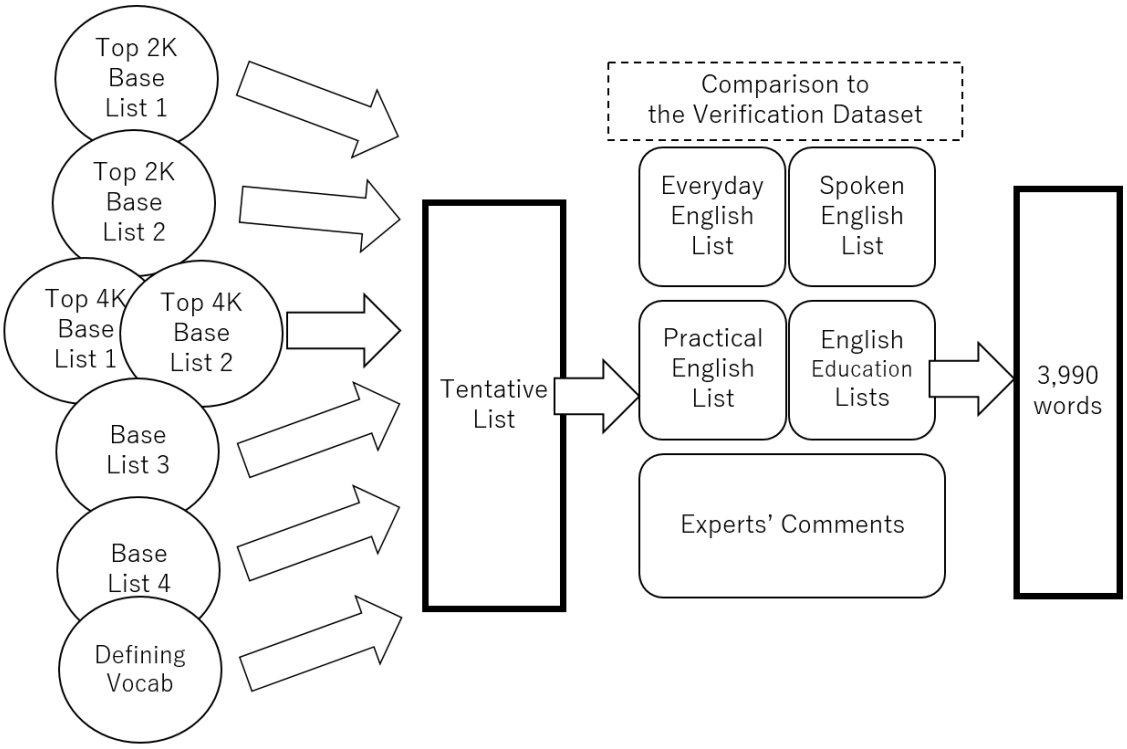

Fig. 2 The word selection process in J2

2.2.3 The 3rd Edition (JACET, 1993)

As J2 came to be widely used, the editors added some extra information to make the list more user-friendly. Thus, five-level frequency bands (based on the *Proceed English-Japanese Dictionary* (Takefuta, 1988)), parts of speech (based on *LDOCE*), and semantic genres (based on the *Longman Lexicon of Contemporary English* (McArthur, 1992)) were added to J2, which was released as J3 (JACET, 1993). The words included in J3 are identical with those in J2.

2.3 Wordlists in the Post-Corpus Age

2.3.1 The 4th Edition (JACET, 2003)

In 1987, the first edition of the *Collins COBUILD English Dictionary*, which was the first fully corpus-based dictionary, was published and attracted international attention. Then, in 1994 the world edition of the British National Corpus (BNC) was released. It included 100 million words of spoken and written samples of contemporary British English. Thus, wordlist developers came to use corpora as new reliable evidence for vocabulary selection. Considering this brand-new trend in applied linguistics, JACET decided to revise their wordlist in 2000.

2.3.1.1 Making the Base-List

Unlike the editors of the previous editions, the J4 editors decided to make a base-list not from other wordlists but directly from a corpus. Thus, they chose to use the BNC as the basic source for their word selection, which was practically the sole corpus publicly available in those days.

First, the J4 editors re-lemmatized the BNC-based 6318-words frequency list (Kilgarriff, 1996), of which a sample is shown below in Table 4.

Table 4 *The BNC wordlist (top five words)*

Rank	Lemma	POS	Freq
1	the	det	6,187,267
2	be	v	4,239,632
3	of	prep	3,093,444
4	and	conj	2,687,863
5	a	det	2,186,369

After the re-lemmatization, the number of words decreased from 6,318 to 5,516, which provided the tentative base-list for J4.

2.3.1.2 Preparation for Adjustments

So as to conduct more transparent and reliable adjustments, the J4 editors decided to obtain external verification data for their self-made corpus (Table 5). Thus, they developed a verification dataset, which is called "JACET8000 Subcorpus" (J8SC).

This self-made corpus included text samples from everyday English, spoken English, American English, English for specific purposes (ESP), and the type of English taught in Japanese schools, all of which the J4 editors thought were not appropriately represented in the BNC. They then created the J8SC-based wordlist, which included approximately 8,000 words.

Table 5 *Materials collected for adjustments in J4*

Data Type	Materials
Everyday English	Children literature*
Spoken English	Cinema scripts*
American English	Newspapers and magazines*
ESP	Science articles and major entries in encyclopedia
	Scripts of science/ politics related TV programs

English in education	English textbooks for secondary school students in Japan
	Exams (Eiken, TOEFL, TOEIC, Center Test [English])

NB: The materials with asterisks include more than one million words of data. Eiken is an English language proficiency assessment test conducted in Japan, which many secondary school students are encouraged to take. Center Test is a national test for those who wish to enter colleges.

2.3.1.3 Adjustments

Using J8SC, the J4 editors conducted three steps of pedagogical adjustments.

