In What Order Should Learners Learn Japanese Vocabulary? A Corpus-based Approach

by

Tatsuhiko Matsushita

A thesis
submitted to the Victoria University of Wellington
in fulfilment of the requirements for the degree of
Doctor of Philosophy

Victoria University of Wellington
2012

Abstract

This thesis attempts to answer the following two main research questions: 1) In what order should learners of Japanese as a second language learn words and characters in order to be able to read Japanese? 2) How will the order vary according to the purpose of learning? To answer these questions, a Vocabulary Database for Reading Japanese (VDRJ) and a Character Database of Japanese (CDJ) were first developed from the Balanced Contemporary Corpus of Written Japanese (BCCWJ) 2009 monitor version (NINJAL, 2009) which contains book texts and internet-forum site texts with 33 million running words in total. Word and character rankings for international students, non-academic learners and general written Japanese were included in these databases. These rankings were proven to be valid for their respective purposes as they provided higher text coverage for the target texts than other texts.

After analysing the use of vocabulary and characters in Japanese, three groups of domain-specific words, namely common academic words, limited-academic-domain words and literary words were extracted. In order to test the expected efficiency for learning these groups of words, an index entitled Text Covering Efficiency (TCE) in different types of texts was proposed.

The TCE represents the expected return per unit of text length from learning a group of words. As such, the TCE score in the target text domain should determine the order in which words in this domain are most efficiently learned. Indeed, the extracted common academic words and limited-academic-domain words showed significantly higher text coverage and TCE scores in academic texts than in other texts. Literary words also provided high text coverage and high TCE scores in literary texts, despite a lower efficiency level than that of academic vocabulary in academic texts. Learning domain-specific words is expected to be much more efficient than learning other words at the intermediate level. At the advanced level or above, learning domain-specific words will be further more efficient

in some domains such as the natural sciences. In sum, the TCE has been shown to provide useful information for deciding on the learning order of various groups of words.

Other findings based on the analyses using the databases and word lists include the features of some indices for dispersion and adjusted frequency, lexical features of different media and genres, indexicality of the distributions of word origins and parts of speech, and the discrepancy between learning orders of words and Kanji. A Lexical Learning Possibility Index for a Reading Text (LEPIX) was also proposed for the simplification of a text as a vocabulary learning resource.

Acknowledgement

It has been four and a half years since I left my previous full-time position at a Japanese university. One of my friends once made fun of me and said I was in mid-life crisis. That might have been true. I thought I was purely eager for research activities, though. I have to admit I felt somewhat lonely leaving my home country. Luckily, I received a lot of support, enabling me to make the most of my new academic life.

First, I express my sincere gratitude to my supervisors, Dr. Peter Gu and Professor Paul Nation. They have given me great support and advice since I came to Wellington for the first time on a windy, sunny day to look for a place to study one year before I entered the university. My first teachers in Wellington were Ms. Susan Smith, Mr. Kieran File and Ms. Alison Hoffman at the English Proficiency Programme. Not only did I learn academic English from them, I also learned how to teach a second language. I also thank a number of great staff members in my university including Dr. Stuart Webb, Dr. Averil Coxhead, Dr. Irina Elgort, Dr. Angela Joe, Prof. Janet Holmes and Dr. Frank Boers. I appreciate their scholarship. I am particularly grateful to Dr. Sky Marsen for her warm encouragement and advice on academic writing. I also thank Dr. Deborah Laurs at the Student Learning Support Service for their advice on writing. I am also thankful to the university's financial support, namely DVC initiative scholarship, Faculty Research Grant and English Language Grant for PhD Study.

My special thanks also go to Dr. Laurence Anthony (Waseda University). He accepted my request to improve his software AntWordProfiler to be able to analyse Japanese. Without this, I could not even start my research. I am also indebted to the institute and people shown below. The National Institute for Japanese Language and Linguistics allowed me to use the Balanced Contemporary Corpus of Written Japanese 2009 monitor version. Dr. Yasuyo Tokuhiro and Prof. Kyoko Murakami (Nagoya University) kindly

offered a digitized version of their edited textbooks as a test corpus. Ms. Yukari Hashimoto Honda gave me information on various corpus analysis tools. Dr. Hiroko Fudano (Kanazawa Institute of Technology) kindly sent me copies of many academic articles which I could not obtain in New Zealand. Dr. Yuriko Kayamoto (Japan Foundation) sent me her printed PhD thesis. Dr. Satomi Kawaguchi (University of Western Sydney) gave me an opportunity to give a seminar talk. Dr. Chihiro Kinoshita Thomson (University of New South Wales) and Dr. Katsuo Tamaoka (Nagoya University) provided references when I applied for the PhD programme. I am greatly indebted to Dr. Tamaoka's useful input on psychology and statistics. He also gave me an opportunity to give a seminar talk. Dr. Etsuko Toyoda (University of Melbourne), Dr. Toshihito Kato (Chunghua University), Dr. Kazuko Komori (Meiji University) and many other research colleagues of mine shared their academic interest and insights with me and sent me useful articles.

I am also grateful to my PhD student colleagues including the Vocabulary Discussion Group members, especially for their input on various software tools, statistical analysis and useful articles. These colleagues include Myq Larsen, Yosuke Sasao, Tatsuya Nakata, Michael Rogers, Joseph Sorrel and Betsy Quero. I am also grateful to Dr. Kazuyo Murata who has given me warm encouragement from time to time. My special thanks also go to Ms. Mitsue Tabata Sandome at the Japanese department for her input on reading comprehension studies, encouragement and many things.

I thank my parents Iwao and Tetsuko for providing education and support. My academic basis has been cultivated from many discussions over dinner and many books in our home since I was a child. I thank my human daughter Miki and a doggy daughter Momiji for relaxing me in many ways.

Last but not least, I thank my wife Jun. You always remind me of the fact that I am not a great scholar but a silly naughty guy through nightly Skype calls. I really could not have flown to Wellington and completed my study without your support from Japan.

A Note of the Description of Japanese and Chinese

Principle

When embedding a Japanese word in a sentence, the word in general Japanese orthography (Hiragana, Katakana and Kanji) is noted first, followed by the transcribed Romanized Japanese notation in single quotation marks with English translation in brackets.

e.g. 本 'hon' (book)

For Chinese words, Pinyin with a number for the tone is used as Romanized Chinese notation.

e.g. 书 'shu1' (book)

Notation of Romanized Japanese

Hepburn style Romanization is the base rule; however, regarding the correspondence to Kana description as important, the other ways are used in the cases shown below.

Short vowel/long vowel/double vowel

Short vowel: Hepburn style e.g. ナイト 'naito'

Long vowel: use '^' e.g. ナイトー 'naito^'

Double vowel: notate the vowels e.g. ナイトウ 'naitou'

Borrowed syllables for loanwords

ティ: ti　ディ: di　フィ: fi　フェ: fe

For notating phonemes, follow the conventional way. For a long vowel, use /R/, for double consonants, use /Q/, for ん, use /N/.

Table of Contents

References

Files in the accompanying CD

1. Appendices
    - Appendix 3-1 Technical Notes (1): How to process Japanese corpus files to create word lists and vocabulary databases (VDRJ as an example)
    - Appendix 3-2 Technical Notes (2): Sign Replacement Correspondence Table for VDRJ
    - Appendix 3-3 Technical Notes (3): How to create a Japanese vocabulary database after the morphological analysis (VDRJ as an example)
    - Appendix 3-4 Technical Notes (4): How to exclude the noise (forms not to include the total tokens of the corpus e.g. signs) and fix the errors in VDRJ
    - Appendix 3-5 Technical notes (5): How to create separate lists for assumed known words
    - Appendix 3-6 Forms excluded from the Vocabulary Database for Reading Japanese
2. VDRJ: The Vocabulary Database for Reading Japanese
3. CDJ: The Character Database of Japanese
4. Domain-specific Words
    - JAWL: The Japanese Common Academic Word List
    - LAD: Japanese Limited-Academic-Domain Words
    - LW: Literary Words
5. Baseword Lists for AntWordProfiler
6. Word Tier Analyser
7. List of Publications and Conference Presentations Related to This Research

List of Tables, Graphs and Figures

Tables

Graphs

Figures

List of Abbreviation

AD	the eight academic domains
AH	arts and other humanities
AKW	Assumed Known Words
ASFR	average sub-frequency ranking
AW	common academic words
AWL	the Academic Word List
BCCWJ	the Balanced Contemporary Corpus of Written Japanese
BCCWJ-T	technical texts in the Balanced Contemporary Corpus of Written Japanese, 2009 monitor version
BLI	Beijing Language Institute
BM	biology and medicine
BNC	British National Corpus
BSB	the Best Seller Books (contained in BCCWJ 2009 monitor version)
Bn	biological natural sciences
CAT	Computer-adaptive test
CBL	Chinese-background learner
CDJ	the Character Database of Japanese
DP	deviation of proportions
EC	economics and commerce
F-JLPT	the former Japanese-Language Proficiency Test
HE	history and ethnology
Ha	humanities and arts
IF	Internet Q & A forum
JASSO	Japan Student Services Organization
JAWL	the Japanese Common Academic Word List
JLPT	the Japanese-Language Proficiency Test
JS-Bn	J-Stage texts in biological natural sciences
JS-NS	J-STAGE (Japan Science & Technology Information Aggregator) academic journal article texts in natural sciences
JS-Tn	J-Stage texts in technological natural sciences
JSPS	Japan Society for the Promotion of Science
KGL	the Ranking for Kanji for General Learners
KIS	the Ranking for Kanji for International Students
KWJ	the Ranking for Kanji in Written Japanese
LAD	the limited-academic-domain words
LEPIX	Lexical Learning Possibility Index for a Reading Text
LFP	Lexical Frequency Profiling
LLR	log-likelihood ratio

LLT	Lexical Level of Text
LP	languages, linguistics and philosophy
LW	the literary words / literary works
MC	Meidai Conversation Corpus
MDS	Multidimensional Scaling
MRQ	main research question
MTT-Bn	Meidai Technical Texts in Biological Natural Sciences
MTT-NS	Meidai Technical Texts in Natural Sciences
MTT-Ss	Meidai Technical Texts in Social Sciences
MTT-Tn	Meidai Technical Texts in Technological Natural Sciences
MWU	multi-word units
NDC	Nippon Decimal Classification
NINJAL	the National Institute for Japanese Language (-2009) or the National Institute for Japanese Language and Linguistics (2009-)
NLRI	the National Language Research Institute
PL	politics and law
POS	Part of Speech
SE	sociology, education and other social issues
SFI	Standard Frequency Index
SRQ	sub-research-question
ST	science and technology
Ss	social sciences
TB	Text Bank in Social Sciences for Intermediate and Advanced Learners of Japanese
TC	text coverage
TCE	Text Covering Efficiency
TIS	Texts for International Students
Tn	technological natural sciences
TTR	type/token ratio
UDP	alternative U (usage coefficient) by applying Gries's DP as dispersion measure
UPC	Utiyama Parallel Corpus
UYN	Utiyama Yomiuri Newspaper Corpus
VDLJ	the Vocabulary Database for Learners of Japanese
VDRJ	the Vocabulary Database for Reading Japanese
WGL	the Word Ranking for General Learners
WIS	the Word Ranking for International Students
WWJ	the Word Ranking for Written Japanese

Chapter 1 Introduction

1.1 Aims and importance of the research

1.1.1 The motive for the research

In Japanese, there is a phrase 中級の壁 'chuukyuu no kabe' which literally means *the intermediate wall.* This phrase refers to the phenomenon where learners cannot feel their own progress (or they really do not make good progress) in their second language learning after they reach the intermediate level. In my personal experience in learning English and Chinese as foreign/second languages, I myself felt that I did not make real progress after the intermediate level even if teachers and friends said I did. In my experience in teaching Japanese as a second language, I also often heard similar remarks from my students. This phenomenon seems common among second language learners of any language.

There are several possible reasons for this phenomenon; however, the most persuasive reason for me is a rapid decrease in text coverage gain after learning core vocabulary. For example, in English, the most frequent 1,000 words (lemmas) cover 72% of text (tokens) in the Brown corpus, but the second 1,000 words only cover 7.7%, and the third 1,000 words only cover 4.3%, and the proportion of each 1,000 words continuously decreases as the word level goes down to low-frequency[1]. Nation (2001) shows other coverage data in different types of texts which all show similar coverage between 71% and 85% by the first 1,000 while it ranges between 4-6% by the second 1,000 words. In Japanese magazine texts, the first 1,000 words provide 60.5% coverage; however, the second 1,000 words only provide 9.5% and the third 1,000 words provide even less at 5.3% (NLRI, the National Language Research Institute, 1962).

The decrease of coverage gain means that learners cannot get a consistent return from learning vocabulary as their learning progresses. At the elementary level, learners will meet

[1] I calculated the percentage myself based on the data shown in Nation (2001, p 15).

the words which they have learned, repeatedly in conversation or written texts as the words they learn at the level are high-frequency vocabulary in general. However, at the intermediate level or above, learners rarely meet words they have learned at that level. For example, in Japanese magazine texts, 500 words are required to gain 1% coverage between the 7,000 and 10,000 word frequency levels. In other words, learning 500 words can only gain one word out of 100 running words on average.

The vocabulary learning burden is heavy. It takes time and energy. Even after years of learning, second language learners will still meet new words from time to time, and there seems to be no end. This will definitely influence learners' motivation. Learners' behaviour is also explained by their conscious or unconscious cost/benefit analysis. There are uncountable elective foreign language courses in the world; however, the number of students decreases as the level goes up in most courses. Many learners quit their learning on the way. One major reason for this will be the low benefit of the high cost of learning.

What is more, most class meeting time is not spent on vocabulary as there are many other things to do in a language course. Vocabulary learning is mostly left to learners' effort. Then, how can teachers assist learners to learn vocabulary, especially at the intermediate level or above? How can we gain efficiency in second language vocabulary learning?

One frequent practice is taking advantage of word (frequency) lists. In learning and teaching Japanese, the former Japanese Language Proficiency Test (F-JLPT) word lists are distributed and exploited widely as a standard. Nevertheless, the usefulness of the list is a little questionable as the word lists were made in the 1980's. The word lists for the current test which started in 2010 are created from the beginning but are not publically available. In addition, the F-JLPT word lists have only four levels with no rankings within each level. Other major publically available word frequency lists are made from magazine texts or newspaper texts (Amano & Kondo, 2000; NLRI, 1962, 2006) but not from book or internet texts in Japanese studies.

Another important consideration is domain-specificity. At the intermediate level or above, the best way to gain higher text coverage is to focus on a particular domain because many of the mid-frequency words (relatively high-frequency words beyond the top 2,000 word level) are used in a limited domain. By working in a particular domain, learners are more likely to encounter the same mid-frequency words repeatedly.

However, looking at the issue from the teachers' side, learner needs are generally various within a group of students; therefore, it is not easy to focus on a particular domain unless the learner needs and purpose of learning are homogeneous to some degree. One solution for this problem is to extract common needs from the learner group and identify the words in common needs. The University Word List (Due & Nation, 1984) and the Academic Word List (Coxhead, 2000) are examples of such attempts for learners of English for academic purposes. Nevertheless, in Japanese, there are few such attempts except technical terms in some particular academic fields[2]. In learning and teaching Japanese for academic purposes, for example, extracting common needs at different stages of the curriculum (Tagine, Dusky, & Assai, 2009; Tagine, Terauchi, Assai, & Motswana, 2007) seems an attractive idea. As the study progresses from university preparatory courses to the first-year university curriculum, second and third year, and postgraduate curriculum, learners' needs will gradually narrow down to a specialised field. What (Japanese) words will suit the common needs at their stages of study?

In Japanese, issues with Kanji (logographic or morphographic Chinese characters) and Kanji words also need to be further investigated. Specifically, the learning orders of words and characters seem not well sorted out in teaching Japanese as a second language. For example, some high-frequency words are written in highly complicated Kanji; therefore, these words are first taught in Kana (Hiragana or Katakana, syllabic phonographic

[2] In Japanese, there are some lists for technical terms as well as academic word lists for high-school students but no successful academic word lists for adult L2 learners. For detailed review of the topic, see 7.1.1 in Chapter 7.

characters) or Romanization for conversational use and the orthography is left to some later stage.

Also, a large portion of Kanji words in Japanese vocabulary create various types of gaps in learning Japanese between Chinese-background learners (CBLs) and non-Chinese-background learners (non-CBLs). Many teachers of Japanese know that the gaps exist; however, there are few studies on the size of the gaps. In Japanese, there is also a large portion of English-origin words[3] which would affect vocabulary learning. How many cognates are there in Japanese at different domains and frequency levels? How can they be converted into learning time?

All the issues mentioned above suggest that there are many things to do to gain higher efficiency in vocabulary learning and teaching in Japanese.

1.1.2 The goal and objectives of this research

The overall goal of this research is to explore the most efficient order for learning and teaching of Japanese vocabulary according to the learners' needs.

To attain this goal, I first create a comprehensive vocabulary database and a character database of Japanese from the Balanced Contemporary Corpus of Written Japanese 2009 monitor version (NINJAL, 2009), for guiding learners and teachers to more efficient learning order of words. Various types of word and character lists are also created from the databases. As a step for creating the databases, some theoretical and practical issues with ordering words are also explored.

Some features of Japanese vocabulary and characters will be investigated from the created databases. The relationship between the learning order of words and characters will be explored as well.

Also, some groups of domain-specific words are to be extracted from the same

[3] In this thesis, I call the loanwords from English as 'English-origin words' or 'Western-origin words'.

corpus as used for creating the databases. How the extracted domain-specific words work in different genres and how we can identify the most efficient learning order of words are also investigated.

A specific use of the databases and word lists is also shown as an example.

1.2 Research questions and organization of the study

The main research questions (MRQs) for this research are:

MRQs: In what order should learners of Japanese as a second language learn words and characters in order to be able to read Japanese? How will the order vary according to the purpose of learning?

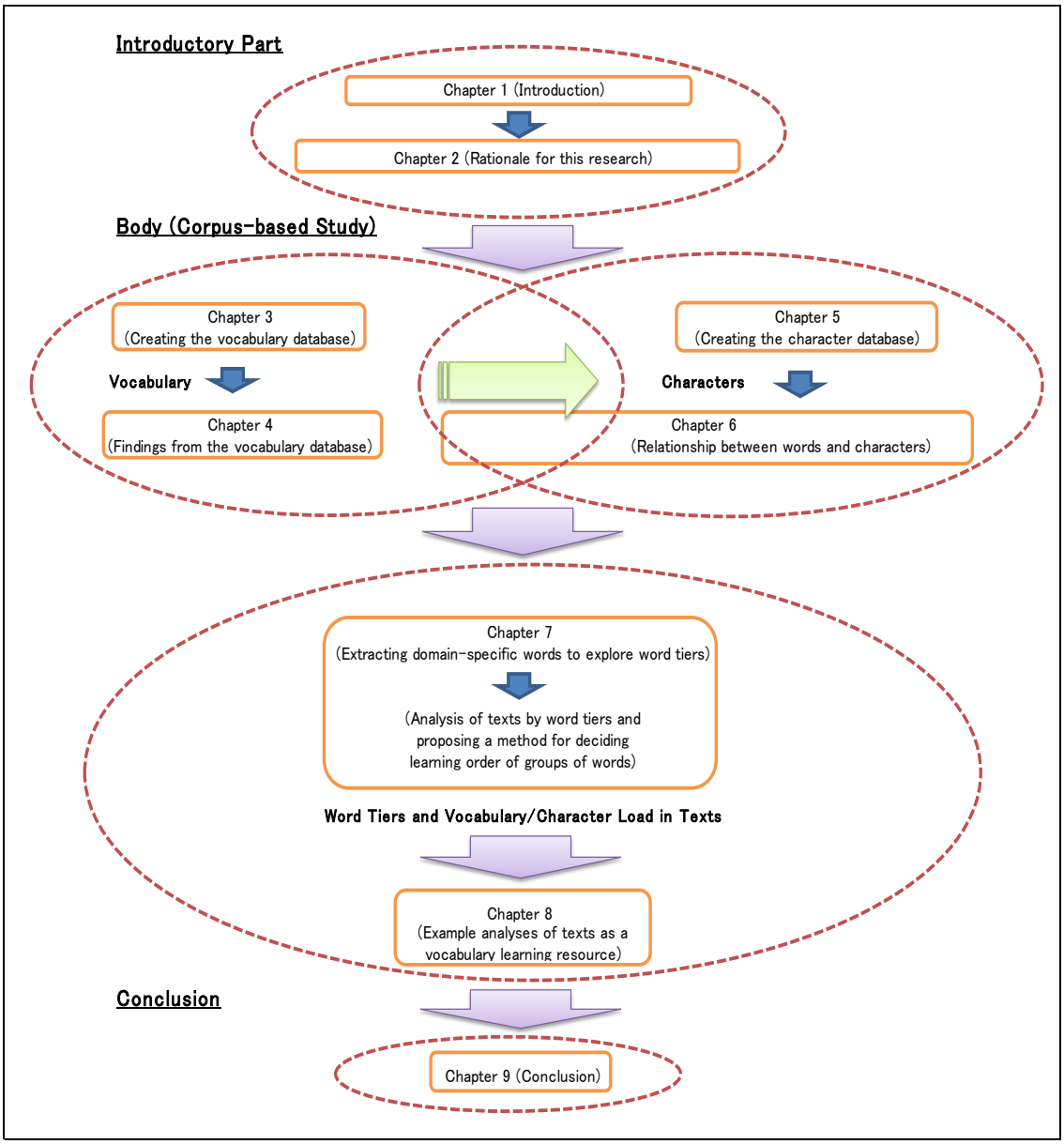
As the corpus used for this study is a written corpus, I limit the range of this research to written receptive vocabulary knowledge, namely the vocabulary knowledge required for reading. To answer the main research questions, there are many sub-research-questions (SRQs). These SRQs will be presented in each chapter.