First Adjustment

The J4 editors compared the BNC-based wordlist and the J8SC-based wordlist to examine the gaps in rank and frequency for each word. The rank gap was calculated by subtracting the J8SC rank from the BNC rank, and the frequency gap was quantified as the log-likelihood values (LL) (Rayson & Garside, 2000). Then the rank adjustment value, which shows to what degree the BNC rank should be adjusted, was calculated by multiplying the rank gap and the frequency gap ratio, which was obtained by dividing an LL value by the constant of 2000, the maximum LL value observed in the given dataset, excluding several outliers.

$$\begin{aligned}\text{Rank Adjustment Value} &= \text{Rank Gap} \times \text{Frequency Gap Ratio} \\ &= (\text{J8SC Rank} - \text{BNC Rank}) \times (\text{LL}/2000)\end{aligned}$$

Take the example of the word “dollar.” Its rank is 2,259 in the BNC and 76 in the J8SC, while its frequency is 3,700 in the BNC and 891 in the J8SC, meaning that the LL value is +788.88. As the rank gap is −1497 and the frequency gap ratio is 39.4% (788.88/2,000), the rank adjustment value is calculated as −590. Thus, the BNC rank is increased by 590. Thus, the final rank of “dollar” is determined as 1669 (2259−590). The J4 editors sorted all the words by this method and chose the top 8,000 words.

Second Adjustment

Next, the J4 editors carefully reexamined the top 1,000 words in their newly made tentative list (TNL) and realized that, even after the first adjustment, they were still greatly different from the words that Japanese secondary school students learn at school. Thus, they decided to perform an additional adjustment using the high school textbook corpus (HSTC), which was a part of the J8SC. This time, the rank adjustment value was

determined as half of the rank gap between the two lists. When a word was not included in the HSTC, its HSTC rank was set at 8,000.

Rank Adjustment Value = Rank Gap × 1/2
= (TNL Rank – HSTC Rank) × 1/2

Take the example of the word *they*. Its rank is 23 in the TNL and 15 in the HSTC. As the rank gap is +8, the rank adjustment value is calculated as +4 (8*1/2). Thus, the TNL rank is increased by 4 and the final rank is set at 19 (15+4). Another example is the word *republican*, which does not occur in the HSTC. Its TNL rank is 519 and its HSTC rank is estimated as 8000. As the rank gap is −7481, the rank adjustment value is calculated as −3,741 (−7481/2). Thus, the TNL rank is lowered by 3,741 and the final rank is set at 3,221.

The J4 editors applied this additional adjustment only to the top 3,000 words in the TNL because that is the estimated size of the vocabulary that Japanese learners learn at secondary schools.

Third Adjustment

The J4 editors thought that the words taught at junior high schools should be included in the top 1,000 words of their list, and group words such as numerals, names of countries and major cities, and names of days and months should be included together. However, as some of these sets of words did not occur so frequently in the corpus, their ranks were much lower even after two levels of adjustments.

Therefore, the J4 editors decided to exclude these words, whose total number reached 250, from the main list and present them in a sub-list, which is called “Plus 250.” With this final adjustment, the editors added 250 words to the main list. Thus, a completely new list, which comprises an 8,000-word main list and 250-word sub-list, was completed and released under the new name of “JACET8000” in 2003.

The series of procedures for vocabulary selection in J4 is summarized in Figure 3 below:

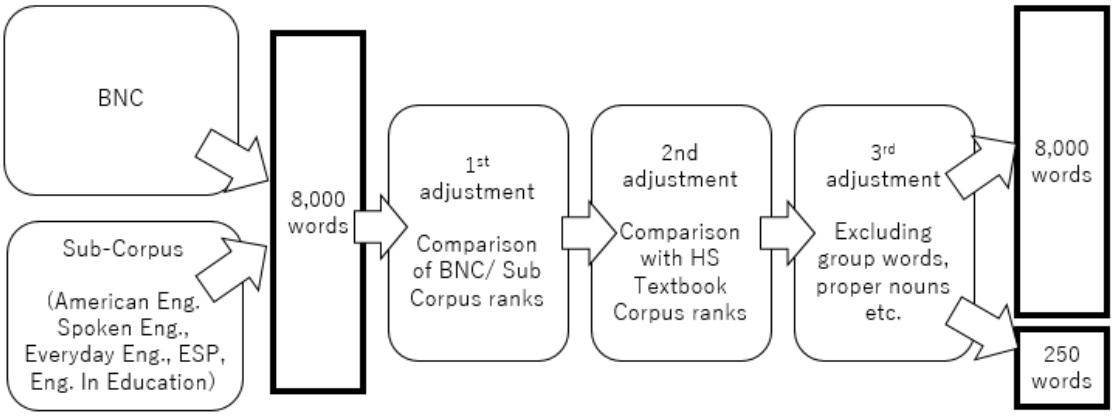

Fig. 3 The word selection process in J4

2.3.2 The 5th Edition (JACET, 2016)

J4 was a great success in the history of wordlist development by the JACET. It came to be used widely not only by learners and teachers but also by many researchers in the fields of TESOL, applied linguistics, psycholinguistics, statistical linguistics, and natural language processing (NLP). However, the J4 editors felt that there remained several substantial demerits in their list.

First, J4 was based only on the BNC for reference data, which does not seem to be entirely rational, especially considering the fact that American English rather than British English is taught at schools in the country.

Second, J4 obtained the frequency data from the whole of BNC, which actually comprises several independent genre modules. Take the example of the word *swim*. Its adjusted frequency per one million words is 13.74 in the whole corpus, but the frequency is 24.46 in fiction, 23.96 in magazines, 14.73 in miscellaneous texts, 14.14 in newspapers, 13.45 in spoken texts, 9.46 in non-academic texts, and only 2.22 in academic texts. This exemplifies the fact that we need to discuss word frequency with attention to genre differences.

Third, the subcorpus compiled for verification of the base-list was limited both in size and balance.

Fourth, J4 presented group words and proper nouns in an independent sub-list, but this undoubtedly reduced the consistency of word selection, for rankings were not assigned to the words in the sub-list.

In light of these limitations, the JACET decided to revise J4 to make it more up-to-date and appropriate for Japanese learners of English. Regarding the first point, the J5 editors decided to use both the BNC and the Corpus of Contemporary American English (COCA), which is a large database including more than 400 million words and practically

the sole corpus of contemporary American English publicly available (see Davies, 2011; Davies, 2015). On the second point, they decided to examine word frequency, not in the whole corpora but in the five genres of spoken transcripts, newspapers, magazines, fiction, and academic texts. As to the fourth point, they drastically expanded the verification dataset. Finally, on the last point, they decided to include all the words in a single list, rather than making a separate main list and sub-list.