The organization of this thesis is as follows. Chapter 2 is a literature review of different aspects of the rationale for this research. Chapter 3 to Chapter 8 are the body of this thesis. In Chapter 3, a vocabulary database is created along with the exploration and explanation of how the database is created. In Chapter 4, lexical features of texts in different media and genres are investigated based on the created vocabulary database. The distributions of word origins, parts of speech and their relationship with register variation are also shown. The distribution of Chinese cognates and potential issues with learning and teaching are also mentioned. In Chapter 5, a character database of Japanese is created and the distribution of Japanese characters is reported. In Chapter 6, the discrepancy between

the learning order of words and characters is discussed. Consequently, I also argue that the learning burden of Japanese vocabulary may not be as heavy as generally perceived. Some important ideas of Kanji learning are proposed to reduce the burden of vocabulary learning. In Chapter 7, I will answer the main research questions. Common academic words, limited-academic-domain words and literary words are extracted first, followed by the exploration of how the vocabulary use will vary according to the genre. A newly developed index entitled Text Covering Efficiency (TCE) is proposed for deciding the learning order of groups of words. Chapter 8 is an extra part of this thesis after answering the main research questions, where a method for simplifying a text by exploiting the databases and word lists is shown. An index called the Lexical Learning Possibility Index for a Reading Text (LEPIX) is also proposed to evaluate how efficient a text is for vocabulary learning. Chapter 9 is the conclusion including a summary, implications and further research directions.

The whole thesis is structured as Figure 1.1.

Figure 1-1 The Structure of the Thesis



Chapter 2 Rationale for this research

2.1 Introduction

This thesis covers several different topics related to the most efficient learning order of words. In this section, relevant previous studies which relate to two or more chapters will be reviewed. Topics only related to a particular chapter will be mentioned in that chapter. Specific topics in this chapter include 1) the necessity of well-validated word and character lists, especially studies on the relationship between text coverage and level of reading comprehension (for discussion in 2.2), features of the Japanese writing system, characters and vocabulary, as well as brief reviews of related topics to study (2.3), 3) the methods for creating word and character lists, especially the importance of dispersion and usage coefficient (adjusted frequency) measures in order to investigate the construct of the whole vocabulary of a language (2.4), and 4) the introduction of some possible applications of word and character lists (2.5). At the end of this chapter, some implications for this research will be summarised.

2.2 Vocabulary in reading

The goal of this section is to claim the necessity of a well-validated word list, which can be derived from the database developed for this study. To attain the goal, several things should be confirmed. These include 1) the importance of the word as a unit of language processing, 2) how text coverage of known words in a text contributes to reading comprehension, and 3) the cognate effect on vocabulary learning.

The first point above should be confirmed because the unit of counting in the proposed list is the word (lexeme). Because the word is a unit of processing, it can also be a unit of learning. The second point is also important, because text coverage is a major measure for this study. The basic assumption is simply "the more known words, the better". I will try to confirm this assumption. The third point is also important, because the Japanese

language has a large proportion of Chinese-origin and English-origin words which will make the learning task distinct for learners with a related language background. This is a frequent topic for curriculum design in teaching Japanese as a second language.

2.2.1 Importance of word in language processing

It seems useful to set the 'word' as a target unit of learning because the word is an essential unit in many models of language including Chomsky's (1965) model. There are also many models proposed for understanding a 'word' at a micro-level. One of the leading models is the 'interactive activation model' (McClelland & Rumelhart, 1981). In this model, there are several different levels of information processing such as the visual feature level, the letter level, the word level, the semantic level and the syntactical level; however, the advantage of the word in processing is emphasised. This finding is in line with the finding known as the 'word superiority effect' (Reicher, 1969).

There are also many models proposed for sentence processing or reading comprehension. Again, in many models, the word is incorporated as an essential unit of processing. For example, in Chujo's (1983) model, sentence processing starts from the input of words. Dijk & Kintsch (1983)'s model is known as a leading model of reading comprehension which incorporates both top down processing (situation model) and bottom up model (textbase model). The word is a basic unit for the latter. Levelt's (1989, 1993) model is also one of the most frequently cited models of language processing. In this model, the mental lexicon plays a crucial role. The lexeme is stored in the lexicon and the lemma derived from a lexeme is the unit of syntactic processing.

There are also some models developed for the bilingual lexicon such as the 'bilingual dual coding model' (Paivio & Desrochers, 1980) and the 'revised hierarchical model' (Kroll & Stewart, 1994), both of which modelled on how the words in the L1 and the L2 and concepts/images are linked in the mental lexicon.

2.2.2 Reading comprehension and lexical coverage of text

To what degree does vocabulary knowledge account for reading comprehension? For this question, Bernhardt (2005) provides a comprehensive review and an answer, that is, around 30% is explained by morpho-syntactic knowledge which she thinks is mostly vocabulary knowledge. This figure is mainly based on European and American studies; however, evidence for higher reliance on vocabulary knowledge exists in Japanese studies.

Koda (1989) reports correlations among different aspects of linguistic knowledge, verbal processing skills and reading comprehension. Vocabulary knowledge showed the highest correlation with reading comprehension at $r = .74$ ($p = .001$). Thus, vocabulary knowledge accounts for 55% of the variance. According to Komori, Mikuni, & Kondo (2004), 47% of reading comprehension is explained by vocabulary size (p.117). According to Noguchi (2008), the test results of the subject 'writing/vocabulary' in the former Japanese Language Proficiency Test (日本語能力試験) in 2005 correlate with 'reading/grammar' at $r = .66$ for Level 1 (advanced), $r = .64$ for Level 2 (intermediate), $r = .78$ for Level 3 (upper elementary) and $r = .80$ for Level 4 (elementary) (p.157). These results show that writing/vocabulary (mostly vocabulary and Kanji knowledge is tested) accounts for more than 40% of the variance at any level.

In Bernhardt's (2005) model, L1 literacy accounts for 20%. The other 50% is unexplained variance including comprehension strategies, engagement, content and domain knowledge, interest, motivation and so on. Here I just confirm that a certain degree—seemingly more than 40% at least—of reading comprehension in Japanese is explained by vocabulary and Kanji knowledge.

In this study, text coverage is a major measure for usefulness of grouped words and/or features of a text domain. 'Text coverage' (or 'lexical coverage', 'vocabulary

coverage' (of text)) is the percentage of the total tokens of a group of words. It is the same as cumulative 'standardized frequency' (frequency per unit) of the group of words if the same unit (e.g. percentage) is used as the measure. Using text coverage for measuring usefulness is based on the simple assumption that the more known words in the target text, the better. Therefore, to evaluate a group of words, text coverage is the most important quantitative criterion in general. For example, Coxhead (2000), Coxhead & Hirsh (2007), Nation & Waring (1997) and Terajima (2010) use text coverage as a measure for assessing a group of domain-specific words.

It is also true that low-frequency words, which provide low text coverage, often carry crucial information in the text (Richards, 1974, p 72). Typically, technical terms are often essential in a particular genre and are not replaceable by another word, but are mostly low-frequency words 'in general'. Nevertheless, most low-frequency words only have a limited usage in a limited domain; thus, those words will not always be low-frequency in the corpus of that particular domain. Thus, this type of domain-specific words can be extracted in a statistical way by comparing the frequencies between the target domain and other general domains (e.g. Chujo & Utiyama, 2006). If a learner works in a particular domain, low-frequency words specific to that domain will be important for the learner. Therefore, after learning core vocabulary, some learners are encouraged to work on domain-specific words depending on her/his purpose. Lexical features in different domains and domain-specific words are major topics for this study. These issues are reviewed and discussed in more detail in Chapters 4 and 7. For these purposes, text coverage also provides important information.

Let us look at text coverage and reading comprehension. In English studies, there is an argument whether there is a threshold level of text coverage by known words to attain a certain level of reading comprehension, and how high the threshold is (Hirsh & Nation, 1992; Hu & Nation, 2000; Komori et al., 2004; Laufer, 1989, 1992; Laufer & Ravenhorst-

Kalovski, 2010; Schmitt, Jiang, & Grabe, 2011). Schmitt et al. (2011) claim that there is no clear threshold level but the relationship between text coverage and comprehension is linear. That is, as coverage increases, comprehension increases. The other studies shown above claim a threshold or necessary vocabulary size for different levels of 'adequate comprehension' at a coverage level between 95% and 98%. For example, Laufer & Ravenhorst-Kalovski (2010) suggest two thresholds of an optimal one at 98% and a minimal one at 95% (both including proper nouns). Komori et al. (2004) deal with Japanese texts. They conclude that there seems a possible threshold at 96%. I do not argue whether the threshold exists or not, but confirm that 98% (one unknown word out of 50 words on average) seems enough for independent reading and 95-96% (one unknown word out of 20-25 words on average) will be enough for some cases.

These figures are important for teaching, because they will tell us how we can choose appropriate reading material for L2 learners. If we use an appropriate vocabulary size test along with an analysis of vocabulary load in the target text, we can judge if the text is at an adequate level for the learners (Chapter 9 in Nation & Webb (2011)). The studies did not directly answer how much unknown vocabulary there should be in a text used for classroom instruction; however, the coverage level must be lower than 98% unless it is for fluency development. 95% or even lower coverage is manageable (Nation, 2001, p 150).

Text coverage accounts for reading performance to a certain extent. Thus, this study claims the necessity of a well-validated word frequency list to estimate text coverage by known words for a particular group of readers, because learners' vocabulary acquisition roughly follows the frequency order (e.g. Beglar, 2010; Read, 1988; Schmitt, Schmitt, & Clapham, 2001). A well-validated word frequency list will also enable us to figure out the minimum number of words needed to reach a certain level of coverage which is used for estimating the level of comprehension (Nation, 2006; Nation & Waring, 1997; Nation & Wang, 1999), as well as to clarify the most efficient learning order of words.

When checking text coverage in Japanese texts, one concern is the relationship between Kanji (the morphographic character used for Japanese orthography as well as other phonographic characters) and the word. This issue is mentioned in 2.3.

In sum, vocabulary knowledge seems to account for more than 40% of the variance in measuring reading comprehension. The required level of text coverage will be at some level between 95% and 98% for adequate reading comprehension. The level will depend on the required level of comprehension and the purpose.

2.2.3 Cognate effect on vocabulary learning

Word origins and Chinese cognates are examined in this study[4] as it is assumed that cognates will have a great effect on Japanese vocabulary learning. The first language (L1) effect on vocabulary learning is not limited to cognates (e.g. Jiang, 2000; Paribakht, 2005). Also, using first language knowledge is thought to be an unavoidable process in second language (L2) learning, especially when there is some similarity between the L1 and L2 or learners lack L2 target knowledge (Ringbom, 2007; Swan, 1997): however, conditions other than cognate effect are not reviewed here since they are not limited to a particular group of learners (e.g. Chinese-background learners) but apply to all learners.

If an L2 word also occurs in learners' L1, it is more likely to be understood and learned easily. Thus, cognates which have the same meaning as the original word can be included in known words for the learners with the relevant language background when calculating the required number of words for a certain level of text coverage. Of course, there will be some 'false friends' or partly deceptive cognates which have totally or partly different meanings and/or usages from the original word; however, research has shown that learners' L1 is basically an advantage in understanding cognates (de Groot & Keijzer, 2000; Lotto & de Groot, 1998). Test validation studies also have shown that there is a largely

[4] When analysing the corpus texts for this study, word origin information is tagged to each word so we can calculate the proportion of word origins in the database. For details, see Chapter 3.

positive cognate effect (Chen & Henning, 1985; Cobb, 2000).

In Japanese studies, there are numerous descriptive contrastive studies on the similarities and differences between Chinese cognates and the original Chinese words (e.g. Agency for Cultural Affairs, 1978; Araya, 1983; Hida & Ro, 1987; Kin, 1987, 1990; Lu, 2000). The Agency for Cultural Affairs (1978) tried to list Chinese cognates which are used in ten elementary and intermediate Japanese textbooks and classify them into four categories of same, similar, dissimilar or zero correspondence on meanings and usages. In this study, there were many wrong judgements on classification which were pointed out and corrected by researchers (Arakawa, 1979; Saito, 1988).

In the 1970's and 1980's, description was mainly made on differences. However, as acquisition studies and psychological studies on the L2 started in Japanese studies in the 1990's, the positive side of cognates was also incorporated into the studies. As European studies reveal, if the form of the cognates is similar to L1, learners' L1 knowledge is usually automatically activated. This is also true of the cognition of Chinese cognates by Chinese-background learners (CBLs) of Japanese. Moreover, as Kanji, the (Chinese) logographic characters, have meanings on their own; the impact on semantic transfer may be stronger than that between European languages. CBLs can access the meaning of vocabulary directly from the orthographical representation as well as through phonological processing, while non-CBLs generally access the meaning through phonological processing (Chikamatsu, 1996; Chiu, 2002; Y. Mori, 1998). Experimental studies also provide evidence which demonstrates that L1 Chinese knowledge has a great impact on semantic processing of Chinese cognates in Japanese (Kayamoto, 2002; Tamaoka & Matsushita, 1999; Tamaoka, Miyaoka, & Matsusita, 2004). The result of the former Japanese Language Proficiency Test has also shown that only CBLs have markedly higher scores in the 'writing/vocabulary' test than in other subjects (Noguchi, 2008). For reading performance, Matsunaga (1999) also demonstrates that intermediate CBLs gain significantly higher

scores than non-CBLs in reading comprehension but not in oral performance.

On the other hand, Hatasa (1992) and Machida (2001) suggest that the advantage for CBLs in understanding vocabulary will not always be an advantage for developing overall Japanese proficiency or reading comprehension. Matsunaga (1999) also emphasises the importance of oral proficiency and phonological processing of Kanji even for developing reading skills. Despite these suggestions, however, no study claims there is no cognate effect but rather claims a large impact on learning Japanese. We need to know the distribution of these cognates first, at what frequency levels and in what kind of domains. This will lead to more useful tests and experiments.

As for Western-origin words, which are fewer than Chinese-origin words in proportion, and have no similarity to Japanese in orthography as they are written in Katakana, if a learner can recode the orthographic representation into phonological information correctly to understand what the original word is, it would be an advantage in learning vocabulary[5]. There seems to be few studies on acquisition of English-origin words by learners of Japanese as a second language; however, there are several studies which prove the advantage for Japanese learners of English in learning English words borrowed by the Japanese language (e.g. Daulton, 1998, 2004). Quackenbush & Oso (1990) demonstrate the phonological 'Japanizing' rules of English-origin words. This is useful for English-background learners to recode the Japanese sound of loanwords into the English one. This is already realised as a form of learning material (The Japanese-language Institute, Japan Foundation, 1995).

In sum, cognates have a large effect on learning L2 vocabulary in general and in Japanese. The effect is mostly positive at least for the short term. Cognates with the same meanings can be included in known words for the learners with the relevant language background when calculating the number of words to attain a certain level of text coverage.

[5] In my own unpublished test, there is certainly an advantage for English-background learners in understanding English-origin words.

2.3 Features of Japanese writing system and the reviews of studies in characters and vocabulary

In this section, I will mainly review relevant studies on Japanese. I will first briefly introduce 1) the features of the Japanese writing system, characters and vocabulary, followed by brief reviews of some related topics, including 2) text coverage by words and characters, 3) the distribution of word origins and their relationship with register variation, and 4) the distribution of part of speech and its relationship with register variation. The second point is related to Chapters 4, 5, 6 and 8, and the third and fourth points are related to Chapters 4 and 7. The research on domain-specific words will be reviewed in 7.1 in Chapter 7.

2.3.1 Features of writing system, characters and vocabulary in Japanese

The complicated writing system is often mentioned as a unique feature of Japanese. Two types of syllabic characters (Hiragana and Katakana), logographic characters (Kanji), the Roman alphabet, and Arabic numbers can be used together in a sentence. Below is an example.

彼 は いつも 7 時 ごろ ダイニング で 洋楽 を BGM に して 朝ごはん を 食べる。

Kare wa itsumo shichi-ji goro dainingu de yougaku o bi^ji^emu ni shite asa-gohan o taberu.

He/(topic marker)/usually/7 o'clock/around/dining room/in/Western music/(case-marker: accusative)/BGM/take…as/morning-meal/eat

(*He usually has his breakfast around 7 o'clock in the dining room while listening to Western music as background music*.)

In this sentence,ダイニング 'dainingu' (dining room) is five Katakana, 彼 'kare' (he), 時

‘-ji’ (o’clock), 洋楽 ‘yougaku’ (Western music), 朝 ‘asa’ (morning) and 食 ‘ta (beru)’(eat) are Kanji, all the other letters are Hiragana except ７ and BGM. As with the word 洋楽 ‘yougaku’ in this example, Kanji are often combined to make up compound words. The semantic transparency of the component Kanji varies depending on the compound. The word 洋楽 ‘yougaku’ (Western music) is somewhat transparent as 洋 ‘you’ has the meaning of *Western* as in 洋食 ‘youshoku’ (Western dishes), and 楽 ‘gaku’ is also a component of the word 音楽 ‘ongaku’ (music); however, the meaning of 洋楽 is not totally transparent because both 洋 and 楽 have other meanings (洋 also means *sea* and 楽 also means *ease* or *pleasure* with the reading ‘raku’.) Processing individual Kanji is a step to word processing. Therefore, Kanji level processing is important as well as word level processing.

For the acquisition of Japanese vocabulary, especially for non-Kanji-background learners, learning words made up of Kanji, the logographic characters, is a substantial barrier because of its complexity of orthographical and phonological forms, meanings and word formation rules (Toyoda, 2007). The issue with Kanji relates to the acquisition of written language in the first place; however, as Matsunaga (1999) suggests, when developing overall skills in Japanese, phonological processing of Kanji is also important.

Moreover, Japanese Kanji has two types of readings: the On-reading and the Kun-reading which can be mutually connected in the mental lexicon mediated by the identical orthographic form. The On-reading is the pronunciation originating in Chinese and the Kun-reading is the Japanese original pronunciation of the same Kanji which shares the same meaning (Table 2-1). To judge if a Kanji should be read in the On-reading or Kun-reading, in many cases, there are contextual clues such as 送り仮名 ‘okuri-gana’ for Kun-reading. (Okuri-gana is generally Hiragana added to a Kanji. Okuri-gana consist of a word together with Kanji and indicate the word is Japanese-origin. In the example above, べる ‘beru’ of 食べる’taberu’ (eat) are okuri-gana. In this case, the character 食 means ‘eat’

while べる ‘beru’ does not carry any specific meaning but is merely a part of the word. There are also some cases that are hard to judge whether a Kanji is read in the On-reading or the Kun-reading, or even can be read in either of the two (e.g. 腕力 ‘wanryoku’, ‘ude-jikara’ or ‘kaina-jikara’ (arm strength)).

Table 2-1 On-reading and Kun-Reading

(Chinese morpheme)	/chu/ 初 : first, beginning
* Sino-Japanese word (Kango)	最初 /sai-**sho**/ = **On-reading**
(/**sho**/ of /sai-sho/ is adapted from Chinese /chu/)	
(Japanese morpheme)	/hajime/ はじめ : first, beginning
* Japanese-origin word (Wago)	初め /**haji**-me/ = **Kun-reading**
(The word 初め only shares the meaning and character but not pronunciation with Chinese /chu/ 初)	

Adult native Japanese users are generally expected to be able to judge if a pronunciation for a Kanji is the On-reading or the Kun-reading since On-reading and Kun-reading have considerably different phonological structures. For example, the second syllable of a two-syllable Kanji has only eight types, namely /i/, /u/, /ki/, /ku/, /chi/, /tsu/, /N/ (ん) and /Q/ (っ) (double consonants). In addition, these phonological differences will consolidate users’ awareness of the relationship between the word origin and register variation, that is, Chinese-origin words (On-reading words) are often used for formal domains and Japanese-origin words (Kun-reading words) are used more for informal domains. For example, in a formal situation, a Japanese speaker will say 集会を延期した for (*We*) *postponed the assembly* while s/he will say 集まりを先に延ばした in a casual daily-life domain. In this case, 集会 ‘shuukai’ (assembly) and 延期する ‘enki-suru’ (postpone) are Chinese-origin (On-reading), and 集まり ‘atsumari’ (assembly, gathering) and 延ばす ‘nobasu’ (postpone) are Japanese-origin (Kun-reading). Note that these two

pairs of words share the same Kanji but have totally different pronunciations. Proficient users of Japanese are thought to have the links between the On-reading and Kun-reading with a single Kanji orthographic representation in their mental lexicon so that they often switch from one to the other depending on the situation. Because of this relationship, it can be predicted that, learning different words linked with a Kanji will help learners learn both written and spoken knowledge of Japanese vocabulary[6]. Inversely, learning Japanese vocabulary may not be efficiently facilitated without this kind of linking.

Also, each individual Kanji has high productivity in compound words which makes the problem more complicated. According to my calculation using the database I developed for this study, 10,053 words[7] (50.3%) are Chinese-origin words and 9,251 words (46.2% of the top 20,000 and 92.0% of Chinese-origin words) are two-Kanji compounds[8]. This result means that there are a large number of Kanji compounds which are combinations of a limited number of (approximately 2,000) Kanji. Each individual Kanji is not always a word but often a component of words, many of which are transparent to some degree, that is, it is possible to infer the meaning of the whole word from the meanings of individual characters. Many Kanji have plural readings which can be connected in the mental lexicon. Therefore, it is important to investigate how many words are covered by how many characters[9].