2.3.2.1 Making the Base-List

The J5 editors examined how often each lemma appeared in each of the five major genres of the BNC and the COCA. They then extracted the words whose average frequency was higher than 1.0 (per one million words) and whose minimum frequency was higher than 0.1 and chose the top 10,000 words after sorting all the words based on their average frequencies. This was regarded as a tentative base-list. By paying attention to genre frequency rather than corpus frequency, J5 editors obtained frequency data that were much more reliable than in the previous projects.

2.3.2.2 Preparation for Adjustments

The J5 editors developed a much larger-scale verification data set, which included a variety of data to reflect both (A) what Japanese college students have already learned in secondary school and (B) what they will learn in college (Table 6).

For secondary school textbooks and definition vocabularies, the J5 editors only collected wordlists, while for other materials they collected raw text data. The sizes of raw text corpora were roughly between 100,000 and 600,000 words, though the sizes of Eiken Pre-1st Grade corpus and the ESP texts corpus were 20,000 words and more than 10 million words, respectively.

The J5 editors tried to include plural text batches in each of the corpora, which made it possible for them to choose vocabulary with attention to the dispersion (range). Thus, they extracted the words occurring in at least two (three in the case of the ESP text corpus) text batches from each corpus. The number of extracted words was between 1,499 (junior high school textbook corpus) and 7,539 words (newspaper corpus).

Table 6 *Materials collected for adjustments in J5*

Data Type	Data Batches	Extracted Words
(A) Text samples reflecting what Japanese college students have learned		
A1: Junior High School Textbooks	18 books (2012)	1,499 words
A2: Senior High School Textbooks	25 books (2012-13)	3,299 words

A3: Senior High School Entrance Exams	5 years (2010-15)	1,640 words
A4: Center Tests	21 years (1994-2014)	3,286 words
A5: Eiken Test (2nd-5th grade)	7 years (2008-15)	2,913 words
A6: Definition Vocabulary	6 dictionaries	3,175 words
(B) Text samples reflecting what Japanese college students will learn		
B1: TOEFL	12 test sets	3,098 words
B2: TOEIC	12 test sets	2,883 words
B3: Eiken Test (Pre-1st Grade)	7 years (2008-15)	3,884 words
B4: Newspapers	36 days (2000, 12, 14)	7,539 words
B5: ESP texts	8 academic genres	6,163 words

NB: Eiken is an English language proficiency assessment test conducted in Japan (see 2.2.1.2). High school students are expected to pass the 2nd level by the time of graduation. For English newspapers, the J5 editors collected articles from *The Japan News* (formerly called *Daily Yomiuri*), a famous English newspaper published in Japan, as they believed that it included more news on Japan and the Japanese than international English newspapers. For ESP texts, the J5 editors collected academic articles in the fields of agriculture, biology, chemistry, engineering, the humanities, the mathematical and physical sciences, the social sciences, medicine, dentistry, and pharmacy.

2.3.2.3 Adjustments

Thus, the J5 editors obtained eleven kinds of wordlists (A1–A6 and B1–B5) in addition to the tentative base-list. Without using the base-list as it was, they decided to choose the words after three steps.

First Adjustment

First, they paid attention to the six lists A1 to A6, which reflect what learners have already learned, namely, more basic vocabulary for Japanese learners, then they extracted the 1,039 words included in all six lists (Level 1), the 404 words in five lists (Level 2), and the 745 words in four lists (Level 3). The total number of words extracted in this process reached 2,188 words.

Second Adjustment

Second, they paid attention to all eleven lists, A1 to A6 and B1 to B5. They then extracted the 148 words included in eight or more lists (Level 4), the 220 words in seven lists (Level 5), the 299 words in six lists (Level 6), the 393 words in five lists (Level 7), the 511 words in four lists (Level 8), the 766 words in three texts (Level 9), and the 1,216

words in two texts (Level 10). The total number of words extracted in this process reached 3,553 words.

Thus, the editors chose 5,741 words in total from their own dataset. However, the words at each level were not rank-ordered. Therefore, they turned to the base-list and determined the rank-order based on the average frequency in the BNC and the COCA.

Third Adjustment

The J5 editors still needed to choose 2,259 words so as to make their list an 8,000-word list, and these remaining words were chosen directly from the base list. Thus, they finally chose 8,000 words in total.

The series of procedures for vocabulary selection in J5 is summarized in Figure 4 below:

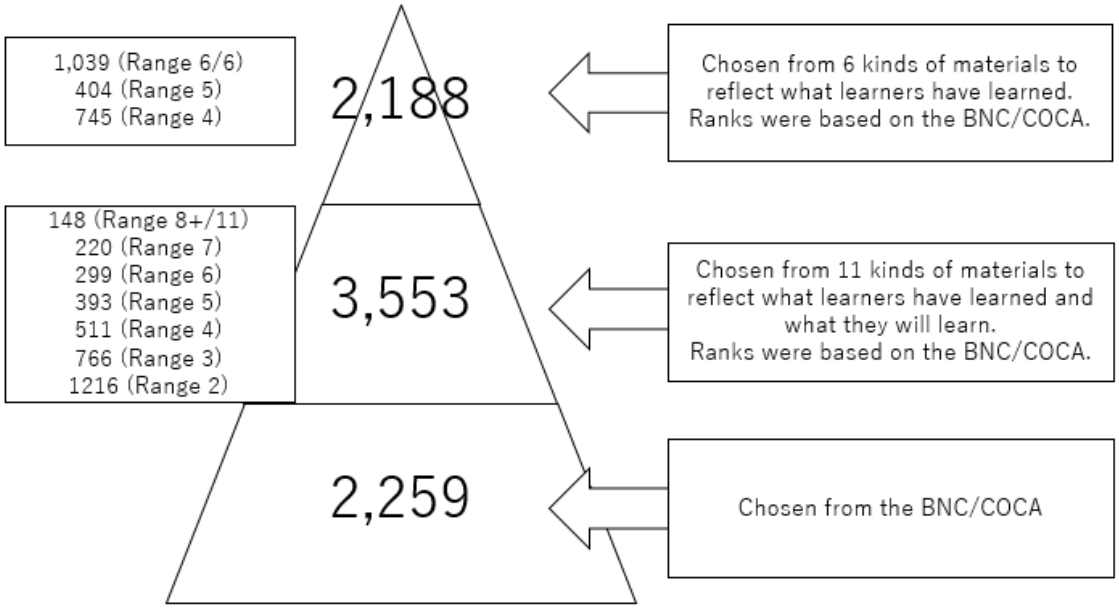

Fig. 4 The word selection process in J5

2.4 Summary of the JACET Wordlist Development

Thus, the JACET has published five wordlists to date. Table 7 below shows the top 20 words in alphabetical order in the different editions of the JACET wordlists.