These relationships also provide an interesting perspective on second language acquisition (SLA) research on Chinese learners of Japanese (or Japanese learners of

---

[6] Toyoda & McNamara (2011) investigate semantic processing of different Kanji sharing the same component by L1 and L2 readers and found L2 semantic processing skills approximate those of L1 readers with increased L2 script knowledge. From this result, they suggest that processing skills with related words sharing a Kanji will also be an interesting topic for further research.

[7] It is counted by the lexeme which is the unit of counting adopted for this study. It is a similar unit to lemma. For more details, see 3.3.3 in Chapter 3.

[8] This is counted based on the Vocabulary Database for Reading Japanese developed for this study. Two-Kanji compounds account for around 13 % text coverage of the Balanced Contemporary Corpus for Written Japanese used for this study. For the details of this database and the corpus, see Chapter 3.

[9] This issue is to be explored in Chapter 6.

Chinese), because Kanji, the logographic character, carries certain meanings but less phonological information. Many teachers of Japanese also know that CBLs read Kanji vocabulary visually and understand its meanings even if they cannot understand the words aurally let alone pronounce them in the target language (Japanese). The same thing often happens when Japanese learners learn the Chinese language. Thus, it is easily predicted and often discussed among teachers of Japanese as a second language that the gap between the knowledge and skills of written and spoken language is larger in CBLs than non-CBLs. CBLs tend to be better at reading compared to their level of listening (Komori, 2005; Noguchi, 2008). This kind of gap caused by the unique relationship between the languages sharing logographic characters seems an aspect not explored in SLA studies of other languages.

This study only focuses on written vocabulary as it aims to provide a basis for measuring knowledge of written vocabulary for future study. By separately focusing on written and spoken lexical knowledge, the relationship between written lexical knowledge and various language skills can be measured.

In sum, both phonographic (syllabic) and logographic (morphographic) characters are used for Japanese orthography. The logographic character Kanji has two types of pronunciation: the On-reading (Chinese-origin) and the Kun-reading (Japanese-origin). The phonological structures and registers of the Chinese-origin words and Japanese-origin words are considerably different. However, different pronunciations are expected to be linked together with a Kanji and its meaning in proficient users' mental lexicon. Also, a limited number of Kanji consist of numerous Kanji compounds; therefore, it seems important for learners to connect different pronunciations with each orthographic form of Kanji. Kanji also create various gaps in learning Japanese vocabulary between written and spoken uses as well as between Chinese and non-Chinese background learners. Therefore, we should assume that written and spoken languages are basically different languages

because listening and reading require considerably different knowledge even for the same word.

2.3.2 Text coverage by words or characters

The most widely spread cumulative text coverage data are from NLRI, the National Language Research Institute (1962, p 26). This has been cited as data for 'general' Japanese for a long time (e.g. Akimoto, 2002; Tamamura, 1984); however, it is questionable whether it can be representative of text coverage of Japanese in general as it is merely based on a set of magazine data published in 1956. It shows 60.5% of the magazine texts are covered by the most frequent 1,000 words[10], 70.0% by the top 2,000, and 81.7% by the top 5,000. These figures are much lower than English and other languages (Tamamura, 1984, p 101).

I myself calculated cumulative text coverage from digitized data from NLRI (2006). This is a word frequency list also made from magazine texts, but published in 1994 which is 28 years later than the data in NLRI (1962). The result is almost the same as NLRI (1962). 59.8% of the words in the texts are covered by the top 1,000 words, 68.8% are covered by the top 2,000 words and 80.1% are covered by the top 5,000 words.

There are also text coverage data from newspaper texts (NLRI, 1970, p 30). The most frequent 1,000 words provide much higher coverage at 73.5%, the top 2,000 words cover 79.9%, and 5,000 words cover 87.6%. These figures are at a similar level to coverage in English (Nation, 2001, p 13–17): however, to the best of my knowledge, this data is not cited in introductory textbooks on Japanese lexicology.

In Japanese studies, there has not been cumulative coverage or frequency data from a large book corpus; however, at least, it is clear that the coverage data will vary depending on the type of texts.

In Japanese, there are also many data on coverage by single characters as Kanji is

[10] The unit of counting is a unit similar to the lemma which consists of a headword and some of its inflected and reduced forms (Nation, 2001).

thought to be an important unit of learning in Japanese. NLRI (1963, p 9) reports that the most frequent 500 Kanji provide 74.5% of the total Kanji tokens in magazine texts. (Note that it is not the text coverage.) It reaches to 90% by the most frequent 1,000 Kanji. On the other hand, the most frequent 1,000 Kanji cover 95 % of the total Kanji tokens used in newspapers (Nozaki, Yokoyama, Isomoto, & Yoneda, 1996; Yokoyama, Sasahara, Nozaki, & Long, 1998). Both coverage by words and characters provide evidence that magazine texts are more diverse in vocabulary and Kanji use. Long & Yokoyama (2005) used four different corpora including texts from newspapers, encyclopaedias and fiction and found the former 1945 'common Kanji' (常用漢字 'Jo^yo^-Kanji', designated by Agency for Cultural Affairs (文化庁) in 1981) account for 97-98% coverage of Kanji tokens in newspapers and encyclopaedias but only 94.4% in fiction texts. This suggests that literary works will contain more low-frequency Kanji.

Understanding a single Kanji of a two-Kanji compound does not mean understanding the word; therefore, text coverage should be calculated by the word in principle to investigate the relationship between the level of reading comprehension and its related factors. However, Japanese has many semantically-transparent compounds whose meaning can be understood or inferred correctly if the component Kanji are known. For example, the word 砂場 'sunaba' (sandbox) is a low-frequency word ranked at 21,237 (Matsushita, 2011a); however, if the words 砂 'suna' (sand) (ranked at 2,726) and 場所 'basho' (place) (ranked at 318) are known, the meaning of 砂場 will be inferred correctly, or at least learned relatively easily. Considering the fact that the top 2,000 Kanji can cover more than 98.6% and 99.7% of Kanji tokens in magazine and newspaper texts respectively (Chikamatsu, Yokoyama, Nozaki, Long, & Fukuda, 2000; NLRI, 1963), it is expected that a limited number of Kanji will cover tens of thousands of words. At least for understanding written texts, it will be useful to know how many Kanji (and other phonographic characters i.e. Hiragana and Katanaka) will provide how high a text coverage by words. This will

enable us to investigate the relationship between the number of known Kanji and reading comprehension.

In sum, text coverage by words is often cited from the data made from magazine texts but not from other texts; however, the coverage figure will be considerably different from domain to domain. There are also many studies on text coverage by character; however, the relationship between the coverage by words and by characters is not clear yet.

2.3.3 Word origins and register variation

In Japanese corpus linguistics, the proportion of word origins in different types of texts has been a topic explored in many studies, probably because it is related to stylistics and lexical changes of Japanese language.

Ito (2002) is a relatively recent study which deals with the relationship between the proportion of word origins and stylistic features of texts. He uses five different corpora including high school textbooks in science and social studies, magazines, educated spoken language, popular song lyrics and children's stories. The result shows that the proportion of Japanese-origin words ranges from 42.2% (textbooks) to 78.0% (children's stories) while the proportion of Chinese-origin words ranges from 55.1% (textbooks) to 18.7% (children's stories). He concludes that Japanese-origin words account for more high-frequency basic words while Chinese-origin words are less basic. He also concludes that the proportion of Japanese-origin words can be a better index for colloquiality than Chinese-origin words as the proportion of Chinese-origin words in pop song texts is exceptionally low as the texts contain many Western-origin words instead.

As for the change of Japanese, one frequent topic is the increase of Western-origin (mostly English-origin) words. Yamazaki & Onuma (2004) show that the proportion of Western-origin words greatly increased from 9.8% to 35.8% of total lemmas (異なり語数) and from 2.9% to 12.2% of total tokens (延べ語数) in magazine texts during the period

between 1956 and 1994. Loanwords are generally thought to be peripheral vocabulary and not to be basic words; however, Kim (2011) examines the process of shifting some loanwords to basic words. The increase of Western-origin words is an important change for teachers of Japanese, because, as discussed in 2.2.3, Western-origin words can be an advantage in understanding and learning Japanese vocabulary.

As for Chinese-origin words, there are some studies which attempt to count what proportion of Chinese cognates are in Japanese vocabulary. These will be reviewed in 4.5.

2.3.4 Part of speech and register variation

Distribution of parts of speech is a major method for identifying register variations. In Japanese studies, Kabashima's law (Kabashima, 1955, 1981) is well-known on this topic. He first excluded function words and categorised the other parts of speech into four groups of 1) nouns, 2) verbs, 3) adjectives, adjectival nouns ('keiyou-doushi' 形容動詞), adverbs and prenoun adjectivals ('rentai-shi' 連体詞), and 4) interjections and conjunctions. He detected regular relationships on the proportions between nouns and the other three and created three formulae. Those formulae largely tell us that the more nouns in a text, the fewer the others. To be precise, the proportions of 1) nouns and 2) verbs or 4) interjections and conjunctions is not expressed in a linear function formula so the logarithm is used for these formula (Kabashima, 1981, p 132–134). The proportions of 1) nouns and 3) adjectives, adjectival nouns, adverbs and prenoun adjectivals are in inverse proportion (linear function). Kabashima claims that the proportion of nouns will increase when writing is done with word limits as nouns carry essential information, thus, they differentiate the registers. For example, the proportion of noun is high in Haiku and newspaper headlines as they have a strict word limit. In other words, parts of speech other than nouns are used for adjusting redundancy. Nouns carry the most important information.

Nishimura (2010) also tries to identify register variations by examining the

proportions of parts of speech in the process of exploring the features of online language use. One feature with her study is that she examines the proportions of sub-categories of function words based on a detailed classification. For example, she found that the proportions of case particles ('kaku-joshi' 格助詞) and れる/られる reru/rareru (a kind of auxiliary verbs which indicates passives/potentials/spontaneous/honorifics) increase as the proportions of adverbial particles ('fuku-joshi' 副助詞) and sentence-final particles ('shuu-joshi' 終助詞) decrease.

2.4 Making a word list

The main purpose of this section is to review relevant literature on the method for making a word list and clarify the points to consider when making a Japanese word list.

There are many word 'frequency' lists in many languages (e.g., BLI, 1986; Eaton, 1940; Juilland, Brodin, & Davidovitch, 1970; Juilland & Chang-Rodrigues, 1964; NLRI, 1962, 2006; Thorndike & Lorge, 1944; Xiao, Rayson, & McEnery, 2009) and some suggestions and practices for using adjusted frequency (or "usage coefficient[11]) (Lyne (1985) and Gries' (2008, 2010) comprehensive review and comparison of indices to be referred in 2.4.2.). However, most of them are the products of a word list with a simple explanation on how the list was created, or the arguments on how mathematically and/or psychologically valid and reliable a word list can be with a specific index. For the purpose of language learning and teaching, to the best of my knowledge, Nation & Webb (2011) seems to be the only comprehensive description which shows how a word list should be made and deals with particular issues with making a word list[12].

Nation and Webb describe six 'steps involved in making a word list' (p. 135-144;

[11] The terms "adjusted frequencies" (Gries, 2008, 2010) and "usage coefficient" (Juilland & Chang-Rodrigues, 1964; Juilland, Brodin, & Davidovitch, 1970; Lyne, 1985) are used in similar contexts.

[12] The vocabulary selection movement arose in the 1920s (Richards, 2001, p 8) and the most significant outcome is Michael West's A General Service List of English Words (West, 1953).

Table 3-1). To summarize, the steps are 1) research question or reason, 2) unit of counting, 3) corpus, 4) criteria for counting words and separate lists, 5) criteria for ordering words and 6) cross-checking the list. The steps deal with making an English word list; however, many of the ideas can be applied to making a word list in another language, with some considerations of the differences between the particular language and English.

In this section, I review important studies on the points of 2) unit of counting, 4) criteria for counting words and separate lists and 5) criteria for ordering words. Specific issues with making Japanese word lists will be presented in this section, but the issues will be discussed in more detail in Chapter 3 to make decisions on how the words should be ordered.

2.4.1 Unit of counting

There are different levels of units to count, namely, the word type, lemma or word family in English. But the idea of lemma and word family does not seem always applicable to Japanese as the structure of the language is different. There are two questions here: 1) What unit is suitable to measuring the written receptive knowledge of Japanese vocabulary? 2) What are the unique issues with making a Japanese list? Which methods or ideas for making an English list can or cannot be applied to making a Japanese list?

Nation & Webb (2011) claim that an inclusive unit such as word family is most suitable for counting receptive knowledge (p.136)[13]. The idea is that if one or two members of the word family are known, little learning is required for receptive use (comprehension) of other family members. For example, if the word *accessible* is known, it is not difficult to

[13] Leech, Rayson, & Wilson (2001) adopt the lemma which only includes the inflections as the unit of counting with no explanation of the reason (p.4). Carroll, Davies, & Richman (1971) adopt the word type, but they also admit that another unit may be suitable for some purposes (p.4). Vermeer (2004) who aims to measure productive knowledge adopts the lemma as the unit of counting (p.179).

understand *accessed* or *accessibility*. The family of *access* includes the members of *accessed, accesses, accessing, accessibility, inaccessible* and *inaccessibility*. In Nation's list made from the British National Corpus, the word families are set at Level 6 in Bauer & Nation (1993) scheme (Nation, 2004, 2006, 2011). This level includes the inflections and the high-frequency, regular, productive and transparent derivational affixes.

The idea that an inclusive unit is suitable for counting receptive knowledge seems reasonable and applicable to making a word list in any language. The ultimate goal of this study is to make some contribution to decreasing the burden of learning vocabulary. If little learning is required for understanding a form, it should be included under the related headword[14].

One problem with this unit is that it does not seem to be easy to make a consistent judgment about what form is included in a family. We have to set criteria to judge if a derivational affix is high-frequency, regular, productive and transparent.

When we apply the idea to Japanese, there is an issue with the nature of Kanji. Each Japanese Kanji has its meaning so that it generally has a strong compounding power. Therefore, it is sometimes hard to decide if a constituent of a form is an affix.

As mentioned in 2.3, the fact that many Japanese Kanji have their On-reading (Chinese-origin) and Kun-reading (Japanese-origin) makes the problem more complicated. Nakano & Nomura (1979), who work on the morphological analysis of large Japanese corpora at the National Language Research Institute (NLRI), also admit that there can be no clear criteria for distinguishing between a word base and an affix in Japanese (p.861). They point out that many of the On-reading (Chinese-origin) units with a single Kanji are problematic, because most morphemes with a Kanji function like a word base semantically while they cannot be an independent word but can be a stable unit when combining with

[14] This idea is basically in line with "the learning burden principle" (Nation & Webb, 2011, p 137) to be mentioned in 2.4.2.

another morpheme. Therefore, if a morpheme with a single Kanji cannot be a word even with its Kun-reading (Japanese-origin), it is reasonable to judge that the form is NOT a word. For example, 教 'kyou'(teaching) is not a word but 教室 'kyoushitsu' (classroom) is a word, because the former is not a free form while the latter is.

On the other hand, there are many Kanji whose Kun-reading can be an independent word (free form) while its On-reading can only be a constituent of a word (bound form). For example, When we read 山 as 'yama', it can be an independent word (mountain); However once it is read as 'san', it appears to be a suffix (Mt.) as it is a high-frequency, regular, productive and transparent bound form as in 富士山 'Fuji-san' (Mt. Fuji) and 御岳山 'Ontake-san' (Mt. Ontake). According to Nation's criteria, 富士 'Fuji' (Fuji) and 富士山 'Fuji-san' (Mt. Fuji) are members of the same word family. Nevertheless, based on this rule, many more affixes must be identified in Japanese than in English. In other words, these forms are judged as affixes from the syntactical viewpoint while they work like a word at the semantic level. Taking the burden for learning the affixes into account, including the derived forms (e.g., 富士山 'Fuji-san' (Mt. Fuji)) in the same family as the word base (e.g., 富士 'Fuji' (Fuji)) does not seem practical.

In addition, it is sometimes difficult to decide if a form is an On-reading or a Kun-reading. For example, 富士山 is sometimes read as 'Fuji yama', and 岩木山 can also be read as either 'Iwaki-san' or 'Iwaki yama'. 腕力 can be read as 'wanryoku', 'ude-jikara' or 'kaina-jikara'. This is a unique issue with Japanese. In these cases, 'san' and 'yama' must belong to different families from the general (Western) linguistic viewpoint where the 'form' means phonological form in general; however, in Japanese written language, one Kanji can be read in two or more ways as shown above. It seems more practical to judge that a pair of readings with a Kanji is one word, particularly where the Kun-reading can be an independent word.

It is also difficult to judge the degree of productivity. A bound form 力 'riki' (power)

of 眼力 ‘ganriki’ (insight) can also form words such as 怪力 ‘kairiki’ (superhuman strength), 馬力 ‘bariki’ (horse power) and 百人力 ‘hyakuninriki’ (tremendous strength), yet its productivity is not as high as 力 ‘ryoku’ (power) of 抵抗力 ‘teikouryoku’ (resistibility) and 理解力 (ability to understand). Thus it is hard to judge if it is a suffix.

In this case, ‘riki’ should probably not be judged as a suffix as the other components of the compounds ‘gan’, ‘kai’ and ‘ba’ are also bound forms (but ‘hyakunin’ is a free form which is exceptional, though). And ‘chikara’, the Kun-reading of 力, is a single word, so that it seems reasonable and practical, at least when analysing written Japanese, to judge that 力 is an independent unit of counting regardless of its reading when it is not combined with a bound form. It can be a suffix as well as a single word.

Nakano & Nomura (1979) also conclude that there cannot be an ‘across-the-board’ rule for a single Kanji with an On-reading such as 車 ‘sha’ (car) of 汽車 ‘kisha’ (train) and 乗用車 ‘jouyousha’ (passenger car) or 性 ‘sei’ (-ty/-ness/condition) of 酸性 ‘sansei’ (acid) and 国際性 ‘kokusaisei’ (internationality). They claim that the form with a clear and substantial meaning should be judged as a word base while the formalized constituent of a form should be judged as an affix.

Overall, Japanese has more affixes than English. Bauer & Nation (1993) identified only 91 affixes from Level 1 to Level 6 (p.262) while Nakano & Nomura (1979) identifies 250 Sino-Japanese prefixes. Besides those, there must be hundreds of Sino-Japanese suffixes and non-Sino-Japanese affixes. Given the fact that these affixes require learning of the form and the substantial meaning, these should also be a unit of counting.

When we analyse Japanese, one practical problem is word segmentation since there is no space between words in Japanese. In fact, we have to use a morphological analyser on a computer for the word segmentation, which means we have no choice but to follow the definition of the dictionary used by the analyser. There are several dictionaries for

morphological analysers, but the most precise and complete one currently is UniDic (Den, Yamada, Ogura, Koiso, & Ogiso, 2009). The developers claim that it follows consistent rules for defining the units and the identity of indexes while other dictionaries reveal lots of problems such as unevenness in defining a unit and failure in handling allomorphs and orthographic variants (Den et al., 2007, p 102–106).

UniDic adopts two units to count: the short unit (短単位) and the long unit (長単位) (Den et al., 2007, p 106–108)[15]. The short unit allows only one combination of two minimal semantic units in principle (e.g., 外 'gai' + 来'lai' = 外来 'gailai') with exceptions that one minimal unit is counted as one short unit or three or more minimal units are counted as one short unit[16]. The long unit allows a longer combination such as 外来語仮名表記 'gairaigo-kana-hyouki' (orthography of loanwords in Kana) or 調査する 'chousa-suru' (to investigate).(For the full set of rules of the units, see (Ogura, Koiso, Fujiike, & Hara, 2009).)

The short unit meets the purpose of this study as it is more inclusive. The long unit seems more suitable for counting productive knowledge as it distinguishes the different conjugated forms. (The dictionary for the long unit is likely to be published soon, but is not available yet.) A further positive feature is that the result of counting by the short unit is comparable with previous studies, because it is developed from and similar to the β unit used in many other studies such as NLRI (1962).

The "multiword unit" is another issue with the unit of counting. Leech, Rayson, & Wilson (2001) identified some sequences of orthographic words such as *so that* and *in spite of* as multiword units to be counted as single words, because they function grammatically as single words. This seems a reasonable idea from the "learning burden principle" (Nation &

[15] UniDic also has the middle unit (中単位), but the dictionary for the unit is not planned to develop (Den et al., 2007, p 107–108).

[16] This seems to be a practical decision because the number of two-Kanji compounds is overwhelmingly more than single Kanji words or words with a combination of three or more Kanji.

Webb, 2011, p 137) to be mentioned in 2.4.2.), because their meanings are not always as transparent as learners cannot easily guess what they mean so that each multiword unit requires some degree of additional learning.

One problem with multiword units is a consistent judgment about identifying multiword units and another is judging their degree of compositionality. More practically, it is hardly feasible for a single researcher to do the task from a large corpus. If multiword units (e.g. *so that*) are counted as single words, at the same time, it is also necessary to omit the frequency counts of the components of the multiword units (e.g. *so* and *that*) by the number of words used for the multiword units. It is extremely time-consuming without a computer program to do the task. The practical solution will currently be counting multiword units separately.