Table 7 *Samples of the entries in different editions of the JACET wordlists*

J1 (4,064)	J2/J3 (3,990)		J4 (8,000+250)		J5 (8,000)	
Word	Word	Lev	Word	Rank	Word	Rank
a	a	1	a	5	a	4
abandon	abandon	4	a.m.	3933	abandon	2437
ability	ability	3	abandon	2100	abandonment	7810
able	able	1	abandonment	7042	abbey	6021
aboard	aboard	4	abbey	4735	abdominal	7271
about	about	1	abdominal	6921	ability	1526
above	above	1	ability	746	able	246
abroad	abroad	4	able	263	abnormal	5023
absence	absence	5	abnormal	5655	abnormality	5431
absent	absent	5	aboard	4988	aboard	4863
absolute	absolute	5	abolish	3832	abolish	3624
absorb	absolutely	4	abolition	4713	abolition	6752
abstract	absorb	4	abortion	3990	abortion	5756
academic	abstract	5	about	38	about	46
accent	academic	5	above	467	above	1091
accept	accent	3	abroad	1475	abroad	917
acceptable	accept	2	abrupt	6984	abrupt	4423
acceptance	acceptable	5	abruptly	4584	abruptly	4709
access	acceptance	5	absence	2207	absence	2567
accident	accident	3	absent	3989	absent	2084

The type of data that the editors use and the methodology they choose to select vocabulary has changed significantly over 33 years, but their strong will and commitment to choose the appropriate vocabulary for Japanese learners of English has never changed. Therefore, they have consistently adjusted the frequency-based lists in various manners, even when they had obtained direct access to large-scale corpora.

3. Consideration of Adopting Other Frequency Integration Methods

3.1 Background

JACET wordlists have become more sophisticated over the four decades, but their current edition (J5) still has several problems that need to be discussed. One is regarding the method to integrate word frequencies obtained from different genre samples. As exemplified in Ishikawa (2015), the method to integrate frequencies directly influences

the word selection and word ranking. Although J5 sorted all the words based on the mean of ten types of genre frequencies and chose the important words, the use of mean values may have skewed the rank of the words. Thus, we discuss which other options are available and how a word rank changes when we adopt each option.

Then, what problem exists in using mean values to integrate the frequencies obtained from different genre samples? Table 8 shows the per-million-word (PMW) adjusted frequencies and mean values of five sample words.

Table 8 *Integration of word frequencies conducted in J5*

	COCA					BNC					Mean
	Sp	Fic	Mag	News	Acad	Sp	Fic	Mag	News	Acad	
time	1,810	1,751	1,575	1,396	1,225	1,872	1,853	1,431	1,406	1,170	1,549
people	3,316	963	1,405	1,735	990	2,137	828	1,011	1,490	946	1,482
way	1,290	1,322	983	808	624	1,303	1,326	907	759	865	1,019
day	847	937	761	716	279	773	807	660	729	223	673
man	804	1389	441	423	247	420	1,457	429	774	296	668

NB: Sp; Spoken; Fic: Fictions, Mag: Magazines; Acad: Academic texts

By using the mean values, J5 editors seemed to have taken it for granted that (1) the ten genres are all independent, (2) there exist no outliers to be excluded, (3) different genres are equal in terms of data reliability, and (4) all genre information is equal in importance. However, these presuppositions have not been necessarily proven, which suggests that we need to investigate (1') whether all ten genres are independent or can be clustered into several sub-groups, (2') how the integrated frequency changes when we exclude the outliers, (3') how it changes when we consider the difference in the sizes of subcorpora, and (4') how it is influenced when we put different weights on different genres so that all genre information may be better represented.

Therefore, in the current analysis, we conduct a clustering analysis to discuss (1'). Moreover, to examine (2') to (4'), we test three alternative indices for integrated frequency—(a) trimmed means, (b) weighted means, and (c) principal components—on a sample set of words to see how the integrated frequency values change according to those measures.

3.2 Research Design

3.2.1 Aim and Research Questions

The aim of the present analysis is to confirm whether ten types of submodules are

independent enough or not and to see how the integrated frequencies of 100 sample words may change when we adopt three alternative frequency integration methods. Our research questions (RQs) are as follows:

RQ1: How are the ten genres clustered? Are these genres independent enough?
RQ2: After adopting the trimmed means, how does the integrated frequency change?
RQ3: After adopting the weighted means, how does the integrated frequency change?
RQ4: After adopting the principal component values, how does the integrated frequency change?

3.2.2 Data

We examine 100 words (unlemmatized) ranked between 6,000 and 6,100 in the original base list made by Mark Davies, who sorted approximately 10,000 words according to the means of the frequencies in 12 genres, including "non-academic" and "miscellaneous" collected only in BNC. It is expected that the influence of adopting different frequency integration methods would be more salient for these relatively low-ranked words than for basic top-ranked words such as "the" and "a."

We re-sorted these words according to the means of the frequencies in the ten genres, which were analyzed during the compilation of J5, and gave new ranks from 1 to 100. Table 9 illustrates the sample words used for the current analysis.

Table 9 *Some of the words used for the analysis*

Ranked 51–100				
purchased	stare	oppose	risky	alpha
dilemma	fighter	developers	tenure	nationwide
fog	dense	expanding	radar	aide
electronics	vocal	demonstrated	one-third	vanilla
crying	trout	patrol	orbit	rockets
beard	dedicated	authentic	administrator	exhibit
fist	shaped	downs	affordable	filed
transit	naval	bullets	wagon	plaza
transform	franchise	foster	traits	workout
promises	spinning	tolerance	tribes	upcoming

3.2.3 Methods

3.2.3.1 Clustering (RQ1)

Concerning RQ1, we conduct a hierarchical cluster analysis that determines the extent to which the 10 clusters are mutually independent and how they can be classified into subgroups. We use the square root of (2-2r) for calculating the initial distance and adopt the Ward method for calculating the distance after aggregation.

3.2.3.2 Trimmed Means (RQ2)

It is known that just a few outliers may sometimes deteriorate the validity of the means. Therefore, concerning RQ2, we test the 10% trimmed means as a possible alternative to the means. This requires us to recalculate the mean after excluding the top 10% and bottom 10% of all the variables. In this case, we exclude the highest and lowest values. Table 10 shows how the integrated frequencies of the sample words listed in Table 8 change after adoption of trimmed means.

Table 10 *Trimmed means (10%)*

	COCA					BNC					TriM
	Sp	Fic	Mag	News	Acad	Sp	Fic	Mag	News	Acad	
time	1,810	1,751	1,575	1,396	1,225	H	1,853	1,431	1,406	L	1,556
people	H	963	1,405	1,735	990	2,137	L	1,011	1,490	946	1,335
way	1,290	1,322	983	808	L	1,303	H	907	759	865	1,030
day	847	H	761	716	279	773	807	660	729	L	697
man	804	1389	441	423	L	420	H	429	774	296	622

NB: TriM: Trimmed Means (10%). "H" and "L" represent the highest and lowest values, respectively, both of which are excluded when calculating the mean values.

As shown in Table 10, the integrated frequency of "day" increases by 23, while that of "people" decreases by 147.

3.2.3.3 Weighted Means (RQ3)

The sizes of the five subcorpora in COCA are the same, while those in BNC are not in accordance, which suggests the possibility that the frequency obtained from a smaller subcorpus may be statistically less reliable than that obtained from a larger subcorpus. Therefore, concerning RQ3, we test using the weighted means by putting corpus-size-based weights on five genre frequencies obtained from BNC.

Table 11 shows the sizes of the five genres in BNC.

Table 11 *Sizes of the five genre subcorpora in BNC (million words)*

	BNC				
Genres	Sp	Fict	Mag	News	Acad
Sizes	10	15.9	7.3	10.5	15.3

Considering this, we can calculate the BNC weighted means as follows:

BNC Weighted Means = {(Sp Freq * 10) + (Fict Freq * 15.9) + (Mag Freq * 7.3) + (News Freq * 10.5) + (Acad Freq * 15.3)} / (10+15.9+7.3+10.5+15.3).

Using this, we can calculate the BNC/COCA weighted means:

BNC/COCA Weighted Means = (BNC Weighed Means + COCA Means) / 2.

Let us calculate this using the example of "people":

BNC Weighted Mean = {(2,137*10) + (828*15.9) + (1,011*7.3) + (1,490*10.5) + (946*15.3)} / 59 = 1,221,
COCA Mean = (3,316 + 963 + 1,405 + 1,735 + 990) / 5 = 1,682,
BNC/ COCA Weighted Mean = (1,682+1,221) / 2 = 1,451.

In this case, the integrated frequency of "people" decreases by 31.

3.2.3.4 Principal Components (RQ4)

A frequency obtained from a subcorpus includes several types of information: frequency in English in general, frequency in a particular genre, frequency in a particular type of English, and a residue. If we can extract only the information directly related to the frequency in English in general from different genre frequencies and combine it, we may be able to obtain a better-balanced integrated frequency value than an ordinary mean and its alternatives.

Therefore, concerning RQ4, we conduct a principal component analysis (PCA) that integrates a set of variables in a balanced way by giving adjusted weights on different variables, so that the integrated values represent the original set of variables in a better way. Here, we define the first component obtained from the PCA (PC1) as one of the alternative integrated frequency indices.

3.3 Results and Discussions

3.3.1 RQ1 Clustering

We obtained a tree diagram (Fig. 5) from the hierarchical cluster analysis.

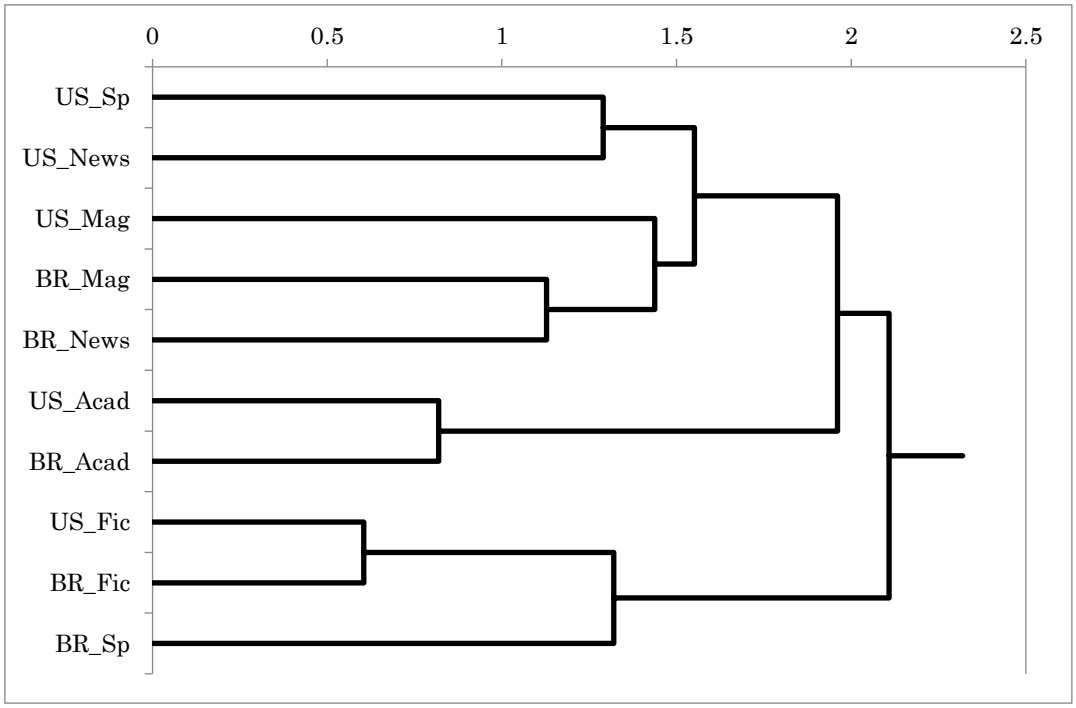

Fig. 5 Tree diagram obtained from a cluster analysis

Fig. 5 shows the following. (1) Many genres are not aggregated at an early state, but US fiction and British fiction as well as US academic papers and British academic papers are aggregated at a considerably early stage. This means that some of the genres have a certain degree of independence from other genres, but others do not. (2) Genres are classified into neither British/American English clusters nor five-genre clusters. This means that the parameters of geographical areas and contextual genres are not mutually exclusive. (3) Fiction and academic papers, both of which are aggregated early, are relatively established as genres, but others are not. (4) The speech genres of British and American English show substantially different features (American speeches are clustered with American news, while British speeches are clustered with US/British fiction). This may be explained by the fact that BNC includes both demographic data, which were collected by directly recording people's natural conversations, and context-governed data such as drafts for varied types of speeches, while COCA includes only the transcribed speeches available online.