In sum, an inclusive unit is suitable for counting receptive knowledge; however, there are several issues with counting Japanese words. One problem is that it is difficult to judge if a unit is a word base or an affix, especially a unit composed of a single Kanji. More affixes occur in Japanese than in English, and those affixes will also be a unit of counting when counting 'words' in Japanese since most Japanese affixes require learning of the form and the substantial learning of the meaning. The most practical solution is to adopt the 'short unit' identified by UniDic (Den et al., 2009).

2.4.2 Criteria for counting words and separate lists

Nation & Webb (2011) claim that decisions about whether a form is counted as a known word should depend on "the learning burden principle", that is to say, "If it does not require previous knowledge (as is the case with most proper names), or it can be figured out from previous knowledge (as is the case with some derived forms and compounds), then it should not be a headword in the lists" (p.137-138). According to this idea, they investigate

transparent compounds (e.g., *lifespan*), proper names, non-words and marginal words (e.g., *eh*), foreign words (e.g., *précis*), abbreviations (e.g., *STD*), homonyms and homographs (e.g., sow ([sou] for sow seeds/ [sau] for female pig)), and then decide to create separate lists for transparent compounds, proper names and non-words and marginal words. The value of separate lists is that they most clearly show what decisions were made and allow adjustment without reading the other lists.

As mentioned at the beginning of this section, no study except for Nation & Webb (2011) deals with this issue, probably because few researchers have paid attention to how, when checking the text coverage by known words, the previous knowledge required to understand the meaning of a word will differ according to the type of word.

2.4.3 Criteria for ordering words

The purpose for ordering the words, which is typically done in the form of word frequency lists, is to show in what order learners should learn them. Then what should be the criteria for ordering them?

The simplest but most powerful idea is that the more words a learner knows in the text, the more effective comprehension becomes. In other words, the higher the text coverage by the known words, the better. Based on this idea, frequency is the most important criterion to order the words. If a learner learns high-frequency words first, s/he can gain the highest text coverage more efficiently.

Then, how can we measure frequency? If a corpus could be designed for each individual learner and the frequency of the words could be checked in the corpus, that would best suit the learner's needs. Yet, this is not a practical idea. To be practical, we can only categorize learner needs and design a corpus to meet each category of needs.

Suppose there are "general learners", what type of people are they? What kind of language do they need to use? On the one hand, learners have different interests and

language use, so that it is not easy to match the corpus domain with each learner's needs. On the other hand, some words are unevenly distributed in a particular domain, even if the whole corpus is a balanced corpus made by a strict sampling procedure. Therefore, to reflect the generalized learners' needs on lexical frequency figures, various types of 'dispersion' indices are often used as a mathematical manipulation. Dispersion indicates how widely and evenly a word is distributed.

It is essential to use a spoken corpus to select basic vocabulary based on frequency data but not solely by subjective selection. There seems to be a general agreement that high-frequency words used in a wide range of domains should be selected for basic vocabulary in principle[17]. Also, there are other factors to select basic vocabulary such as ease or difficulty of learning, necessity and coverage of semantic field (Richards, 1974; West, 1953)[18].

To judge which index is appropriate for ordering words, the relevant literature on the construct of the whole vocabulary of a language and specific statistical indices for dispersion and adjusted frequency (or "usage coefficient") are reviewed below.

2.4.3.1 The construct of vocabulary knowledge in the language as a whole in terms of word frequency and dispersion

Frequency is a very important index to order the words in general, but dispersion seems as important as frequency. Let us look at what dispersion is and why it is important.

---

[17] Nation & Webb (2011) are concerned that criteria other than calculations such as frequency or *range* are often applied in an ad hoc rather than a principled way (p.148).

[18] West (1953) refers to five factors (other than frequency) which are considered to be vocabulary selection. Those are: 1) Ease or difficulty of learning (= Cost), 2) Necessity, 3) Cover, 4) Stylistic level, 5) Intensive and emotional words (which West claims are of secondary importance for foreign learners.) (p. ix-x). In the context of making a word list for South Asian countries, Richards (1974) proposed four principles: a) Frequency and range, b) Availability and familiarity (e.g., concrete words which are easy to recall), c) Coverage (e.g., words needed for basic science concepts), d) Meaning priorities (p.79).

There are wide and narrow usages of the term ‘dispersion’. In the wider sense, whether the frequencies in the sub-corpora are counted (e.g., Juilland’s *D* (Juilland & Chang-Rodrigues, 1964)) or not (‘*range*’ or ‘document frequency’), indices which show how widely a word is used are all called dispersion (Leech et al., 2001, p 17–18). In the narrower sense, dispersion does not include *range* but just means indices which sub-frequencies are used to calculate. In this thesis, following Leech et al.'s (2001) wide sense, all of these are defined to be a kind of dispersion which is used in addition to frequency.

It is taken for granted that learning high-frequency words earlier is a good way to gain text coverage efficiently. Nevertheless, even a high-frequency word may not be so useful for a learner if it is used only in a limited domain not related to the learner.

Gries (2008, 2010) argues from a psycholinguistic viewpoint that frequencies in isolation are not perfect predictors of aspects of processing but can also be misleading, because there are different distributional patterns. He, therefore, advocates the importance of a dispersion measure. Nation & Webb (2011) also claim that the *range* of a word, which is one of the dispersion measures based on the definition here, is more important than frequency because the most generally important words are used in a wide range of texts (p.142).

From the viewpoint of text coverage, if a word list contains a lot of unevenly-used words, text coverage can only be higher in limited domains. Supposing there are learners with broad learning goals, who will encounter various texts in various domains, it is necessary to identify the important words whatever the learners’ major or needs domains are. To do this, it is necessary to introduce a dispersion measure which shows how evenly a word is distributed in different domains. If dispersion is used in combination with frequency, narrowly-ranged words can be downgraded properly in order to gain higher average text coverage with various texts in various domains.

Dispersion is expected to have a high degree of correlation with frequency. Carroll

(1971) reports that D_2[19], a measure of dispersion over 17 subject categories, correlates with the logarithm of *F* (total frequency) at .8538 in a sample of 56 words of widely varying frequency (p.xxix). It may look high in general; however, D_2 only accounts for 73% of the variance of the total frequency. This shows that some words are unevenly distributed. Generally speaking, high-frequency words have high dispersion, that is, they tend to be used in a wide range of domains. Some high-frequency words have low dispersion as they are used in more limited domains than others. Also, as Gries (2010) points out, some dispersion measures may vary depending on the number of sub-sections. It should be noted that the greater the number of sub-sections is, the higher the correlation between dispersion and total frequency will be.

It is also important to measure "general" frequency and sub-frequencies as well as dispersion to identify keywords or domain-specific words in a text or a domain. In keyword studies, a keyword is generally defined as a word without which readers cannot understand the whole passage, in other words, a word which carries a greater amount of information than other words in the text (Kabashima, 1981, p 119–125). Keywords are generally extracted by some keyness index (e.g., log-likelihood ratio), that is, words which have a much higher frequency in a particular text than in a collection of texts are regarded as keywords. This means that keywords in a passage are generally low-frequency words in a collection of texts and low dispersion words. There seems a trade-off between general importance and keyness in a text or a domain. Inevitably, in any sense, to measure the general importance is essential to identify specificity. The construct of sub-sections is also an important issue because the meaning of generalness and specificity will change depending on the construct of sub-sections.

Nonetheless, in many previous studies in Japanese linguistics, frequencies in a magazine corpus (NLRI, 1962) have been substituted for "general" frequencies (e.g. NLRI,

[19] Carroll's dispersion index is known as D_2 (Carroll, 1970, p 62) which is calculated by a different formula from Juilland's *D* (Juilland & Chang-Rodrigues, 1964; Juilland, Brodin, & Davidovitch, 1970).

1984; Tamamura, 1984). What is more, there is little discussion of the classification of sub-sections. To the best of my knowledge, no existing Japanese word list has been created totally based on a combination of objective criteria including dispersion. Excluding or downgrading unevenly distributed words all depended on subjective judgement by so-called experts (e.g., Butler, 2010; Japan Foundation & Association of International Education, Japan, 2002; Komiya, 1995; Muraoka & Yanagi, 1995; Oka, 1992; Tamamura, 1987). That was mainly due to the limitation of workload. Nowadays, we should pursue more objective ways to make word lists based on frequency and dispersion as computer technology has been developing[20].

In sum, dispersion is used to measure how widely and/or evenly a word is used, and it includes *range* and other indices in this thesis. Dispersion is vital for identifying the general importance of a word, which is inevitably important for identifying specificity in a text or a domain.

2.4.3.2 Indices for dispersion and adjusted frequency

The use of frequency and dispersion to rank words in a large corpus has at least a fifty-year history (Carroll, Davies, & Richman, 1971; Juilland et al., 1970; Juilland & Chang-Rodrigues, 1964; Leech et al., 2001; Lyne, 1985; Nation & Webb, 2011). There have been a few indices for dispersion which are used to calculate various types of adjusted frequency (or "usage coefficient") to decide on the ranking of words. Gries (2008, 2010) warns that researchers should be more aware of the differences of different indices and the importance of empirical validation studies on a large corpus, and offers a comprehensive review and some empirical studies of the indices.

[20] To calculate dispersion, sub-frequencies must be counted which is nowadays done by computer programmes such as AntWordProfiler (Anthony, 2009) which is adapted from Range (Nation & Heatley, 2002). AntWordProfiler was only available for alphabetical characters before Version 1.200w, but has been available in Unicode (UTF-8) since Prof. Anthony improved it by accepting my request in 2009.

The simplest index for dispersion is *range*[21] which counts the number of sub-sections where the word appears but does not take account of the sub-frequency. That is, the count of *range* is simply the number of sub-sections a word occurs in. Vander Beke (1932) primarily ranked the items according to *range* and secondarily according to frequency (Lyne, 1985, p 101). In Nation's (2006) list made from the British National Corpus, *range* over ten sub-sections was also adopted as the primary criterion to order the words[22].

As Lyne (1985) shows, however, *range* cannot discriminate words with quite different distributions[23]. For example, when a word has sub-frequencies of (25, 25, 25, 25, 25, 25) in five sections and another word has (1, 1, 1, 1, 98), both words are given the range of 5 while their dispersion (Juilland's *D*) figures are 1.000 and .525 respectively (p. 131–144).

In addition, *range* can only be sensibly applied when the sub-sections are equally-sized; however, if the sub-sections are equally-sized designed by genre, the total frequency figure may not be able to account for language users' different levels of contact with different genres. It would be a flaw when we use a balanced corpus where the texts are sampled in a strict way to reflect the reality. Or if we manage to divide the whole balanced corpus into equally-sized sub-sections, then some domains will have more sub-sections than others as people will generally not evenly work within different genres. In this case, the *range* figure will not reflect in how many unique genres the word is used. Given these, it seems that a dispersion measure where sub-frequencies are taken into account is necessary.

One of the earliest mathematical dispersion indices is Juilland's *D* (Juilland &

[21] In the information sciences, it is generally called "document frequency".

[22] See also p. 82.

[23] Lyne (1985) admits a certain degree of practical usefulness of *range* by showing an example analysis (p.133-134).

Chang-Rodrigues, 1964). Carroll (1970) proposed D_2 and Rosengren (1971) proposed *S* as an alternative to Juilland's *D*. (Lyne, 1985) compares *D*, D_2 and *S*, and concludes that Juilland's *D* is the most appropriate dispersion measure. Lyne applied these indices to both fictitious and his own factual data, and concludes that Carroll and Rosengren's criticisms of Juilland's *D* are unjustified or of little practical significance (p.117). Lyne's criticism of D_2 and *S* is mainly on that these indices generally return higher dispersion values than *D* but overpenalise the distribution which includes zero(s) in one or more sub-sections. (Leech et al., 2001) inherited (Lyne, 1985)claim (p.18) and adopts Juilland's *D* as well as *range* to their word frequency list.

Gries (2008) gives a comprehensive review of various dispersion measures including *range*, *D*, D_2 and *S*, and proposes an alternative index *DP* (deviation of proportions). He supports Lyne's claim about the treatment of distribution patterns which include zero(s) in sub-section(s); however, he also points out some flaws of dispersion measures other than *DP*. For example, some indices require equally-sized sub-sections, which is often not realistic. Juilland's *D* is also applied to equally-sized sub-sections in their own data sets (Juilland & Chang-Rodrigues, 1964; Juilland et al., 1970), but both Lyne and Gries claim that relative frequency can be used with Juilland's *D* when the sub-sections are not equally-sized (Gries, 2008, p 411; Lyne, 1985, p 116). Gries also points out some flaws of Juilland's *D* (and some other indices) as below.

a) Juilland's *D*, in some cases, returns a negative value even though its expected value is within the range from 0 to 1.
b) Range of figures of Juilland's *D* and some other measures depend on the number of sub-sections as they divide a value by the number of sub-sections in the process.
c) Juilland's *D* and some other measures are not sensitive enough. For example, Juilland's *D* does not distinguish between the two distribution patterns of (4, 2, 1, 1, 0) and (3, 3, 2, 0, 0).

Gries proposed *DP* which can resolve the problems mentioned above.

However, there are also some concerns about *DP*. First, it is not clear how it can be integrated with a frequency value to compute adjusted frequency. (Contrary to other indices, *DP* gives 0 when a word is totally evenly distributed and gives 1 for the opposite.) This creates a problem since the current study needs a measure to order words according to a value. Second, as Gries himself points out, it does not return the maximal value 1 even when all occurrences are in one sub-section. This may mean it does not have enough sensitivity in some cases. Third, as for his criticism mentioned in c) above, it is not easy to tell which pattern is more evenly distributed. This should not be used to show a lack of discriminatory power without evidence.

Gries (2010) further explores the differences between 29 different dispersion measures by applying them to the spoken component of the British National Corpus World Edition, checks intercorrelations of the measures. The result shows that the dispersion measures are classified into five different clusters. He also applies the measures to check the external validity with some psycholinguistic data but concludes that none of the dispersion measures reaches really high levels of predictive power, which was to be expected.

In sum, among all the dispersion measures, Juilland's *D* and Gries *DP* seem to be the most adequate measures which can be applied to the current study. *DP* seems more valid mathematically; yet, it is not clear how it works when it is applied to large corpus data. Particularly, we should be aware how those indices can contribute to the word rankings which are the central concern for this study.

As dispersion measures vary, there are also quite a few adjusted frequency (usage coefficient) measures, one of which will be the major criterion to order the words for the current study.

Juilland's *U* is simply the product of *F* (total frequency of the whole corpus) and *D* (dispersion): $U = F \times D$. Carroll (1970) devised a complicated formula from the viewpoint of probability to propose the Standard Frequency Index (SFI). As mentioned above, however, Gries (2008) and Lyne (1985) criticise Carroll's dispersion measure D_2, and, as a consequence, do not support SFI, either. Lyne clearly states that he supports Juilland's *D* but points out some problems of *U*. He still prefers *U* rather than other indices available at that time, but proposes that it should be applied to 'undifferentiated' (not-classified-by-genre) sub-sections so that there cannot be many sub-sections which have zero or very low occurrences (Lyne, 1985, p 125–129).

The main problem with *U* is, as Muller (1965) points out, that it does not differentiate the distribution patterns having different frequencies in one sub-section and the same frequencies in the other sections such as (1, 1, 1, 1, 1), (1, 1, 1, 1, 3) and (1, 1, 1, 1, 5), because the latter distribution pattern has higher frequency but lower dispersion, and vice versa (Lyne, 1985, p 125). Particularly, whatever the frequency is in one section, if all the other sections have zero (cases such as (0, 0, 0, 0, 1) and (0, 0, 0, 0, 5)), both *D* and *U* will be zero (ibid.). Lyne claims that Juilland's *D*, which he prefers, reacts more vigorously to the skewness (a measure of the asymmetry of the distribution) of distribution than D_2 and *S* (Lyne, 1985, p 129). This nature might be a flaw, but can also be a strength as it does not react to sampling bias too much while it cannot work well in the very low-frequency range.

Leech et al. (2001) order the words only by the total frequency, and show the dispersion figures of *D* and *range* over 100 sub-sections of the British National Corpus separately from the frequency. They do not adopt any adjusted frequency as a criterion for ordering words. They do not give the reason; however, they may have accepted Lyne's concern about *U* since they accepted Lyne's proposal about the dispersion measure.

As mentioned above, Gries (2008) proposed a new dispersion measure *DP* but did not propose how it can be integrated with the total frequency to develop an adjusted

frequency. Gries (2010) tested a few adjusted frequency measures on two data sets of reaction times from lexical decision tasks by native speakers (Baayen, 2008; Balota & Spieler, 1998), but found no significant difference between the correlation values with the measures so that he reserves his own opinion about it.

In Nation's list (explained in Nation (2006) and Nation & Webb (2011)), words are basically ordered by *range* as the first criterion, and frequency as the second; however, as mentioned, it is not appropriate for differently-sized sub-sections. In addition, it penalizes too much when sub-sections contain zero. Therefore, this approach does not seem appropriate to be applied to the current study.

In sum, for this study, Juilland's *U* and some combination of frequency and Gries DP are possible measures for adjusted frequency to be used for ordering words. Carroll's SFI may also be worth applying.

Let me repeat here: the main concern for this study is the ranking of words, because it shows the order of usefulness of learning. For this purpose, it is more important to check how much an unevenly-distributed word is penalized by each index rather than the mathematical conformation for the whole corpus data. The reasons for penalizing unevenly-distributed words for this study are as follows.

1) An unevenly-distributed word is less important for people who operate within the sub-genres which have less occurrences of the word.
2) In light of a possible application of the "law of diminishing marginal utility" of a word in a text, i.e. the more the occurrences of a word in a text, the less important each occurrence will be. When we compare two words with the same total frequencies, the more evenly-distributed word is likely to have more importance as a whole.

In other words, the degree of importance for ordering words in this study is not only the

matters of the psychological properties or mathematical, statistical behaviour, but also the problem of how the frequencies or usefulness of words should be assessed in different genres at different frequency levels. Text coverage can be one of the criteria; however, a subtle difference of text coverage by a small group of words in a large corpus is not likely to be a good tool for assessing different measures. The fact that there is no good balanced spoken Japanese corpus makes us more pessimistic about the solution to this problem by checking the text coverage. For the time being, a somewhat subjective judgment such as comparing word rankings of unevenly-distributed words by different indices may be a more valid way to judge which measure is more adequate. Experimental rankings are to be examined in 3.3.5 to decide on which index to use this study.

2.5 Application of word lists and Kanji lists

2.5.1 Advantages of word lists and Kanji lists

The purpose of this section is to clarify the potential uses of word lists and vocabulary databases I developed for this study by reviewing previous studies. The advantage of word lists and Kanji lists are basically the same. The difference is only on the unit of learning; therefore, 'word lists' or 'vocabulary lists' in this section include Kanji lists.

Various types of word lists have been created for teaching and learning. The most representative purpose is to show the target words to learn, often with the order of words by importance (typically by frequency). But the advantages of word lists are not limited to these. Nation & Webb (2011) list seven values of word list research as below (p.132-134).

1) Designing courses
2) Setting learning goals
3) Guiding the creation of simplified texts

4) Analysing the vocabulary in texts

5) Analysis of lexical richness

6) Creating specialized word lists

7) Guiding the construction of vocabulary tests[24]

If I apply these seven uses to the so-called Deming cycle which consists of the four steps of Plan-Do-Check-Study (Deming, 1994), we found that the seven uses are mainly useful for planning and checking stages of learning or teaching. Bearing this in mind, I sort out various uses of a vocabulary database and word lists derived from the database, for the possible users, namely learners, teachers, course designers and researchers.

2.5.2 Application to learner-directed learning

Vocabulary lists are useful for learner-directed learning. "Vocabulary is not explicitly taught in most language classes, and students are expected to 'pick-up' vocabulary on their own without any guidance" (Oxford & Crookall, 1990, p 9). There are a couple of possible reasons for this. First, learners' vocabulary needs will vary, especially after learning core vocabulary. Second, it takes too much time and energy for teaching and learning in class. Third, it is often thought that vocabulary is more suitable for self-directed learning than other skills, and it may be true. These sound negative reasons for self-directed vocabulary learning; however, there are also positive reasons.

If a word list suits learner's needs and level, it will facilitate extensive, self-directed, structured vocabulary learning. Gu & Johnson (1996) claim that self-initiation and selective attention in vocabulary learning are positive predictors of both vocabulary size and general proficiency (p.668). Kojic-Sabo & Lightbown (1999) claim the importance of self-awareness, self-monitoring, organization and active involvement of the learner in the

[24] 3), 4) and 6) are exemplified in this thesis in Chapter 8, 4, and 7 respectively. As for 7), a Japanese vocabulary size test was created and the data was collected and analysed, but not included in this thesis.

acquisition process (p.190). Sanaoui (1995) also provides evidence that a self-initiated, extensive structured approach to vocabulary learning is significantly more successful in retaining vocabulary.

In particular, several studies suggest that raising awareness of learners' vocabulary learning strategies is useful (Cohen, 1990; Gairns & Redman, 1986; Hulstijn, 2001). If a word list is provided to learners with suitable suggestions for use, it will be effective. Gu (2003) states "Good learners seem to be those who initiate their own learning, selectively attend to words of their own choice, studiously try to remember these words, and seek opportunities to use them." Merely giving learners a word list as material for rote memorization will deprive them of their own choice; however, using a high-frequency word list as a check list, or a selected specialised list of words with the explanation for selection criteria and usefulness will raise learners' awareness.

As Nation & Webb (2011) suggest, a word list contributes to controlling the vocabulary load of an extensive reading text. Extensive reading is mainly an independent mode of learning as well as a classroom activity (e.g. Mikami & Harada, 2011).

Also, word lists can be uploaded to web-sites for selective use. For learning English, for example, Tom Cobb's Compleat Lexical Tutor site[25] provides word lists with various selective learning devices (Cobb, 1996). For learning Japanese, the Reading Tutor site (Kawamura, Kitamura, & Hobara, 1997) provides the lexical profile of a text on a web page as well as a bilingual glossary using the former Japanese Language Proficiency Test word lists. As the Compleat Lexical Tutor does, if a web-site also provides a self-checking vocabulary test with appropriate feedback, that will also facilitate self-directed vocabulary learning (Matsushita, 2011b). Word lists can also contribute to this.

2.5.3 Application to course design and teaching

All the seven values listed in Nation & Webb (2011) introduced at the beginning of

[25] www.lextutor.ca.

this section can be applied here. Vocabulary learning can be or should be incorporated in the curriculum. Depending on the purpose of learning and the conditions given, we can identify specific words to learn, set a specific number of words as a learning goal and use word lists for checking materials, tests and learner outcomes.

Once the learning goal is set and specific words are identified as target words, we can analyse the vocabulary load of material for teaching. If the lexical level of texts is too high for the learner group, we can simplify the texts and identify possible target words in each text specifically. We can also analyse learners' compositions (or transcribed conversation texts) to detect learners' lexical level. For these purposes, word frequency lists provide essential information. It is important to use the same word lists for checking the vocabulary load of the text, selecting test items and checking learners' language.

Word lists can also be directly applied to teaching. Folse (2011) claims advantages of word lists which match the purposes of learning, based on previous studies which compare studying words in a word list versus various kinds of contexts (Laufer & Shmueli, 1997; Prince, 1996). Townsend & Collins (2008) also show that teaching academic words to middle school students had a significant effect on increasing knowledge of academic words.

2.5.4 Application to research

For research purposes, a vocabulary database and word lists can also contribute to tests and experiments as well as analysing informants' language. In order to develop tests, a well-validated word list is necessary for appropriate sampling of the test items. For experiments, various lexical factors such as frequency, dispersion or word length must be well controlled. A good database and word lists can provide this information (Gilquin & Gries, 2009; Gries, 2010). Frequency data is one of the strong predictors of reaction time. And thus, reaction times can also be employed for validating a word frequency list (New, Brysbaert, Veronis, & Pallier, 2007).

As Nation & Webb (2011) suggest, word lists can also serve for checking lexical

diversity. For example, Laufer (1994) examines learners' lexical development using the Lexical Frequency Profiling (LFP) exploiting word lists.

Word lists are also applicable to exploring register variations. For example, we can check what kinds of texts contain more academic words (or any group of words by part of speech, word origin and/or frequency level). If we do this on groups of texts, we may be able to detect lexical features from particular groups to identify register variations. For example, Ito (2002) is such a study by checking the proportion of word origins.

2.6 Conclusion of Chapter 2

The main goal of this study is to explore the most efficient learning order of words. In this chapter, some theoretical and methodological issues are investigated by reviewing relevant literature. Below is a summary of main points. (Chapters related to the points are shown in square brackets at the end of each point.)

1) The word is an essential unit of language processing. [Basic to the whole thesis]
2) Vocabulary knowledge seemingly accounts for at least 40% of variance in reading comprehension in Japanese. [Basic to the whole thesis]
3) Required level of text coverage for adequate reading comprehension will be at some level between 95% and 98% depending on the purpose. [Chapters 4, 6 and 8]
4) Cognates have a large effect on learning L2 vocabulary in general and in Japanese. The effect is mostly positive at least for short-term. [Chapters 4 and 7]
5) Both phonographic (syllabic) and logographic (morphographic) characters are used for Japanese orthography. The logographic character Kanji, the Chinese character has On-reading (Chinese-origin) and Kun-reading (Japanese-origin). The phonological structures and registers of the two are considerably different. [Basic to the whole thesis]
6) A limited number of Kanji consist of numerous Kanji compounds; therefore, it seems

important for learners to connect different pronunciations with each orthographic form of Kanji. [Chapters 5 and 6]

7) Kanji also create various gaps in learning Japanese vocabulary between written and spoken uses as well as between Chinese and non-Chinese background learners. Written and spoken languages should be studied separately because listening and reading require considerably different knowledge even for the same word. [Basic for the whole thesis but particular important for Chapters 4, 5, 6 and 7]
8) Text coverage by words is often cited from the data made from magazine texts in Japanese studies; however, the coverage figure will be considerably different from domain to domain. There are also many studies on text coverage by character; however, the relationship between the coverage by words and by characters is not clear yet. [Chapters 4, 5 and 6]
9) Japanese-origins words account more for high-frequency basic words while Chinese-origin words are less basic. Western-origin words increase markedly in Japanese these several decades. [Chapters 4 and 7]
10) According to Kabashima's law, some groups of parts of speech decrease as nouns increase. Proportions of some groups of function words can also be indices for register variations as well as nouns etc. [Chapters 4 and 7]
11) For making a word list, an inclusive unit is suitable for counting receptive knowledge; however, it is difficult to judge if a unit is a word base or an affix, especially a unit composed of a single Kanji. The most practical solution is to adopt the 'short unit' identified by UniDic. [Mainly Chapter 3]
12) For some categories of words such as proper nouns which require little previous knowledge to understand, it may be better to create separate lists from a general word list. [Mainly Chapter 3]
13) Adjusted frequency (usage coefficient), which is a combination of dispersion and

frequency, is an adequate measure for ordering words. Juilland's *D* or Gries *DP* seems to be the most adequate measures for dispersion. Thus, Juilland's *U* (product of frequency and *D*) and some combination of frequency and Gries *DP* are possible measures for adjusted frequency. Carroll's SFI may also be worth applying. [Chapters 3 and 5]

14) Word lists (and Kanji lists) have various advantages in learning, teaching and researching (Japanese as) a second language. Applications include self-directed learning, curriculum design, checking vocabulary load of a text, simplification of a text, creating vocabulary tests, controlling variables of experiments and tests, exploring register variations, and so on. [Mainly Chapters 4, 7 and 8]

Chapter 3 Making and validating the Vocabulary Database for Reading Japanese: How should we order the words?

3.1 Introduction

As shown in 2.5, using word lists has a number of advantages in second language learning and teaching. To show what words are necessary for learners to attain a certain purpose, it is essential to refer to vocabulary data based on a corpus which reflects the target domain for the learners. If the frequency list reflects the learner's target domain, the frequency ranking will basically show the most efficient order of learning vocabulary. Word lists will also provide useful data for developing vocabulary tests in the target domain and measuring how difficult or easy the vocabulary in a text is for the learners. For this purpose, it is important to create vocabulary data based on a corpus which has high representativeness of the target domain. If a test is made from biased vocabulary data, the result will be distorted.

Various word lists have been created in the field of teaching and learning Japanese as a second language[26]; however, for the purposes mentioned above, there are some problems of corpus size, age and methods with the existing lists as will be mentioned in 3.2. To resolve the problems, a vocabulary database was created. Based on the database, word lists were created by different combinations of indices.

In the following sections, firstly, significant studies on existing Japanese word lists are reviewed. Secondly, the methods for creating the vocabulary database and the word lists for this research are described in detail. The URL to download the database is also shown. Thirdly, the validity of the word lists derived from the created database is examined. In particular, some indices for ranking words are compared based on the text coverage of some test corpora in different genres. Different weightings on sub-frequencies depending

[26] For a more comprehensive introduction to Japanese word lists, see (Kai, 2000, 2002).

on different purposes will also be examined. Lastly, general and Japanese specific issues with making word lists are mentioned as remaining issues.

The features of Japanese vocabulary arising from the analysis of the database will be described in the following chapters.

3.2 Significant research

3.2.1 Problems with existing Japanese word lists

Among all the Japanese word lists to be made for second language learners of Japanese, the most influential one must be the former Japanese Language Proficiency Test (F-JLPT) word lists (Japan Foundation & Association of International Education, Japan, 2002) made up of the four lists from Level 4 (beginner level) to Level 1 (advanced)[27]. The words were selected by an expert committee; however, the basic references adopted for selecting Level 4 and 3 (elementary) words were eleven types of Japanese elementary textbooks where the vocabulary was subjectively selected. After selecting words which occur in four or more textbooks, the committee made an adjustment to fix the words by checking other references including the National Language Research Institute (NLRI) (1984). This reference is a check list where each Japanese word is checked if it is adopted in seven types of word lists most of which are based on subjective selection. The only objective data of the seven lists was NLRI (1962) and the other six lists are made by subjective selection based on unclear criteria. The selected words overlap to some degree; however a considerable number of words do not overlap. The cause of the differences is not clear because the selection criteria for each list are not clearly described.

[27] The former Japanese Language Proficiency Test (F-JLPT) was conducted from 1985 to 2009. For the current JLPT which started in 2010, new word lists were created. But the lists have not been made public. According to Akimoto & Oshio (2008), the JLPT committee members classify the words subjectively to each of the new five categories from N5 to N1 in reference to the objective data including Amano & Kondo (1999, 2000) and NLRI (2006).

NLRI (1962), the only objective data among the seven lists, is a word frequency list based on a corpus which consists of ninety types of magazines published in the 1950s. This data was cited by many studies such as Ishiwata (1970) and Tamamura (1984) on Japanese as NLRI (1962) was the only large scale general vocabulary survey at that time. But, it contains flaws partly because large corpus research was not developed in Japanese studies before the late 1990's.

First of all, the total number of token in the NLRI (2006) is not enough. It only contains around 533 thousand running words, and the makers consider the frequency is statistically reliable for approximately only 7,200 words[28]. The word ranked at 6,843rd in the list, which is the lowest ranking in the list, only has 7 occurrences. The F-JLPT word list contains around 8,000 words from Level 4 to1, and the test specification stated that approximately 10,000 words including the 8,000 words would be the target vocabulary at Level 1. In the current Japanese Language Proficiency Test, approximately 15,000 words are targeted at N1, the most advanced level (Akimoto & Oshio, 2008). Keeping in mind that the British National Corpus contains 100 million words and the Bank of English contains hundreds of millions of words, a corpus of merely 533 thousand (0.533 million) words is clearly not large enough. In English studies, Brysbaert & New (2009) claim that a corpus of 16–30 million words is needed for reliable word frequency norms for most practical purposes (p.980).

Secondly, existing word frequency lists including NRLI (1962) do not have sub-frequency data which enable us to calculate dispersion or mix the sub-frequencies. Checking the words in order of frequency, words with significantly uneven distribution are found quite often even in the high-frequency range. Taking dispersion into account is necessary to fix this problem. NRLI (1962) also has sub-frequency data on five genres; however, the number of words in the sub-corpora is significantly unequal. There are 57338,

[28] In the NLRI (1962), approximately 780 words are ranked at the lowest ranking where 10 percent of the words are estimated to be missed from the list due to error (NLRI, 1962, p.21, 26, 224-227).

94417, 98608, 97285, and 185135 words in the five genres of 1) Critique/Entertainment and Culture (評論・芸文), 2) Commonalty (庶民), 3) Utility/Popularized Science (実用・通俗科学), 4) Life/Women (生活・婦人), and 5) Amusement/Hobby (娯楽・趣味) respectively[29]. The classification of sub-genres also has a problem as it has a sub-genre titled "General". Moreover, the data is not provided as a digitized version so that it is not possible to process the data electronically.

Thirdly, the NRLI (1962) is too old. It is based on a survey in the 1950s, but the lexical change in Japanese is large. For example, loanwords mainly from English have increased markedly at all the frequency levels (Matsushita, 2009; Yamazaki & Onuma, 2004).

Fourthly, the survey is based on magazines and so it cannot represent general Japanese. There are some indices for register variation and domain-specificity such as the proportions of nouns, verbs and affixes, and the text coverage curves, which indicate the features of magazine texts as relatively casual but containing more words for specific genres and advertisements (Matsushita, 2009, 2010; Nishimura, 2010). Many magazines are edited for people who have special interest in some area such as fishing or golf. It is thus a problem to regard this as typical written language.

The F-JLPT word lists were created by taking all the major word lists at that time into account so it is likely to be better than the others. But it still has the problems mentioned above. In addition, the list excludes the names of foods and vegetables and some place names. This may be because those words are thought to be inappropriate for worldwide testing; in any sense, however, they are still essential words for learners and teachers (Kawamura, 2006). The database and the word list should include those words as well.

After the F-JLPT word lists, among a few published word lists, a word familiarity list (Amano & Kondo, 1999), a newspaper frequency list (Amano & Kondo, 2000) and a

[29] The numbers of words were calculated by adding the numbers of content words and function words based on Table 13 in NLRI (1962, p 314).

magazine word frequency list (NLRI, 2006) are notable and comprehensive. Nevertheless, these do not meet the needs of current learners and teachers, either. The word familiarity list misses many new words and low-frequency words as the measurement was only done for the words contained in a dictionary. The other lists are not sufficient, either. Considering the lexical features of those genres, either newspaper word lists or magazine lists cannot be representative of the whole Japanese vocabulary. Newspapers are too formal while magazines are too casual and contain too many domain-specific words (Matsushita, 2009, 2010; Nishimura, 2010) (Lexical features of different media will be mentioned in Chapter 4). In addition, these two have too many current words which may not be used so frequently after a certain period of time. For example, the words 政府 'seifu' (government) and 国民 'kokumin' (member of a nation) are ranked at 91st and 205th respectively in the newspaper list (Amano & Kondo, 2000), 520th and 559th in the list made from books and internet-forum sites (Matsushita, 2011)[30] and 1457th and 1487th in the magazine list (NLRI, 2006). These words occur more in newspapers. On the other hand, the words like 楽しむ 'tanoshimu' (enjoy) and タイプ 'taipu' (type) are used more in magazines. They are ranked at 2185th and 3078th respectively in the newspaper list (Amano & Kondo, 2000) and 834th and 900th in the list made from books and internet forum sites (Matsushita, 2011), while they are ranked at 292nd and 240th in the magazine list (NLRI, 2006).

It thus seems useful to create a vocabulary database and word lists based on a corpus which meet the four criteria shown below.

1) It is large enough.
2) It includes data by which texts can be classified into sub-genres to calculate dispersion.
3) It is recent.
4) It includes various types of texts to reflect the needs of the users such as academic

[30] Matsushita (2011) is the list created for this study. The ranking used for the comparison here is 'the *U* (usage coefficient) ranking for written Japanese including assumed known words' which will be described in the later sections in this chapter.

prose, literary works and internet language.

In current Japanese linguistic studies, the only corpus which meets the four criteria is the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (NINJAL, the National Institute for Japanese Language, 2009). In this research, a new vocabulary database entitled Vocabulary Database for Reading Japanese (VDRJ) is created from the BCCWJ 2009 monitor version.

3.2.2 Research questions

The main research questions (MRQs) are:

MRQs: In what order should learners of Japanese as a second language learn words and characters in order to be able to read Japanese? How will the order vary according to the purpose of learning?

The sub-research-questions (SRQs) in this chapter are as follows.

SRQ 1) How can a Japanese vocabulary database and word lists be created to identify target words for learners at different levels of proficiency?

SRQ 2) What index is the most appropriate among existing indices to rank words in the best order for learning vocabulary for reading a wide range of Japanese texts?

SRQ 3) Is the most appropriate word ranking criteria different depending on the target learners such as general learners or international students? If yes, what are the more suitable criteria for those different learner groups?

SRQ 4) Are the created word lists better for text coverage than existing ones?

3.3 Process and techniques for making a vocabulary database for reading Japanese

In this section, the steps and tools for making the Vocabulary Database for Reading

Japanese (VDRJ) specially created for this study are described in detail. As shown in 2.4, Nation & Webb's (2011) six 'steps involved in making a word list' (p. 135-144; Table 3-1) is the most comprehensive guideline for making word lists. The vocabulary database for this study, from which various word lists can be created, basically follows Nation and Webb's steps but consider how to apply it to Japanese where necessary. To summarize, Nation and Webb's steps are 1) research question or reason, 2) unit of counting, 3) corpus, 4) criteria for counting words and separate lists, 5) criteria for ordering words and 6) cross-checking the list. In this section, I start to describe the target users of the database and the word lists as is related to Step 1 the research question stated above, followed by describing Step 3 the corpus, in conjunction with the divisions of the sub-corpora, as the choice of corpus is related to the target learners. Then, the other steps are described by following the order of Nation and Webb's steps. Step 6 cross-checking the list will be discussed in 3.3.5 as well as in Chapters 4 and 8. For technical notes, see Appendices from 3-1 to 3-5.

Table 3-1 Nation and Webb's six 'steps involved in making a word list (Nation & Webb, 2011, p 135)

1	Decide on the research question the list will be used to answer, or the reason for making the list.
2	Decide on the unit of counting you will use – word type, lemma, word family. This decision should relate closely to your reason for making the list.
3	Choose or create a suitable corpus. The makeup of the corpus should reflect the needs of the people who will benefit from the use of the list. For example, if you are designing a list for very young learners, the corpus should include the typical uses of language that young learners would meet and use. The size of the corpus will also depend on the nature of the word list. Brysbaert & New (2009) present data suggesting that for high-frequency words a 1,000,000-token corpus is sufficient. For low-frequency words, a corpus over 30,000,000 tokens is needed.
4	Make decisions about what will be counted as words and what will be put into separate lists. For example, will proper nouns be a part of the list or will they be separated in the counting?
5	Decide on the criteria that will be used to order the words in the list. These could include range, frequency and dispersion, or some summative measure like the standard frequency index (Carroll, Davies and Richman, 1971).
6	Cross-check the resulting list on another corpus or against another list to see if there are any notable omissions or unusual inclusions or placements.

3.3.1 The target users of the database and the word lists

To begin with, the target users of the database and word lists need to be identified. The database is basically for researchers and teachers of Japanese. For the word lists, which are extracted from the database, "general" learners and international students in Japanese universities are mainly targeted for this research.

It is not easy to identify "general" learners of Japanese as a second language. Here they can only be simply defined as "non-specialist learners who have the most common features with all learners of Japanese". They can partly be academic, but they mainly learn Japanese for non-academic purposes.

3.3.2 The corpus set and the divisions of the sub-corpora

The texts in the corpus BCCWJ are sampled in a careful manner[31] so it can be regarded as a representative set of book texts and internet forum texts of contemporary Japanese. All the sampled texts are published during the period between 1986 and 2005. The corpus does not contain magazine texts and newspaper texts as they are not included in the 2009 monitor version[32]. It may be a weakness of the corpus set, while it can also be considered a strength at the same time in that it will contain less unstable current vocabulary.

The whole corpus set contains approximately 33 million running words made up of the book corpus containing approximately 28 million and the internet forum site (Yahoo Chiebukuro) corpus containing approximately 5 million. Half of the book texts are sampled from books published between 2000 and 2005, and the other half is sampled from library books published between 1986 and 2005 to be stored at libraries in the Tokyo area.

---

[31] For detailed sampling principle and method, see Maruyama (2009) and Kashiwano et al. (2009).

[32] The complete BCCWJ which includes magazine and newspaper texts was completed in October 2011.

As mentioned in the previous section, all the texts need to be divided into sub-corpora in order to calculate a dispersion index. There are two main methods to divide a corpus into sub-corpora when the texts for different genres are not equally-sized. One way is to divide the whole corpus into equally-sized texts regardless of the genre. The number of sub-corpora can be either small or large with this method, but the number of sub-corpora in each genre may not be equal. The other way is to classify it into sub-corpora based on its content even though the sizes of the sub-corpora are not equal. This method will be more sensitive to different lexical quality of each sub-corpus while the differences on statistical features of sub-corpora may be greater, and/or the statistical figures of the sub-corpora may have different sensitivities even though the standardized frequency (average frequency per unit) is applied to the analysis.

For this research, the latter way, the content-based division is adopted. As it is more important to detect the different lexical quality of different genres than the evenness of the statistical sensitivity. As a result, literary texts which make up more than 8 million tokens among the whole corpus of 33 million are merely counted as one sub-corpus, because the whole corpus was compiled by a strict sampling way to make a "balanced" corpus so that it reflects the fact that people seem to read more literary books than the other genres.

The next question is: what criteria should be used to classify the texts? Because one of the main target users of the word lists is international students, the texts are placed into sub-corpora on the basis of academic genre. There are two main references for the classification: 1) the classification for the applications for the Japanese national grant-in-aid[33], and 2) the classification for statistics of affiliations of international students[34]. Based on these, all the academic genres are classified into the four large academic domains of 1) Arts and Humanities, 2) Social Sciences, 3) Technological Natural Sciences and 4)

[33] For the current classification, see (JSPS, 2010a, 2010b).

[34] For the current classification, see JASSO (2010).

Biological Natural Sciences, and then each of the four domains are classified into seven academic fields that come to 28 academic fields in total[35]. This is applied to the classification of the book corpus but not to the internet-forum one (Table 3-2) [36].

Each text in the corpus has two types of codes in terms of the content: Nippon Decimal Classification (NDC) (Mori & Japan Library Association (revised edition), 1995) and C-code[37]. NDC is a book classification code adopted in most Japanese libraries. C-code is a target audience code given by the publisher. The last two digits of C-code almost correspond to the hundreds and tens digits of NDC. Taking advantage of these codes, I created a correspondence table between NDC/C-code and the 28 academic fields to classify all the texts in the book corpus (Table 3-3)[38].