The results of the cluster analysis may question the appropriateness of regarding the

ten genres as equal and independent and integrating them as ordinary mean values.

3.3.2 RQ2 Trimmed Means

By calculating 10% trimmed means, we obtained new integrated frequency values. Table 12 lists the top five words whose frequency ranks increased the most in comparison with the ranks based on the means, as well as the same number of words whose frequency values decreased the most.

Table 12 *Words whose ranks changed significantly*

	#	Word	Freq		Rank		
			Mean	TriM	Mean	TriM	Dif
Rank-up	1	naval	10.28	10.37	68	47	↑ 21
	2	promises	10.75	10.64	60	40	↑ 20
	3	dedicated	10.41	10.37	66	48	↑ 18
	4	loyal	12.15	12.44	39	21	↑ 18
	5	dense	10.64	10.37	63	46	↑ 17
Rank-down	1	kissed	14.24	8.67	17	74	↓ 57
	2	capitalism	14.69	9.70	12	60	↓ 48
	3	distinguish	14.42	10.09	16	54	↓ 38
	4	systematic	12.03	8.88	41	72	↓ 31
	5	fist	10.93	7.75	57	84	↓ 27

NB: Dif: Difference in the ranks. "↑" and "↓" represent the rank-up and rank-down, respectively.

It was suggested that adopting a trimmed mean causes (1) word ranks to change a lot (the rank of "kissed" decreases by 57) and (2) the degree of rank-down (27–57) to be larger than that of rank-up (17–21). Table 13 presents the 10 genre frequencies of two sample words.

Table 13 *Genre frequencies of words whose ranks changed significantly*

	COCA					BNC				
	Sp	Fic	Mag	News	Acad	Sp	Fic	Mag	News	Acad
naval	11.77	5.93	~~16.96~~	15.14	12.17	~~2.91~~	5.41	5.65	14.81	12.07
kissed	3.50	51.37	4.73	2.94	0.72	2.31	~~72.79~~	0.69	3.06	~~0.33~~

Table 13 shows that exceptionally high or low values, some of which may be influenced

by the topic of the texts included in the subcorpus, are excluded and more appropriate integrated values are obtained.

3.3.3 RQ3 Weighted Means

By calculating weighted means (WtdMean), we obtained a new set of integrated frequency values, as shown in Table 14.

Table 14 *Words whose ranks changed significantly*

	#	Word	Freq		Rank		
			Mean	WtdMean	Mean	WtdMean	Dif
Rank-up	1	farming	14.74	14.12	11	22	↑ 11
	2	guides	11.38	10.60	49	60	↑ 11
	3	electronics	11.15	10.34	54	65	↑ 11
	4	purchased	11.30	10.60	51	61	↑ 10
	5	attacking	12.31	11.95	36	46	↑ 10
Rank-down	1	systematic	12.03	13.09	41	30	↓ 11
	2	fist	10.93	11.86	57	47	↓ 10
	3	kissed	14.24	16.69	17	7	↓ 10
	4	stare	10.70	11.59	61	52	↓ 9
	5	density	13.01	13.84	30	23	↓ 7

Table 14 shows that adopting a weighted mean, which reflects the difference in the size of BNC subcorpora as regards the statistical reliability of the frequency data, causes (1) word ranks to change less than when adopting a trimmed mean (maximum degree of change is 11) and (2) the degree of rank-down (10–11) to be somewhat larger than that of rank-up (7–11). Table 15 presents 10 genre frequencies of the two sample words.

Table 15 *Genre frequencies of the words whose ranks changed significantly*

	American					British				
	Sp	Fic	Mag	News	Acad	Sp	Fic	Mag	News	Acad
Weights	---	---	---	---	---	10	15.9	7.3	15.3	10.5
farming	5.05	4.57	16.02	14.65	22.51	8.73	**6.03**	21.89	**31.62**	16.37
systematic	3.12	1.05	7.01	3.84	48.28	2.71	**1.19**	5.51	**3.54**	44.09

In this calculation method, greater focus is placed on the frequencies in fiction and news. In the case of "farming," the focus is on the highest value (31.62), leading to an

increase in the rank. However, in the case of "systematic," the focus is on the lowest value (1.19), leading to a decrease in the rank.

3.3.4 RQ4 Principal Component

When conducting a PCA, we usually obtain a well-balanced integrated value as the first principal component (PC1), which should put positive loads on all the variables. With the current dataset, we obtained four principal components (PC1, PC2, PC3, and PC4) whose Eigen values are higher than 1.0, but none of them put positive loads on all the variables. Figures 6–9 show the loads in the four principal components.