Also, '3' at the thousands digit of C-code means that the book is written for experts so that the book text can be regarded as a technical text. Therefore, all the book texts in 28 academic genres are classified into technical texts and the other general texts[39] making up 56 sub-corpora.

---

[35] The number of sub-divisions for VDRJ (i.e. 4 large divisions and 7 sub-corpora in each of the four divisions, 28 sub-corpora in total) is the same as the corpus for Coxhead (2000); however, the sub-divisions of VDRJ corpus is different from Coxhead's one, In Coxhead's corpus, only one of the four sub-divisions is in science while VDRJ corpus has two science sub-divisions out of the four.

[36] There are some fields which are not easy to classify. In NDC, the book classification code adopted in most Japanese library, psychology is classified a part of "philosophy and thoughts" which generally thought to be a part of humanities while it is classified as a part of social science in the Japanese Grant-in-aid classification and as a part of natural science in many western countries. In addition, there are many books on fortune telling which are classified as psychology. I classify academic books on psychology into social science but books on fortune telling or similar into humanities. Similarly, I had some difficulties with classification in the field of education, information science, home science and so on. For details, see Table 3-2 and 3-3.

[37] For the current classification, see Maruyama (2009b).

[38] Classifying and merging the texts into sub-corpora took three months as done by the author alone.

[39] 7 and 8 at the thousands digit of C-code means reference books for primary and middle school students, that are somewhat academic; however they are classified as general as they do not seem 'technical' for adult learners. The number of these texts is only 6 among more than ten thousand texts.

Table 3-2 The Classification of Domains and Fields for VDRJ

Domain/Field	The Ten Domains	Code for the Ten Domains	The 28 Academic Field Code	Notes
Literary Works/Imaginative Texts	Literary works	LW	a6_G	All classified as general texts of a6
Humanities and Arts				
Languages and Linguistics	Languages, Linguistics and Philosophy	LP	a1	
Philosophy and Religion			a2	
History	History and Ethnology	HE	a3	
Ethnology			a4	
Fine Arts	Arts and Other Humanities	AH	a5	
Literature (non-Literary/non-imaginative texts)			a6_T	All classified as technical texts of a6
Other Humanities and Arts			a7	
Social Sciences				
Politics	Politics and Law	PL	s1	
Law			s2	
Economics	Economics and Commerce	EC	s3	
Commerce and Business			s4	
Sociology and Social Issues	Sociology, Education and Other Social Issues	SE	s5	Including welfare, labour, gender issues
Education			s6	Including pedagogy on each subject
Other Social Matters			s7	Including transportation, media, current issues
Technological Natural Sciences				
Mathematics	Science and Technology	ST	t1	
Physics			t2	
Astronomy, Earth and Planetary Science			t3	
Chemistry, Metal and Mine			t4	
Technology (Architecture, Civil Engineering)			t5	
Technology (Mechanics, Electricity, Marine Engineering)			t6	
Other Technological Natural Sciences			t7	manufacturing, library science, part of domestic science
Biological Natural Sciences				
Biology	Biology and Medicine	BM	b1	
Agriculture			b2	Including forestry, fishery, animal husbandry, veterinary
Pharmacy			b3	
Medicine			b4	
Dentistry			b5	
Nursing			b6	
Other Biological Natural Sciences			b7	environmentology, part of domestic science
Internet Q & A Forum (Yahoo Chiebukuro)		IF		

Table 3-3 The correspondence between NDC/C-code and the Domains/ Fields in VDRJ

NDC	Genres in Nippon Decimal Classification (NDC) Newly revised 9th edition	The Four Academic Domains (*1)	The Ten Domain Code (*2)	The 28 Academic Field Code (*3)
000	General works (000-090 except for 007)	+	+	+
007	General works (Information science)	Tech.	ST	t7
010	Libraries/Library and information science	Tech.	ST	t7
020	Books. Bibliography	Human.	AH	a7
030	General encyclopedias	+	+	+
040	General collected essays	+	+	+
050	General serial publications	+	+	+
060	General societies	+	+	+
070	Journalism. Newspapers	Social.	SE	s5
080	General collections	+	+	+
090	Rare books. Local collections.	+	+	+
100	Philosophy	Human.	LP	a2
110	Special treatises on philosophy	Human.	LP	a2
120	Oriental thought	Human.	LP	a2
130	Western philosophy	Human.	LP	a2
140	Psychology (except for 147 and 148)	Social.	SE	s7
147	Psychology (Parapsychology, pychicism)	Human.	LP	a2
148	Psychology (Physiognomy, divination)	Human.	LP	a2
150	Ethics. Morals	Human.	LP	a2
160	Religion	Human.	LP	a2
170	Shinto	Human.	LP	a2
180	Buddhism	Human.	LP	a2
190	Christianity	Human.	LP	a2
200	General history	Human.	HE	a3
210	General history of Japan	Human.	HE	a3
220	General history of Asia	Human.	HE	a3
230	General history of Europe	Human.	HE	a3
240	General history of Africa	Human.	HE	a3
250	General history of North America	Human.	HE	a3
260	General history of South America	Human.	HE	a3
270	General history of Oceania/General history of Polar Regions	Human.	HE	a3
280	General biography	Human.	HE	a3
290	General geography/Description travel	Human.	HE	a4
300	Social science	+	+	+
310	Political science	Social.	PL	s1
320	Law	Social.	PL	s2
330	Economics (except for 335 and 336)	Social.	EC	s3
335	Economics (Corporate management)	Social.	EC	s4
336	Economics (Business management)	Social.	EC	s4
340	Public finance	Social.	EC	s3
350	Statistics	Social.	EC	s3
360	Society	Social.	SE	s5
370	Education	Social.	SE	s6
380	Customs, folklore and ethnology	Human.	HE	a4
390	National defence. Military science	Social.	PL	s1
400	Natural science	+	+	+
410	Mathematics	Tech.	ST	t1
420	Physics	Tech.	ST	t2
430	Chemistry	Tech.	ST	t4
440	Astronomy. Space sciences	Tech.	ST	t3
450	Earth sciences	Tech.	ST	t3
460	Biology	Bio.	BM	b1
470	Botany	Bio.	BM	b1
480	Zoology	Bio.	BM	b1
490	Medical sciences (except for 492.9, 497, 498, 499)	Bio.	BM	b4
492.9	Medical sciences (Clinical medicine, diagnosis/treatment/nursing)	Bio.	BM	b6
497	Medical sciences (Dentistry)	Bio.	BM	b5
498	Medical sciences (Hygienics, public hygiene, preventive medicine)	Bio.	BM	b7
499	Medical sciences (Pharmacy)	Bio.	BM	b3
500	Technology. Engineering (except for 509)	+	+	+
509	Technology. Engineering (Industrial economy)	Social.	EC	s4
510	Construction. Civil engineering	Tech.	ST	t5
520	Architecture. Building	Tech.	ST	t5
530	Mechanical engineering	Tech.	ST	t6
540	Electrical engineering	Tech.	ST	t6
550	Maritime engineering. Weapons	Tech.	ST	t6
560	Metal and mining engineering	Tech.	ST	t4
570	Chemical technology	Tech.	ST	t4
580	Manufactures	Tech.	ST	t7
590	Domestic arts and sciences	Tech.	ST	t7
591	Domestic arts and sciences (Home economics and management)	Social.	EC	s4

Table 3-3 (Continued)

NDC	Genres in Nippon Decimal Classification (NDC) Newly revised 9th edition	The four academic domains (*1)	The ten domain code (*2)	The 28 academic field code (*3)
592	Domestic arts and sciences (Home technology)	Tech.	ST	t7
593	Domestic arts and sciences (Clothing, sewing)	Tech.	ST	t7
594	Domestic arts and sciences (Handicraft)	Tech.	ST	t7
595	Domestic arts and sciences (Hair dressing, cosmetics)	Bio.	BM	b7
596	Domestic arts and sciences (Food, cooking)	Bio.	BM	b7
597	Domestic arts and sciences (Housing, furnishing and supplies)	Tech.	ST	t7
598	Domestic arts and sciences (Home hygienics)	Bio.	BM	b7
599	Domestic arts and sciences (Child rearing)	Bio.	BM	b7
600	Industry and commerce	+	+	+
610	Agriculture (except for 611)	Bio.	BM	b2
611	Agriculture (Agricultural economics)	Social.	EC	s3
620	Horticulture (except for 621)	Bio.	BM	b2
621	Horticulture (Horticultural economics/administration/management)	Social.	EC	s3
630	Sericulture. Silk industry (except for 631)	Bio.	BM	b2
631	Sericulture. Silk industry (Sericultural economics/administration/management)	Social.	EC	s3
640	Animal husbandry (except for 641)	Bio.	BM	b2
641	Animal husbandry (Livestock economics/administration/management)	Social.	EC	s3
650	Forestry (except for 651)	Bio.	BM	b2
651	Forestry (Forestry economics/administration/management)	Social.	EC	s3
660	Fishing industry. Fisheries (except for 661)	Bio.	BM	b2
661	Fishing industry. Fisheries (Fishery economics/administration/management)	Social.	EC	s3
670	Commerce	Social.	EC	s4
680	Transportation services	Social.	SE	s7
690	Communication services	Social.	SE	s7
700	The arts. Fine arts	Human.	AH	a5
710	Sculpture. Plastic arts	Human.	AH	a5
720	Painting. Pictorial arts. Shodo	Human.	AH	a5
730	Engraving	Human.	AH	a5
740	Photography and photographs	Human.	AH	a5
750	Industrial arts	Human.	AH	a5
760	Music. Theatrical dancing	Human.	AH	a5
770	Theater. Motion pictures	Human.	AH	a5
780	Sports and physical training	Bio.	BM	b7
790	Accomplishments and amusements	Human.	AH	a5
800	Language	Human.	LP	a1
810	Nipponese	Human.	LP	a1
820	Chinese. Other Oriental languages	Human.	LP	a1
830	English	Human.	LP	a1
840	German	Human.	LP	a1
850	French	Human.	LP	a1
860	Spanish	Human.	LP	a1
870	Italian	Human.	LP	a1
880	Russian	Human.	LP	a1
890	Other languages	Human.	LP	a1
900	Literature	Human.	LW/AH+	a6/a7+
910	Nipponese literature	Human.	LW/AH+	a6/a7+
920	Chinese literature/Other Oriental literature	Human.	LW/AH+	a6/a7+
930	English and American literature	Human.	LW/AH+	a6/a7+
940	German literature	Human.	LW/AH+	a6/a7+
950	French literature	Human.	LW/AH+	a6/a7+
960	Spanish literature	Human.	LW/AH+	a6/a7+
970	Italian literature	Human.	LW/AH+	a6/a7+
980	Russian literature	Human.	LW/AH+	a6/a7+
990	Literatures of other languages	Human.	LW/AH+	a6/a7+

+ The C-code is also referred to decide on the domain/field.
Within the NDC range between 910-990, in principle, texts with the unit digit 1,2 or 3 of NDC go to a6 (literary works), the
If NDC and C-code do not agree on whether the text is on literature, the judgement depends on the content. Texts which
The texts classified as literature by NDC and the last two digits of C-code are 95 (review, essay, others) go to a7.
Texts on social issues or thoughts are mainly referred to C-code. The last two digits 30 go to s7, 36 go to s5, exceepting
Except for the case with +, the field is decided by C-code where NDC seems inappropriate (misclassification).

There seem to be no established criteria for deciding on the number of sub-corpora to calculate the dispersion; however, some of the 28 academic fields do not seem to have

enough number of tokens to get a reliable dispersion figure (See Table 3-4). On the other hand, the four academic domains also seem inappropriate for the base of dispersion measure as the number four is too small to calculate the dispersion.

Nation & Webb (2011) describe how Nation's word list (Nation, 2006)[40] had been developed based on the classification of British National Corpus composed of ten sub-sections with which he checked the range of each word to rank words. In light of this, I also tried to divide the whole corpus into the same or similar number of sub-corpora. There are two reasons for this decision. First, the purposes of this study are similar to Nation and Webb's ideas. Nation's list (Nation, 2006) serves for checking text coverage and developing Vocabulary Size Test (Beglar, 2010; Nation & Beglar, 2007). Likewise, this word lists are also designed for checking text coverage and developing vocabulary size test. Second, BCCWJ, the main corpus for this study, is designed in the light of the design of the British National Corpus.

To divide the whole corpus into ten, the literary work texts are extracted from the book corpus as one domain first, and then the remainder of the book corpus divided into eight with the consideration of combining close fields together and balancing the number of tokens. Adding the internet-forum site corpus as one domain, the ten domains for the dispersion measure were completed. The result is shown in Table 3-5.

3.3.3 Word segmentation and the unit of counting

As mentioned in 2.4.1, the unit of counting cannot help but be influenced or limited by the tools for word segmentation as there is no space between words in general Japanese orthography. To create the database and word lists, word segmentation must be done first by choosing an appropriate morphological analyser and a dictionary for the analyser.

[40] This set of lists can be downloaded from the "Resources" section of Nation's web-site http://www.victoria.ac.nz/lals/staff/paul-nation.aspx.

Table 3-4 Numbers of Types and Tokens by Field in VDRJ *The corpus is made from books and internet forum sites contained in NINJAL (2009).

Field	Code for the ten domains	G (General)		T (Technical)		Total	
		G Type	G Token	T Type	T Token	Type	Token
Literary Works/Imaginative Texts	LW	**68,446**	**8,251,999**	--	--	**68,446**	**8,251,999**
Humanities and Arts							
Languages and Linguistics	LP	21,252	403,305	7,831	102,504	23,708	505,809
Philosophy and Religion		36,253	1,503,013	9,269	125,917	38,229	1,628,930
History	HE	49,700	2,096,004	11,835	138,139	51,514	2,234,143
Ethnology		39,759	1,083,009	3,040	19,666	40,150	1,102,675
Fine Arts	AH	35,501	967,809	5,042	39,744	36,177	1,007,553
Literature (G=Literary works=Imaginative texts)		--	--	5,592	36,852	5,592	36,852
Other Humanities and Arts		46,304	1,973,098	683	3,414	46,337	1,976,512
The Whole of Humanities and Arts		**88,953**	**8,026,238**	**23,787**	**466,236**	**92,810**	**8,492,474**
Social Sciences							
Politics	PL	26,299	920,841	8,814	115,166	27,900	1,036,007
Law		16,502	511,059	10,074	333,946	19,542	845,005
Economics	EC	20,015	684,404	12,534	367,555	23,525	1,051,959
Commerce and Business		22,087	846,432	10,788	310,716	24,489	1,157,148
Sociology and Social Issues	SE	30,362	1,318,930	12,960	333,772	33,008	1,652,702
Education		20,157	621,050	10,417	262,063	22,675	883,113
Other Social Matters		18,993	424,164	4,114	36,168	19,652	460,332
The Whole of Social Sciences		**54,613**	**5,326,880**	**29,386**	**1,759,386**	**60,762**	**7,086,266**
Technological Natural Sciences							
Mathematics	ST	3,497	40,397	1,959	19,472	4,352	59,869
Physics		2,368	25,239	1,280	9,430	2,920	34,669
Astronomy, Earth and Planetary Science		8,181	101,565	2,583	21,765	9,035	123,330
Chemistry, Metal and Mine		4,682	37,469	2,553	23,275	6,017	60,744
Technology (Architecture, Civil Engineering)		16,242	307,617	7,662	114,099	18,443	421,716
Technology (Mechanics, Electricity, Marine Engineering)		12,993	195,762	5,495	72,049	14,820	267,811
Other Technological Natural Sciences		18,530	399,470	8,426	145,175	21,018	544,645
The Whole of Technological Natural Sciences		**32,125**	**1,107,519**	**15,864**	**405,265**	**36,309**	**1,512,784**
Biological Natural Science							
Biology	BM	14,680	262,283	4,064	41,071	15,672	303,354
Agriculture		14,932	238,989	3,376	28,584	15,860	267,573
Pharmacy		3,610	24,703	1,103	10,197	4,017	34,900
Medicine		16,657	485,896	5,955	82,800	17,961	568,696
Dentistry		1,740	11,551	874	3,814	2,174	15,365
Nursing		2,348	19,255	2,491	23,505	3,744	42,760
Other Biological Natural Sciences		28,254	943,822	6,749	74,567	29,490	1,018,389
The Whole of Biological Natural Science		**40,160**	**1,986,499**	**13,117**	**264,538**	**42,674**	**2,251,037**
Internet Q & A Forum (Yahoo Chiebukuro)	IF	**54,215**	**5,224,852**	--	--	**54,215**	**5,224,852**
The Whole of VDRJ		**135,794**	**29,923,987**	**46,996**	**2,895,425**	**144,231**	**32,819,412**

Note 1: Published books and library books are added together.
Note 2: The figures contain number of signs. Unidic and MeCab were used for word segmentation. No additional processing was made for extracting noises.
Note 3: If the C-code of a text is 3,000-3,999, it is counted as a technical text.

Table 3-5 Numbers and Ratios of Tokens by the Ten Domain Classification

Domain	Code for the Ten Domains	Number of Tokens	Proportion
Literary Works/Imaginative Texts	LW	8,251,999	25.1%
Languages, Linguistics and Philosophy	LP	2,134,739	6.5%
History and Ethnology	HE	3,336,818	10.2%
Arts and Other Humanities	AH	3,020,917	9.2%
Politics and Law	PL	1,881,012	5.7%
Economics and Commerce	EC	2,209,107	6.7%
Sociology, Education and Other Social Issues	SE	2,996,147	9.1%
Science and Technology	ST	1,512,784	4.6%
Biology and Medicine	BM	2,251,037	6.9%
Internet Q & A Forum	IF	5,224,852	15.9%
Total		32,819,412	100.0%

The combination of a morphological analyser and a dictionary adopted for this study is MeCab (Kudo, 2009a) and UniDic (Den et al., 2009). MeCab seems the newest and most accurate analyser. It still produces errors, but the error rate for recognizing lexemes with UniDic is approximately 1.5% which is 1.2% lower than Chasen (Kudo, 2009b, p 31) which was the representative analyser used for many previous studies. The error rate is the most important criterion for choosing the analyser. UniDic is primarily compiled for analysing BCCWJ. It is a very comprehensive dictionary which returns types of information such as orthographic form, phonological form, conjugation type, lexeme, part of speech, word-origin type and so on.

The unit of counting adopted for this study is what is called a 'lexeme' of the 'short unit' (短単位) defined by UniDic (Den et al., 2009). This is quite an inclusive unit. It is similar to the word family in English to some extent; however, there are some points which do not allow simple comparison with English.

To begin with, the 'short unit' is a similar unit to the morpheme but is allowed to combine with another morpheme only once in designated cases. (For the complete rules of the units, see Ogura, Koiso, Fujiike, & Hara (2009).) This unit must be close to the unit of processing meaning which meets the purpose of this study. One good point with this unit is that it is comparable with other studies as it is adopted for many studies since a similar unit called β unit is used in NLRI (1962), one of the most influential Japanese vocabulary frequency list in twentieth century.

Here we should note that an affix such as 学 ‘gaku’ (-logy) for 社会学 ‘shakai-gaku’ (sociology) or 室 ‘shitsu’ (room) for 会議室 ‘kaigi-shitsu’ (meeting room) is also the short unit to be counted as one unit for this study. But, 学 for 医学 ‘igaku’ (medical science) or 室 ‘shitsu’ for 教室 ‘kyoushitsu’(classroom) is not the unit, because 医 ‘i’ (medical) or 教 ‘kyou’ (teaching) is not a free morpheme while 社会 ‘shakai’ (society) and 会議 ‘kaigi’ (meeting) are free morphemes. Taking all the criteria into account, the unit of counting for this study is similar to the one for Nation’s list, where Level 6 in Bauer & Nation (1993) is adopted for affixed forms[41], except that more affixes are counted for this study[42]. As discussed in 2.4.1, Japanese affixes have more varieties to express more substantial meanings (e.g. ‘room’ in the above example) than English, these affixes should be a unit of counting for this study as the affixes require learning of the form and the meaning.

For some compound verbs, UniDic allows a combination of two verbs at most as compound verbs often derive different meanings from the original verbs. That is, verbs such as 受け入れる ‘uke-ireru’ (accept) which is the combination of 受ける ‘ukeru’ (receive) and 入れる ‘ireru’ (put into) can be counted as a unit.

The ‘lexeme’ for this study includes the following.

a) Conjugated forms of verbs and adjectives

e.g. 読む ‘yomu’ and 読み ‘yomi’ (read)

b) Phonologically changed forms

e.g. やはり ‘yahari’ and やっぱり ‘yappari’ (also, still, after all)

c) Some cognates with different orthographic forms

e.g. 足 ‘ashi’ and 脚 ‘ashi’ (foot, leg)

---

[41] The Level 6 of Bauer & Nation (1993) definition of affix includes all inflections and the most frequent, productive, and regular prefixes and suffixes (p. 255-261). The stems to which affixes are added must be able to stand as free forms (e.g., *administrator* and *administrative* cannot be members of the same word family because *administrate* is not a free form). See also 2.4.1.

[42] In this study, 753 affixes are identified while only 91 affixes are identified in English from Level 1 to 6 in Bauer & Nation (1993). See also Chapter 4.

The criterion c) seems problematic because a learner requires a lot of extra knowledge for different orthography. UniDic also returns 'orthographic form' where different forms are all counted separately; however, this unit does not seem appropriate for assessing written receptive knowledge because "The most sensible unit when counting for receptive knowledge is the word family…The idea behind using the word family as the unit of counting is that if one or two members of the family are known, then little learning is required for receptive use (comprehension) of other family members." (Nation & Webb, 2011, p 136). Therefore, I accept the compromise to use the lexeme as the unit of counting. When an item can be written in two or more forms, users are recommended to check the frequencies of different forms by a concordance[43] or the Kanji database made in Chapter 5.