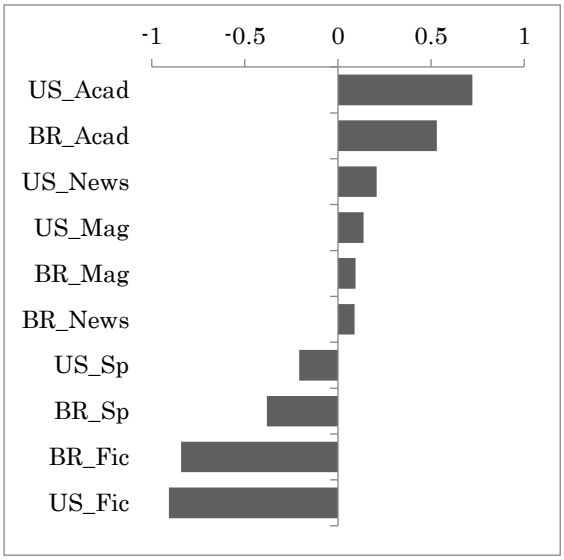

Fig. 6 PC1 loads

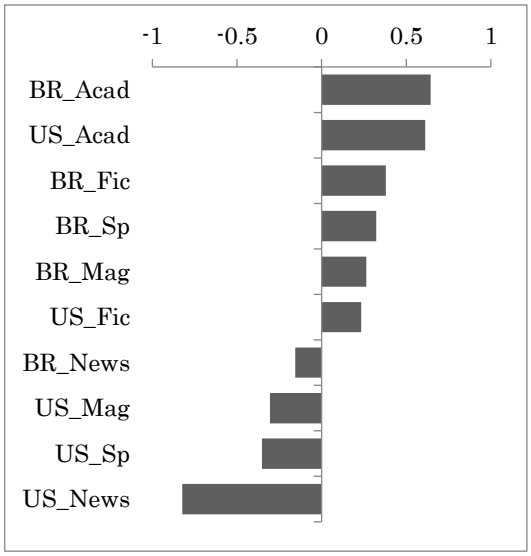

Fig. 7 PC2 loads

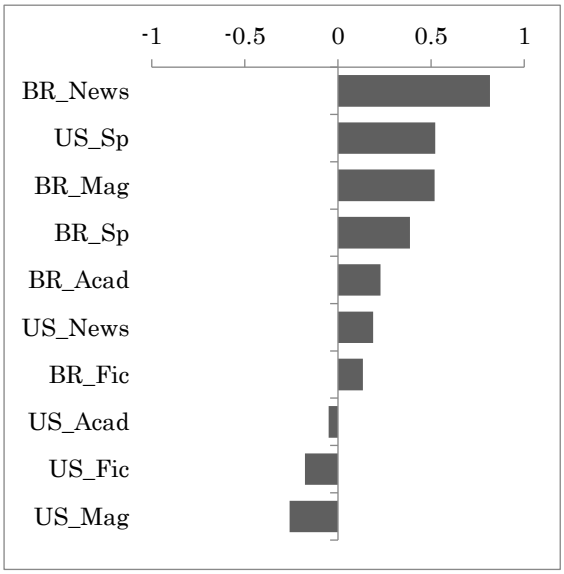

Fig. 8 PC3 loads

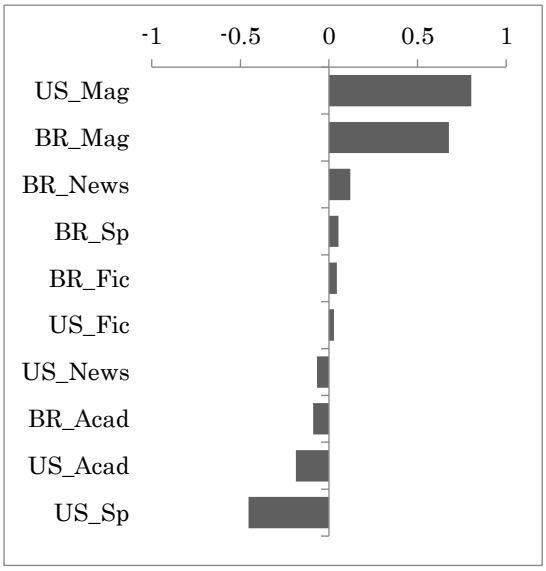

Fig. 9 PC4 loads

It seems that PC1 and PC2, which are similar in quality, represent the axes of *(mainly American) academic texts vs. fiction* and that of *academic texts in general vs. American media English*, respectively. Moreover, PC3 is assumed to concern the axis of *British media/spoken English vs. American written English*, and PC4 shows the axis of *magazines vs. American spoken English*.

From the viewpoint of English for general purposes, it would not be appropriate for us to choose one of these as a well-balanced integrated frequency value.

4. Conclusion

The current study surveyed the history of JACET wordlists. The first three editions were developed by combining existing data-based wordlists, but since the 2000s, JACET wordlists have become fully corpus-based, which renders them more reliable as material for learning the "pedagogical vocabulary for Japanese learners of English."

However, obtaining direct access to genre frequencies rather than corpus frequencies seems to cause wordlist editors a new problem, namely, the problem of how to integrate different genre frequencies in a reasonable way. J5 editors used the means as an index for the genre-integrated frequency, but its validity is not necessarily clear. Therefore, the current study paid attention to how different genres are interrelated and then tested three possible alternatives for the means.

Our quantitative analyses showed that (1) the ten genres are not necessarily mutually independent and exclusive, and some of them (e.g., US fiction and British fiction) are qualitatively identical (RQ1); (2) adoption of a trimmed mean, which controls the influence of outliers, leads to a considerable change in word ranks (RQ2); (3) adoption of a weighted mean, which controls the influence of the difference in the size of BNC subcorpora, also leads to a change in word ranks, but its effect is rather restricted in comparison with adoption of a trimmed mean (RQ3); and (4) as the internal variance among variables is very large, it was impossible to obtain some principal components that could be alternatives to the integrated frequency index (RQ4).

An important finding is that rank-changes are seen with most of the sample words (95% for trimmed means and 83% for weighted means). Tables 16 and 17 respectively show the number of words that are ranked-up or ranked-down and the extent to which their ranks changed.

Table 16 *Number words with changed ranks*

	Rank-up	Rank-down	Same Ranks
Trimmed	58	37	5
Weighted	45	38	17

Table 17 *Width of the rank change*

	Rank-up				Rank-down			
Width of the rank change	1-	5-	10-	20-	1-	5-	10-	20-
Trimmed	19	16	21	2	12	9	7	9
Weighted	25	17	3	0	19	14	5	0

A corpus-based word selection is always related to the frequency. It should be noted that the choice of a frequency integration method may strongly influence the word selection and word ranking.

Another important finding is that although the two alternative methods we tested here are based on different principles—exclusion of outliers and control of the difference in the size of subcorpora—several words are commonly ranked-up or ranked-down. Table 18 lists these words.

Table 18 *Words commonly ranked up or down*

Commonly ranked-up words	Commonly ranked-down words
accurately, administrator, dense, distress, expanding, filed, foster, naval, oppose, orbit, promises, sandwich, shaped, skilled, steadily, suspicion, tenure, tolerance, torn, transform, within	developers, downs, fog, positively, score, scoring, speculation, threatened, vanilla

Existence of these words may question the validity of means as a frequency integration method adopted in J5. If we rank the words according to several different frequency integration measures and choose only words that are commonly ranked highly, we may be able to make the word selection much more reliable.

These are suggestive results, and we have to be careful about easy overgeneralization of the findings obtained in the present study. In a future study, we aim to expand the number of sample words used for the analysis and the number of alternative methods to verify the replicability of our findings. We also need to investigate the possibility of using several alternative methods in combination (e.g., excluding outliers while

simultaneously considering the difference in the size of the subcorpora).

Although it is not clear how JACET wordlists will change in the future, their next edition can become more reliable through adoption of an appropriate method to integrate genre frequencies.

Bibliography

Arc (1983). Jitsuyo eigo goi risuto. *Bessatsu English Journal*. Tokyo, Japan: Arc. [Practical English wordlist].

Brezina, V., & Gablasova, D. (2013). Is there a core general vocabulary? Introducing the new general service list. *Applied Linguistics*, *36*(1), 1–22.

Browne, C. (2013). The new general service list: Celebrating 60 years of vocabulary learning. *The Language Teacher*, *7*(34), 13-16.

Carroll, J. B., Davies, P., & Richman, P. (1971). *The American heritage word frequency book*. Boston, MA: Houghton Mifflin.

Davies, M. (2011). N-grams data from the Corpus of Contemporary American English (COCA). Retrieved from http://www.ngrams.info

Davies, M. (2013). The 100,000 word list of COCA. Retrieved from http://www.wordfrequency.info/intro.asp

Francis, N., & Kučera, H. (1982). *Frequency analysis of English usage: Lexicon and grammar*. Boston, MA: Houghton Mifflin.

Hindmarsh, R. (1980). *Cambridge English lexicon: A graded word list for materials writers and course designers*. Cambridge, England: Cambridge University Press.

Ishikawa, S. (2007). Eigo kyoiku no tame no kihongo o do erabuka: Kopasu gengogaku kara no shiten. *The English Teachers' Magazine*, *55*(13), 10-13. [How should we choose the basic English words to be taught at schools? A viewpoint of corpus linguistics].

Ishikawa, S. (2013). The ICNALE and sophisticated contrastive interlanguage analysis of Asian learners of English. In S. Ishikawa (Ed.), *Learner corpus studies in Asia and the world, 1* (pp. 91-118). Kobe, Japan: Kobe University.

Ishikawa, S. (2015). A new corpus-based methodology for pedagogical vocabulary selection: Compilation of "HEV1800" for Japanese high school students. *Journal of the Chubu English Language Education Society*, *44*, 41-48.

Ishikawa, S. (In press). A reconsideration of the construct of "a vocabulary for Japanese learners of English": A critical comparison of the JACET wordlists and new general service lists. *Vocabulary Learning and Instruction*, *7*.

JACET Kyozai Kenkyu Iinkai (JACET, Committee for the Research of Teaching Materials). (1981). *Daigaku ippan kyoyo katei ni okeru eigo kodoku yo kyokasho no*

arikata. Tokyo, Japan: JACET. [A report on how textbooks for English reading classes at college general education should be].

JACET Kyozai Kenkyu Iinkai (JACET, Committee for the Research of Teaching Materials). (1983). *"Eigo kodoku yo kyokasho no arikata" ni tsuite no anketo chosa hokoku: JACET kihongo dai 2 jian o chushin ni*. Tokyo, Japan: JACET. [The results of the survey on *A report on how textbooks for English reading classes at college general education should be*: With a focus on the second version of JACET list of basic words].

JACET Kyozai Kenkyu Iinkai (JACET, Committee for the Research of Teaching Materials). (1993). *JACET 4000 basic words*. Tokyo, Japan: JACET.

JACET Kihongo Kaitei Iinkai (JACET, Committee for Revision of the JACET Wordlist). (2003). *JACET list of 8000 basic words*. Tokyo, Japan: JACET.

JACET Kihongo Kaitei Tokubetsu Iinkai (JACET, Special Committee for Revision of the JACET Wordlist) (2016). *The new JACET list of 8000 basic words*. Tokyo, Japan: Kirihara Shoten.

Kilgarriff, A. (1996). BNC lemmatized frequency list. Retrieved from https://www.kilgarriff.co.uk/bnc-readme.html

Kiyokawa, H. (1976). Kogo eigo no kihongo risuto sakusei no kokoromi. *Senshu Language Laboratory Bulletin*, *5*, 28-38. [A new list of basic spoken words].

Kučera, H., & Francis, N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.

McArthur, T. (1992). *Longman lexicon of contemporary English*. London, England: Longman.

Palmer, H. E., West, M. P., & Faucett, L. (1936). *Interim report on vocabulary selection for the teaching of English as a foreign language: Report of the Carnegie Conference, New York 1934, and London 1935*. London, UK: P. S. King & Son.

Pheby, J. (1981). *The Oxford-Duden pictorial English dictionary*. Oxford, England: Oxford University Press.

Procter, P. (1978). *Longman dictionary of contemporary English* (1st Ed.). London, UK: Longman.

Rayson, P., & Garside R. (2000). Comparing corpora using frequency profiling. *WCC '00 Proceedings of the Workshop on Comparing Corpora*, *9*, 1-6.

Takefuta. Y. (1988). Keyword 5000. In K. Hasegawa, T. Shimaoka, I. Koike, Y. Takefuta (Eds.). *Proceed English Japanese dictionary*. Tokyo, Japan: Fukutake Shoten.

West, M. (1953). *A general service list of English words: With semantic frequencies and a supplementary word-list for the writing of popular science and technology*. London, England: Longmans, Green.

Yodonawa, M. (1983). *Koko eigo goi no jittai to gakushu goi no arikata*. Tokyo, Japan: Tokyo Metropolitan Institute of Educational Research. [Actual state of English vocabulary taught at high schools and how vocabulary for learners should be].

Zen-eiren (The National Federation of the Prefectural English Teachers' Organization) (Ed.). (1967). *Zen-eiren: Koko kihon eitango katsuyo shu*. Tokyo, Japan: Kenkyusha. [Zen-eiren: Basic English vocabulary for high school students and its use].