3.3.4 Criteria for counting known words and making separate lists: The idea of "Assumed Known Words"

3.3.4.1 Forms excluded from the database

First of all, some forms such as signs for enumeration should be excluded from the database. Single phonographic characters (Hiragana, Katakana, alphabet and other foreign characters) judged as signs by the tools of MeCab and UniDic are excluded from the list[44]. Some of them are incorrectly analysed as a lexeme by the analyser. In this case, excluding single characters seems appropriate so that the frequency count will be less distorted. Most signs are not counted as a word with the software tools[45]; however some signs not automatically excluded by the software must be excluded manually. Signs which have a specific meaning (e.g., (株) for 'Inc.' or 'Co. Ltd.', 々 for repeating the previous character)

[43] The vocabulary database will include different orthographic forms in magazine texts with the frequency of each form cited from NLRI (2006).

[44] Non-sign single characters such as particles が or は are of course included in the database.

[45] AntConc (Anthony, 2007) and AntWordProfiler (Anthony, 2009) are the main tools for creating the database and the word lists.

are included in the list. For specific signs excluded from the database, see Appendix 3-2.

3.3.4.2 Assumed Known Words

After excluding signs, "Assumed Known Words" must be identified. This is one of the key concepts for this study as it directly relates to the learning burden. As mentioned in 2.4.2, Nation & Webb (2011) claim that decisions whether a form is counted as a known word should depend on "the learning burden principle". That is, words such as proper names should not be a headword in the frequency list as they require little previous knowledge to be understood (p.137-138). Based on this idea, Nation created separate lists whose words are counted as known words when measuring text coverage. The separate lists are for transparent compounds, proper names and non-words[46] such as ah, hmm or eh[47]. This study also follows this idea and creates separate lists for Assumed Known Words; however, there are a few problems to consider when applying this idea to Japanese.

In this study, three separate lists for Assumed Known Words are created: 1) Proper nouns, 2) Hesitations or fillers and 3) Miscellaneous. The words in these lists are assumed known words so that they are counted as known words when measuring the coverage of text. Transparent compounds are not identified except for numerals.

3.3.4.2.1 Proper nouns

From the viewpoint of statistical analysis, proper nouns can be the most substantial issue. In English, most proper nouns are easy to identify as their initial letter is capitalized, while there is no such rule for Japanese proper nouns. Nevertheless, these words seem easy to be identified from other types of contextual clues such as という 'toiu' i.e. リクルートという会社 'Rikuru^to toiu kaisha' (a company called Recruit). Thus, most proper nouns

[46] He also calls the items "hesitations etc." on a different page (Nation & Webb, 2011, p 141).

[47] He also considers foreign words and abbreviations which are included in the general list.

are counted as known words for this study.

Some high-frequency proper nouns are put into the general list but not into the proper noun list as Assumed Known Words since they require previous learning to understand their meaning. Nation (2011) uses the words *London*, *Paris*, *Rome* as examples. These words are 'assumed to require more background information on the part of the reader than other proper names' (p.139). The word 東京 'Toukyou' (Tokyo) is generally expected to include knowing that it is the capital of Japan without any explanation. Then, the question is: What proper nouns are shared with the background information by the majority of users of Japanese? Checking frequency lists, high-frequency proper nouns seem mostly taught in primary schools in Japan, or names with the current issues used in the media. In this study, for country names and prefecture names, the cut-off point was set at 7.0 occurrences per million tokens. The words with 7.0 or more occurrences per million tokens are put into the general list of VDRJ[48] as most of them seem known to the majority of users of Japanese. Aware of these criteria, other place names and historic persons' names are classified with some adjustment. Commonly used family and given names are mostly put into the proper noun list. Some names which can be either a place name or person's name such as 川口 'Kawaguchi', 上野 'Ueno' or 美保 'Miho' are also put into the proper noun list even if each of them has 7 or more occurrences per million. Some examples of the lowest-frequency words in the general list and the highest-frequency words in the proper noun list are in Table 3-6 and 3-7. For more detailed criteria for choosing the proper nouns to be put in the proper noun list, see Appendix 3-5.

3.3.4.2.2 Hesitations or fillers

Hesitations or fillers such as えー 'e^', うー 'u^' are separately put into the hesitations list. Though fillers have a certain function in the interaction, they seem

[48] Some single or compound abbreviated words of high-frequency proper nouns such as 伊 'i' (Italy) or 北米 'hokubei' (North America) are also put into the general list even if it only has less than 7 occurrences.

understandable without previous knowledge. Only 10 words are listed. The fillers provide very little coverage in the written text. See also Appendix 3-5.

Table 3-6 Ten examples of the lowest-frequency proper nouns in the general list

Written form	Reading	Meaning	Token per Million
イングランド	Ingurando	England	7.47
和歌山	Wakayama	Wakayama (Prefecture)	7.44
元禄	Genroku	Genroku (period in Edo era)	7.44
マッカーサー	Makka:sa:	MacArthur	7.35
スウェーデン	Suwe:den	Sweden	7.20
アルゼンチン	Aruzenchin	Argentina	7.10
国鉄	Kokutetsu	Japanese National Railways (company)	7.07
パレスチナ	Paresuchina	Palestine	7.07
ナポレオン	Naporeon	Napoleon	7.04
シンガポール	Shingapo:ru	Singapore	7.01

Table 3-7 Ten examples of the highest-frequency proper nouns in the Assumed Known Word list *

Written form	Reading	Meaning	Token per Million
日本橋	Nihonbashi/Nipponbashi	a bridge in Tokyo/Osaka	6.98
東海道	Toukaidou	a highway from Tokyo to Osaka	6.83
屋久	Yaku	Yaku Island	6.80
山梨	Yamanashi	Yamanashi (Prefecture)	6.77
広東	Kanton	Guangdong / Canton (Province in China)	6.74
大津	Ootsu	a city name (in Shiga Prefecture)	6.71
スコットランド	Sukottorando	Scotland	6.68
ソビエト	Sobieto	Soviet (Union)	6.64
釈迦	Shaka	Shakyamuni (the Budda)	6.64
ＥＵ	i:yu:	European Union	6.52

*Assumed Known Words means the words which do not require previous knowledge to understand.

3.3.4.2.3 Miscellaneous words

A list for ‘miscellaneous words’ was also created. This list is mainly for wrongly

analysed forms which do not make sense. Many of them are a part of proper nouns or expletives[49]. 346 items are identified. Half of them are one-timers so that they provide very little coverage of text. The reason why these items are included in the database is that they seem to be the counterparts of the excluded single character items. To count the tokens, these items should also be included in the database separately from the general list.

3.3.4.2.4 Transparent compounds and numerals

Transparent compounds can theoretically be assumed known but not identified for this study except for the numerals, because making a transparent compound list does not seem a practical idea. Japanese has lots of Kanji compounds (see 2.3 in Chapter 2). The majority of them are made up of two Kanji. Many Kanji are considerably productive in forming words as there are only about two thousand commonly used Kanji which produce tens of thousands of Kanji compounds. What is more complicated, each component of those Kanji compounds cannot always be regarded as a morpheme, let alone a word. Most Kanji have the basic meaning which is sometimes quite abstract and generates various meanings according to the combination with the other components. Also, many Kanji have two or more phonological forms even if they keep the same meaning (see 2.3 in Chapter 2), which leads to the difficulty in identifying a morpheme[50]. There is another practical reason for the decision namely that it would be difficult to compare the results with other studies if transparent compounds are separated as known words because no other Japanese studies followed that procedure. Alternatively, this study investigates how many characters cover how many words in Chapter 6.

Only for the numerals, transparent compounds are identified. These are not put in a

[49] Most wrongly analysed single character items are excluded from the database as mentioned above.

[50] Morioka (1984) proposes the concept of "Kanji morpheme" (p.168-170). It was not totally established in Japanese linguistics; however, some of his ideas are widely acknowledged in the field.

separate list but included as family members under the least frequent part of the compound. For example, 二十 “nijuu” (twenty) goes under 十 “juu” (ten) as it is less frequent than 二 “ni”. This decision is somewhat arbitrary but practical. Numerals are high-frequency words which affect the results of counting.

Before identifying the transparent numerals, this study used 簡略モード (simple mode) of the software NumTrans (Yamada, 2008) for the segmentation of numerals[51]. This choice is also important as the way for counting numerals since there are several ways to express numbers in Japanese. The simple mode is the most common way among the three choices of detailed mode, simple mode and no transformation.

3.3.4.3 Words not assumed known

3.3.4.3.1 Foreign words and abbreviations

Nation and Webb (2011) also consider foreign words (e.g. précis in English) and abbreviations (p.139-140). In this study, based on their idea, foreign words (e.g. “European Union” in a Japanese text) and abbreviations (e.g. “EU” in a Japanese text) are not separated but included in the general list because knowing the word ヨーロッパ連合 “Yo^roppa Rengou” (European Union) or even ヨーロピアン・ユニオン “Yo^ropian Yunion” does not mean knowing the words “European Union” or “EU” as they have different forms which need to be learned.

3.3.4.3.2 Homonyms, homographs and other form-related words

Homonyms and homographs are basically classified according to MeCab and UniDic’s judgements but are manually checked and corrected as far as possible where necessary. In particular, within the top 20,000 words, if a word was thought to have two or more completely different meanings, the usage of the word was scrutinized using a

[51] The software is mounted on the user interface software 茶まめ Chamame (Ogiso, 2009). For more detailed rules for the number segmentation, see Yamada & Koiso (2008).

concordance, and its frequency figure was corrected where the rank of the word was largely influenced[52].

In Japanese, there are many homonyms in loanwords from Western languages such as コート 'ko^to' (coat/court) and ドラッグ 'doraggu' (drag/drug) and some Japanese-origin or Chinese-origin ones such as 私 'watakushi' (I/private) and 大 'dai' (university (affix) / large-size). Some of the loanword-type homonyms such as リング 'ringu' (ring for circle/boxing) and マッチ 'matchi' (match for fire/game) are originally homonyms in English while the majority of homonyms are phonologically unified when they become Japanese which has less phonemes. For any type of homonym, UniDic tries to add a tag to distinguish them in meaning, but unfortunately it often fails. Thus, it requires manual checking.

Japanese also has many words which share the orthography but have different meanings. These words are often called homographs. In Japanese, however, many of this type of words are a pair of Kun-reading and On-reading with a Kanji or Kanji compound (e.g. 金 'kane/kin' (money/gold)) so that they should be called cognates. The Kun-reading word borrowed the orthography from Chinese so that the pair has a historical relationship. In addition, in psychology, words with related meanings are generally called cognates but not homographs. Since the words are from different origins but tied with each other through orthography, here I name them 'written cognates'. Both 'kane' and 'kin' are Japanese words while 'kane' and the Chinese word 金 /jin1/ are phonologically originating in different languages. The former can be called 'intralingual-written cognates' while the latter can be called 'interlingual-written cognates'. The On-reading word 'kin' and the word /jin1/ are general cognates as they share the same phonological origin.

Only cases such as the words 'kome' (rice) and 'bei' (America) both of which share the same Kanji 米 but have no semantic relationship between them, the pair of words can

[52] This is an exhausting job. This checking and correction alone took two to three months, but it is never completed. If the word segmentation and tagging had no errors, this job would be much easier.

be categorized as homographs. The Kanji 米 was given to represent the meaning *America* as it is a part of the word 亜米利加 where only pronunciation was borrowed from the Kanji to represent the sound *America*. In other words, the word 'bei' has no semantic relationship with 'kome'.

The categories of these form-related words are shown in Table 3-8 and 3-9. 'Partial-cognate compound' in Table 3-9 means a compound whose components are originating in Chinese but the word does not exist in Chinese. 'Interlingual-written cognate' means a word which is used in Chinese in the same or similar orthography while the pronunciation is Kun-reading, Japanese-origin pronunciation. If every component of a word shares the meaning and orthography but not phonology with the original Chinese character, the word can be called an 'Interlingual-written-partial-cognate compound'.

Table 3-8 Categories for Intralingual Form-related Japanese Words

Category	Phonological form	Orthographical form	Meaning	Examples
Homonym	same	same	different	"ko:to" コート (coat/court) "doraggu" ドラッグ (drag/drug)
Homophone	same	different	different	"kawa" 川/皮 (river/leather) "tou" 塔/十 (tower/ten)
Intralingual-written cognate	different	same	same/related	"kuda/kan" 管 (tube)、 "moto/hon" 本 (basis/book *1)
Homograph	different	same	different	"kome/bei" 米 (rice/America)

*The word "hon" usually means a 'book' which was derived from 'basis' historically.
It can also be a component of a Kanji compound which means 'basis' as is in 基本 "kihon" (basics).

Table 3-9 Categories for Interlingual Form-related Words between Chinese and Japanese

Category	Phonological form	Orthographical form	Meaning	Examples (Japanese/Chinese)
Cognate	related	same/similar	same/related	"gakushuu"/"xue2 xi2" 学習/学习 (learning) "goudou"/"he2 tong2" 合同 (combined/contract)
Partial-cognate compound	related with each component	same/similar for each component	related/ different	"taisetsu"/"da4-qie4" 大切/大−切 (important/big-cut)
Interlingual-written cognate	different	same/similar	same/related/different	"kuda"/"guan3" 管 (tube) "baai"/"chang3 he2" 場合/场合 (case) "ugoku"/"dong4" 動く/动 (move)
Interlingual-written-partial-cognate compound	different	same/similar for each component	related/ different	"tokei"/"shi2-ji4" 時計/时−计 (clock/time-measure)

Corpus software is good at dealing with forms but not at meanings generally. The distinction for form-related pairs of words shown in Table 3-8 basically follows UniDic, but in some cases still needs manual correction. Most of these words are included in the general list anyway. (The distinction shown in Table 3-9 is not directly concerned with making the vocabulary database described in this chapter; however, it is related to discussions on word origins in Chapters 4 and 7.)

3.3.4.4 Remaining issues with cognates and loanwords

The idea of Assumed Known Words is also important in terms of understanding cognates or loanwords[53]. As mentioned in the previous chapters, more than half of the Japanese vocabulary is cognates or loanwords. For adult Chinese learners, many of the written Kanji words require little previous learning of Japanese to be understood. The same approach can be applied to loanwords from English (Gairaigo) for English speaking learners (Daulton, 2004). This advantage (or disadvantage) is not for all learners; however, considering the fact that there are notably high proportion of Chinese-origin and English-origin words in Japanese, and Chinese and English background learners of Japanese, it will be useful for measuring actual learning burden to identify the words which share the same basic meaning and form between Japanese and learners' languages. This issue is to be discussed in 4.5 and 7.4.5 in Chapters 4 and 7.

3.3.5 Criteria for ordering words (1): Index

The sub-research-questions here is: SRQ 2) What index is most appropriate among existing indices to rank words in the best order for learning vocabulary for reading a wide range of Japanese texts?

[53] Cognates share a common etymological origin. Loanwords are words directly borrowed from a language, and the use is basically not changed.

The most general criterion for ranking words is frequency. It is due to the idea that the most important words are the words which learners encounter most frequently in their lives. As discussed in 2.4.2, however, dispersion is also an important criterion, because some high-frequency words only occur in limited domains which are not very relevant to some learners. Given this, basic words should be the words which have a high-frequency in a wide range of domains. Generally, learners will be benefited by learning this type of words first. There are some mathematical indices for ranking words; however, these typically involve a kind of adjusted frequency calculated by some combination of the total frequency in a large corpus and the dispersion calculated based on the sub-frequencies of the sub-corpora made by dividing the whole corpus[54].

Among a few adjusted frequency measures, as discussed in 2.4.2, Juilland's *U* (usage coefficient) (Juilland & Chang-Rodrigues, 1964) and Carroll's Standard Frequency Index (SFI) (Carroll, 1970) are possible adjusted frequency measures for this study. In addition, Gries *DP*, as an alternative to Juilland's *D*, can also be applied to the formula of Juilland's *U*. In sum, the three indices shown below are to be tested in this section.

1) Juilland's U[55]: $U = F \times D$
2) Alternative *U* (U_{DP}) by applying Gries's *DP* as dispersion measure:

$$U_{DP} = F \times (1\text{-}DP)$$

3) Carroll's SFI: $\text{SFI} = 10 \times (\log_{10} U_m + 4)$

F: the frequency of a given word in the whole corpus

[54] Ordering words by Range as the first criterion is also a possible method (Nation, 2006; Vander Beke, 1932); however, as discussed in 2.4.3.2., it is not suitable for this study since it requires equally-sized sub-corpora, and it penalise sub-sections with zero frequency too much.

[55] For users' convenience, as Leech, Rayson, & Wilson (2001) do, dispersion figures will be shown after multiplying by 100 in the complete database.

$D = \left(1 - \frac{V}{\sqrt{n-1}}\right)$

V (variation coefficient) = $\frac{\sigma}{\bar{f}}$

σ: Standard deviation of sub-frequencies

$\bar{f}$: the mean sub-frequencies $\bar{f} = \frac{F}{n}$

n: Number of sub-corpora

$DP = (\sum_1^n |Po - Pe|)/2$

Po (observed percentage) = $\frac{f_j}{F}$

f_j: Frequency of a given word in sub-corpus j

Pe (expected percentage) = $\frac{s_j}{N}$

s_j : Total number of words in sub-corpus j

N: Total number of words in the whole corpus

When computing *UG*, *F* is multiplied by (1-*DP*), because the value of *DP*, opposed to Juilland'*D*, will be 0 when a word is totally evenly distributed in each sub-corpus.

$U_m = (1{,}000{,}000/N)[FD_2 + (1 - D_2)f_{min}]$

$D_2 = H/\log n$

$H = \log P - \left(\sum_j p_j \log p_j\right)/P$ ($p_j \log p_j = 0$ for $p_j = 0$)

$P = \sum_j P_j \quad P_j = \frac{f_j}{s_j}$

$f_{min} = \left(\sum s_j f_j\right)/N$

The rankings of words by these indices are compared as follows to decide on the index to order words for this study.

3.3.5.1 Method[56]

For all the words (lexemes) excluding assumed known words, *U*, U_{DP} and SFI are calculated, and then ranking is given to each word by the indices. The number of lexemes is 111,285 excluding 30,700 assumed known words; however, there are tens of thousands of low-frequency words which have little practical importance but would influence statistical analysis. Therefore, after excluding 'words which occur only once' (one-timers) in the whole corpus, different ranges of words such as the most frequent sixty thousand or twenty thousand words should be tested by statistical analysis.

Specifically, the following four-step procedure was conducted.

1) Spearman's rank correlation coefficients (Spearman's *Rho*) were computed between *D*, *DP*, D_2 and the adjusted frequency indices (Table 3-10 to 3-12)[57]. Correlation coefficients were computed not only for the sixty thousand and twenty thousand words, but also for words ranked from 5,001 to 20,000 by *F* (total frequency). This was to avoid influences from some extreme frequency figures in the high-frequency range. To see the nature of the indices, Spearman's rank correlation coefficient is vital as the purpose of this study is to seek the best order to learn the Japanese vocabulary.

2) The number of words which have a gap in ranking by 1,000 or more between the indices was counted (Table 3-13). This will explain which index will be more sensible to skewness (a measure of the asymmetry of the distribution), kurtosis (a measure of flatness of the distribution), or uneven distribution. Again, ranking is rather more important than the index figure itself because the ranking shows the proposed order of learning. The base word lists for checking the text coverage will be created by k, i.e. 1,000 words so that the ranking gap less than 1,000 will have less importance.

[56] All the analyses in 3.3.5 were done before the wrongly-segmented items are corrected.

[57] Pearson's correlation coefficient cannot be applied to the indices as they do not follow the normal distribution. (Kolmogorov-Smirnov test for normality, $p < .001$ for all the indices.)

3) Among the most frequent 20,000 words, the most frequent ten words were listed from each of the word groups which consist of words with a gap in ranking by 1,000 or more between *U*-U_{DP}/*U*-SFI/U_{DP}-SFI (from Table 3-14 to 3-20). The average of the sub-frequencies of each word was computed, and then the indices, the average sub-frequency and sub-frequency rankings were compared between the words. For better comparison, words close to the rankings of 3,000/ 6,000/ 9,000/12,000/15,000/18,000 were added to the analysis as benchmarks.

4) Skewness (a measure of the asymmetry of the distribution, absolute value is used for the analysis here) and kurtosis (a measure of flatness of the distribution) for the most frequent 20180 words (with 48 occurrences or more in the whole corpus) were computed, and then Spearman's rank correlation coefficients between skewness/kurtosis and the other indices were computed for the words with *Range* ten, eight, six, four and two. By doing so, it is expected to examine which index is more sensitive to skewness and kurtosis.

3.3.5.2 Results and Discussion

1) Correlation coefficients (Spearman's *Rho*) between dispersion and adjusted frequency indices are shown from Table 3-10 to 3-12.

Table 3-10 Correlations (Spearman's *Rho*) between Dispersion and Adjusted Frequency Indices for the Words excluding One-timers in VDRJ N=61,056

	D	*DP*	D_2	*U*	U_{DP}	SFI
D	1	.923***	.986***	.826***	.774***	.787***
DP	.923***	1	.938***	.823***	.822***	.803***
D_2	.986***	.938***	1	.887***	.846***	.856***
U	.826***	.823***	.887***	1	.982***	.992***
U_{DP}	.774***	.822***	.846***	.982***	1	.995***
SFI	.787***	.803***	.856***	.992***	.995***	1

***. Correlation is significant at the 0.001 level (2-tailed).

Table 3-11 Correlations (Spearman's *Rho*) between Dispersion and Adjusted Frequency Indices for the Most Frequent 20000 Words in VDRJ N=20,000

	D	*DP*	D_2	*U*	U_{DP}	SFI
D	1	.911***	.986***	.540***	.510***	.479***
DP	.911***	1	.920***	.496***	.501***	.444***
D_2	.986***	.920***	1	.593***	.568***	.538***
U	.540***	.496***	.593***	1	.991***	.994***
U_{DP}	.510***	.501***	.568***	.991***	1	.994***
SFI	.479***	.444***	.538***	.994***	.994***	1

***. Correlation is significant at the 0.001 level (2-tailed).

Table 3-12 Correlations (Spearman's *Rho*) between Dispersion and Adjusted Frequency Indices for the Words with the Frequency Ranking from 5,001 to 20,000 in VDRJ N=15,000

	D	*DP*	D_2	*U*	U_{DP}	SFI
D	1	.905***	.984***	.521***	.468***	.423***
DP	.905***	1	.915***	.482***	.491***	.402***
D_2	.984***	.915***	1	.565***	.522***	.478***
U	.521***	.482***	.565***	1	.981***	.989***
U_{DP}	.468***	.491***	.522***	.981***	1	.988***
SFI	.423***	.402***	.478***	.989***	.988***	1

***. Correlation is significant at the 0.001 level (2-tailed).

As shown from Table 3-10 to 3-12, for the dispersion measure, D_2 performs similarly to *D* on this data. This result agrees with Gries (2010). *DP* is slightly different from the other two indices; however, adjusted frequencies (usage coefficients) are remarkably highly correlated with each other. This result is also consistent with Gries (2010).

Among the three tested ranges of words, the widest range which includes the top sixty thousand words returned the highest correlation coefficients, the top twenty thousand words returned the second highest, and the 15,000 words excluding the top 5,000 words returned the lowest among the three for Spearman's *Rho* (Table 3-10, 11 and 12). This means that, between the indices, there is no great difference in adjusted frequencies and rankings in the low-frequency range over the 20,000 word level as well as within the top

5,000 words, while the range between the 5,000 and 20,000 word levels will have more differences between the indices.

However, even for the words ranked from 5,001 to 20,000, the adjusted frequencies (usage coefficients) *U*, U_{DP} and SFI still highly correlate with each other at .98 or higher. These results mean, at least for this set of data, there seems no significant difference between the indices overall.

2) Therefore, the main concern is now for the words which are considerably differently ranked by different indices. The question here is: Which index is the most appropriate for ranking the words which have considerably great gaps in rankings by different indices? The number of words which have a gap in ranking by 1,000 or more between the indices is shown in Table 3-13. Example words are shown in Tables 3-15 to 3-20.

Table 3-13 Number of Words with the Ranking Gap of 1,000 or More between Adjusted Frequency Indices in the Most Frequent 20,000 Words

Ranking Gap (*)	*U*-U_{DP} (%)	*U*-*SFI* (%)	U_{DP}-*SFI* (%)
+1,000 or more	2,083 (10.4)	2,086 (10.4)	1,817 (9.1)
-1,000 or less	1,430 (7.2)	51 (0.3)	1,020 (5.1)

* Greater number in ranking here means lower ranking, i.e., '*U*-*UDP* = +1,000' means the ranking of *U* is lower than that of

Table 3-13 shows that *U* tends to give lower rankings to more words than the other two indices but to give higher rankings to fewer words. This tendency is particularly striking when *U* is compared with SFI. Only 51 words have a higher *U* ranking than SFI while 2,086 words have a lower *U* ranking than SFI. U_{DP} is in-between. It gives lower rankings by 1,000 to fewer words (1,430 words) than *U* (2,083 words) while to more words (1,817 words) than SFI (1,020 words). These results mean that *U* will be most sensitive to skewness and kurtosis. In other words, *U* tends to give lower rankings to unevenly distributed words. SFI tends to provide higher rankings to unevenly distributed words,

probably because, as Lyne (1985) indicates, D_2 generally provides higher figures than *D*. In other words, compared to other measures, SFI weights less with dispersion but more with the total frequency. U_{DP} does not tend to penalize unevenly distributed words so much as *U*; however, there are considerably high proportions (7.2 % against *U* and 9.1% against SFI) of words which have higher rankings by U_{DP} than by *U* or SFI, therefore, it is necessary to further examine which words are penalized or not penalized by these indices in the following step.

3) Table 3-14 shows the rankings by the indices and the sub-frequencies for the bench mark words. Tables from 3-15 to 3-20 show the rankings by the indices and the sub-frequencies for the most frequent ten words from each of the word groups which consist of words with a gap in ranking by 1,000 or more between *U*-U_{DP}/*U*-SFI/U_{DP}-SFI. Table 3-14 is for the benchmark words.

For the Tables 3-14 to 3-20, the codes for the ten sub-sections are as follows (See also Table 3-4). LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.

Table 3-14 Rankings of the Benchmark Words as Reference to the Comparison with the Words from Table 3-15 to Table 3-20

Lexeme in Kanji & Romanization		English Translation	*F* Ranking	*U* Ranking	U_{DP} Ranking	*SFI* Ranking	Ave. Freq. Rank. in 10 Sub Corpora	LW Freq. Ranking	LP Freq. Ranking	HE Freq. Ranking	AH Freq. Ranking	PL Freq. Ranking	EC Freq. Ranking	SE Freq. Ranking	ST Freq. Ranking	BM Freq. Ranking	IF Freq. Ranking
人格	jinkaku	character, personality	2,995	2,912	3,010	2,911	3,435	4,770	1,127	4,246	3,668	2,165	4,227	1,329	5,212	2,958	4,647
残酷	zankoku	cruel	5,991	5,464	5,390	5,580	7,143	4,864	7,531	3,663	3,468	8,926	12,194	4,776	7,726	10,384	7,898
破滅	hametsu	ruin	8,979	7,817	7,736	8,211	9,106	8,397	6,868	8,139	4,540	5,540	7,203	9,652	12,457	14,082	14,178
航行	koukou	navigation	11,969	10,997	11,472	11,208	14,283	8,397	14,826	6,468	19,965	11,493	6,649	13,162	11,182	20,616	30,068
論調	ronchou	tone of argument	14,993	13,353	14,272	13,812	14,178	24,024	16,911	19,633	12,950	5,965	7,533	7,106	14,198	14,082	19,380
現況	genkyou	present condition	17,866	16,172	17,648	16,571	17,066	32,207	25,325	19,633	22,670	6,211	7,533	10,182	11,182	14,082	21,630

The lexemes here are selected based on the following criteria. 1) Noun not meaning concrete things. 2) Orthographically stable (generally written in the fixed combination of Kanji. 3) Dispersion (*D*) figure is between 70 and 80.

For relatively evenly distributed words (.70 < *D* <.80) such as the words in Table 3-14, there are no great gaps in rankings between indices, and the rankings have no great gaps from the average sub-frequency ranking as well. Nevertheless, as shown in the Tables 3-15 to 3-20, there are great ranking gaps between the indices for unevenly distributed words.

Table 3-15 Ranking Comparison of the Most Frequent 10 Words with *U* Ranking Lower than U_{DP} Ranking by 1,000 or More

Lexeme in Kanji & Romanization		English Translation	*F* Ranking	*U* Ranking	U_{DP} Ranking	*SFI* Ranking	Ave. Freq. Rank. in 10 Sub Corpora	LW Freq. Ranking	LP Freq. Ranking	HE Freq. Ranking	AH Freq. Ranking	PL Freq. Ranking	EC Freq. Ranking	SE Freq. Ranking	ST Freq. Ranking	BM Freq. Ranking	IF Freq. Ranking
出品	shuppin	display, exhibition	713	7,125	2,624	2,538	14,393	21,623	38,534	9,273	4,630	11,493	15,875	12,235	12,457	17,688	124
落札	rakusatsu	successful bid	788	11,857	3,011	3,122	20,989	32,207	25,325	24,622	22,670	6,454	10,193	20,647	30,863	36,771	141
ヤフー	yafu:	Yahoo	1,080	14,575	3,728	3,904	26,465	54,591	38,534	19,633	46,296	28,382	13,695	17,819	8,746	36,771	182
図書	tosho	books, publications	1,592	3,056	2,007	2,328	2,723	2,477	2,209	2,482	1,338	3,680	4,074	2,444	157	4,475	3,898
オークション	o:kushon	auction	1,651	11,720	4,587	4,553	20,826	11,118	25,325	50,344	12,155	28,382	15,875	17,819	10,195	36,771	274
入札	nyuusatsu	bidding	1,793	5,573	4,124	3,359	10,889	15,214	20,059	14,500	17,977	3,602	5,576	8,823	8,746	14,082	311
預金	yokin	money on deposit	2,180	4,847	3,365	3,628	8,696	7,872	10,165	2,100	7,461	3,533	286	6,468	21,003	25,415	2,659
顧客	kokyaku	costomer, client	2,214	4,828	3,825	3,608	6,661	8,397	3,446	12,922	10,841	2,772	259	5,629	2,406	15,590	4,347
彼氏	kareshi	boy friend	2,268	8,389	4,949	4,547	15,122	8,257	11,954	24,622	12,950	28,382	15,875	6,868	30,863	11,048	397
ID	aidi:	ID	2,445	7,182	5,514	4,416	15,272	21,623	38,534	19,633	26,415	6,742	7,932	14,324	5,212	11,877	429

Table 3-16 Ranking Comparison of the Most Frequent 10 Words with *U* Ranking Lower than SFI Ranking by 1,000 or More

Lexeme in Kanji & Romanization		English Translation	*F* Ranking	*U* Ranking	U_{DP} Ranking	*SFI* Ranking	Ave. Freq. Rank. in 10 Sub Corpora	LW Freq. Ranking	LP Freq. Ranking	HE Freq. Ranking	AH Freq. Ranking	PL Freq. Ranking	EC Freq. Ranking	SE Freq. Ranking	ST Freq. Ranking	BM Freq. Ranking	IF Freq. Ranking
出品	shuppin	display, exhibition	713	7,125	2,624	2,538	14,393	21,623	38,534	9,273	4,630	11,493	15,875	12,235	12,457	17,688	124
落札	rakusatsu	successful bid	788	11,857	3,011	3,122	20,989	32,207	25,325	24,622	22,670	6,454	10,193	20,647	30,863	36,771	141
ヤフー	yafu:	Yahoo	1,080	14,575	3,728	3,904	26,465	54,591	38,534	19,633	46,296	28,382	13,695	17,819	8,746	36,771	182
オークション	o:kushon	auction	1,651	11,720	4,587	4,553	20,826	11,118	25,325	50,344	12,155	28,382	15,875	17,819	10,195	36,771	274
入札	nyuusatsu	bidding	1,793	5,573	4,124	3,359	10,889	15,214	20,059	14,500	17,977	3,602	5,576	8,823	8,746	14,082	311
預金	yokin	money on deposit	2,180	4,847	3,365	3,628	8,696	7,872	10,165	2,100	7,461	3,533	286	6,468	21,003	25,415	2,659
顧客	kokyaku	costomer, client	2,214	4,828	3,825	3,608	6,661	8,397	3,446	12,922	10,841	2,772	259	5,629	2,406	15,590	4,347
彼氏	kareshi	boy friend	2,268	8,389	4,949	4,547	15,122	8,257	11,954	24,622	12,950	28,382	15,875	6,868	30,863	11,048	397
ID	aidi:	ID	2,445	7,182	5,514	4,416	15,272	21,623	38,534	19,633	26,415	6,742	7,932	14,324	5,212	11,877	429
発送	hassou	shipping	2,465	7,937	5,562	4,708	14,841	20,583	20,059	13,630	22,670	9,582	4,319	25,244	14,198	17,688	436

Nine out of the ten words in Table 3-15 and 3-16 are overlapping. Considering the fact that the words with a 1,000 or more ranking gap between U_{DP} and SFI (Table 3-19 and 3-20) do not overlap with the words in Table 3-15 and 3-16, *U* provides rankings to unevenly distributed words differently from the other two indices.

Including the terms for auctions such as 出品 'shuppin' (display, exhibit), 落札 'rakusatsu' (successful bid), オークション 'o^kushon' (auction), 入札 'nyuusatsu' (bidding) and 発送 'hassou' (shipping), seven words in Table 3-15 and eight words in

Table 3-16 have distinctively high-frequency only in the sub-corpus IF (Internet Q & A forum site corpus). These words must be downgraded substantially as the gap between IF and the other corpora is very large.

One possible criterion for judging which index downgrades unevenly-distributed words properly is the average sub-frequency ranking[58], which is lower (i.e. greater in ranking number) than the rankings by other indices (*U*, U_{DP} and SFI) for all the other words in Table 3-15 and 3-16 except for 図書 'tosho' (books, publications). (Only the *U* ranking for 図書 'tosho' (3056) is lower than the average sub-frequency ranking (2,723).) For example, 出品 'shuppin' (display, exhibit) is ranked at 124 in IF while lower than 10,000 in seven sub-corpora out of the ten. The average sub-frequency ranking for the word is 14,393, to which the overall ranking by *U* is the closest at 7,125 while the word is ranked at 2,624 and 2,638 by U_{DP} and SFI respectively. Even the lowest ranking among the three (7,125 by *U*) seems too high, let alone the rankings by U_{DP} and SFI. Considering the fact that the sub-sections of this corpus are differently-sized ones classified based on genre[59] and media, and that the words only frequently used in a domain are not so necessary for learners who don't need to read texts from the domain, the rankings by *U* seem more appropriate than the rankings by the other two indices.

Then, what words have much "higher" *U* rankings than U_{DP} or SFI rankings? Closely comparing the ranking figures between the words in Table 3-15/16 and 3-17/18, three things can be pointed out.

[58] Some people may think that the average sub-frequency ranking can be the overall ranking instead of using adjusted frequency; however, there are at least two problems with the idea. One is that the ten sub-frequencies will be weighted totally the same. Considering the fact that the whole corpus is a balanced corpus where the weight for language users is reflected, the total frequency should also be taken into account. Second is that the ranking in a sub-corpus greatly depends on the number of words and it will influence the average ranking too much. It is also a problem that not all the words are listed in every sub-corpus.

[59] Lyne (1985) called the sub-sections classified based on genre as 'differentiated' sections (p. 126).

Table 3-17 Ranking Comparison of the Most Frequent 10 Words with *U* Ranking Higher than U_{DP} Ranking by 1,000 or More

Lexeme in Kanji & Romanization		English Translation	*F* Ranking	*U* Ranking	U_{DP} Ranking	*SFI* Ranking	Ave. Freq. Rank. in 10 Sub Corpora	LW Freq. Ranking	LP Freq. Ranking	HE Freq. Ranking	AH Freq. Ranking	PL Freq. Ranking	EC Freq. Ranking	SE Freq. Ranking	ST Freq. Ranking	BM Freq. Ranking	IF Freq. Ranking
合弁	gouben	joint management	10,001	11,363	13,324	11,270	17,172	27,231	38,534	5,937	19,965	2,445	3,998	17,819	8,746	25,415	21,630
車種	shashu	model of a car	10,003	11,797	13,270	11,553	20,074	25,441	38,534	19,633	19,965	28,382	8,850	15,839	4,148	36,771	3,181
物作り	monodzukuri	manufacturing	10,005	13,909	16,120	13,073	24,248	54,591	20,059	50,344	22,670	11,493	1,968	4,867	6,942	25,415	44,135
前章	zenshou	previous chapter	10,014	8,784	10,359	9,426	15,392	54,591	4,701	6,079	10,841	5,745	5,425	8,143	5,405	8,857	44,135
箇年	-kanen	-year	10,038	9,929	11,585	10,160	14,876	27,231	11,954	6,795	17,977	6,211	3,998	6,468	3,376	20,616	44,135
フォーラム	fo:ramu	forum	10,040	10,069	11,383	10,068	12,817	27,231	20,059	21,804	12,155	5,745	6,649	3,064	5,405	11,877	14,178
塩基	enki	alkali, base	10,054	14,050	18,824	14,635	24,050	27,231	38,534	34,869	46,296	28,382	13,695	25,244	2,095	2,528	21,630
自明	jimei	self-evident	10,066	8,953	10,280	9,408	10,971	18,883	4,843	12,922	5,680	5,745	9,461	5,148	6,630	15,590	24,812
シンポジウム	shimpojiumu	symposium	10,068	8,989	10,632	9,524	13,097	24,024	4,989	12,922	7,993	4,621	13,695	4,776	6,630	7,181	44,135
上述	joujutsu	above mentioned	10,072	9,247	11,179	9,735	13,866	35,907	5,717	7,698	26,415	4,621	3,688	6,868	6,630	11,048	30,068

Table 3-18 Ranking Comparison of the Most Frequent 10 Words with *U* Ranking Higher than SFI Ranking by 1,000 or More

Lexeme in Kanji & Romanization		English Translation	*F* Ranking	*U* Ranking	U_{DP} Ranking	*SFI* Ranking	Ave. Freq. Rank. in 10 Sub Corpora	LW Freq. Ranking	LP Freq. Ranking	HE Freq. Ranking	AH Freq. Ranking	PL Freq. Ranking	EC Freq. Ranking	SE Freq. Ranking	ST Freq. Ranking	BM Freq. Ranking	IF Freq. Ranking
至難	shi'nan	extremely difficult	13,933	11,109	11,195	12,134	17,172	27,231	38,534	5,937	19,965	2,445	3,998	17,819	8,746	25,415	21,630
瞭然	ryouzen	obvious [lit.]	14,021	11,097	10,944	12,193	20,074	25,441	38,534	19,633	19,965	28,382	8,850	15,839	4,148	36,771	3,181
旅費	ryohi	traveling expenses	14,433	11,768	11,509	12,829	24,248	54,591	20,059	50,344	22,670	11,493	1,968	4,867	6,942	25,415	44,135
噛み合う	kamiau	mesh, in gear	14,657	11,687	11,551	12,765	15,392	54,591	4,701	6,079	10,841	5,745	5,425	8,143	5,405	8,857	44,135
あながち	anagachi	(not) necessarily	14,763	11,805	11,773	12,897	14,876	27,231	11,954	6,795	17,977	6,211	3,998	6,468	3,376	20,616	44,135
填補	tempo	supplementation	15,062	19,424	24,437	20,447	12,817	27,231	20,059	21,804	12,155	5,745	6,649	3,064	5,405	11,877	14,178
ジャンボ	jambo	jumbo, jumbo-sized	15,077	12,075	11,780	13,138	24,050	27,231	38,534	34,869	46,296	28,382	13,695	25,244	2,095	2,528	21,630
似通う	nikayou	resemble closely	15,311	12,289	12,270	13,379	10,971	18,883	4,843	12,922	5,680	5,745	9,461	5,148	6,630	15,590	24,812
正論	seiron	sound argument	15,313	12,602	12,150	13,638	13,097	24,024	4,989	12,922	7,993	4,621	13,695	4,776	6,630	7,181	44,135
難題	nandai	difficult problem	15,435	12,491	12,411	13,499	13,866	35,907	5,717	7,698	26,415	4,621	3,688	6,868	6,630	11,048	30,068

Firstly, the words which have 1,000 or more high *U* rankings do not appear within the top 10,000. This means U_{DP} and SFI do not penalize as many unevenly distributed high/middle-frequency words as *U*.

Secondly, the ranking gaps between the indices in Tables 3-17 and 3-18 are not as large as the ones in Tables 3-15 and 3-16. This means, even for low frequency words, U_{DP} and SFI do not penalize unevenly distributed words so much as *U*.

Thirdly, the words which have 1,000 or more 'high' *U* rankings, in contrast to the words which have 1,000 or more 'low' *U* rankings in Table 3-15/16, have no single domain where the sub-frequency ranking is distinctively high. For example, no words in Table 3-17/18 have a sub-frequency higher (smaller in figure) than 1,000, while all the words which have 1,000 or more lower *U* rankings in Table 3-15/16 have a sub-frequency higher than 1,000. Interestingly, for all the words where the *U* ranking is lower than the U_{DP} or SFI

ranking (Table 3-15/16), the total frequency (*F)* ranking is always higher than rankings by the adjusted frequencies (*U, U_{DP}* and SFI), while the words where the *U* ranking is higher than the U_{DP} or SFI ranking (Table 3-17/18) do not always have the higher *F* ranking than rankings by the adjusted frequencies (*U, U_{DP}* and SFI). This means, for the latter group of words, some words are highly unevenly distributed but some are not. Five out of the ten words where the *U* ranking is higher than the U_{DP} ranking (Table 3-17) have the higher *U* ranking than *F* ranking. What is more, nine out of the ten words where the *U* ranking is higher than the SFI ranking (Table 3-18) have the higher *U* ranking than *F* ranking. This suggests that *U* tends to penalize the words which are distinctively frequently used in only one single domain while U_{DP} and SFI tend to penalize words with wider unevenness.

Before moving to the next step, let us check the words which have a great gap in ranking between U_{DP} and SFI.

Table 3-19 Ranking Comparison of the Most Frequent 10 Words with U_{DP} Ranking Lower than SFI Ranking by 1,000 or More

Lexeme in Kanji & Romanization		English Translation	*F* Ranking	*U* Ranking	U_{DP} Ranking	*SFI* Ranking	Ave. Freq. Rank. in 10 Sub Corpora	LW Freq. Ranking	LP Freq. Ranking	HE Freq. Ranking	AH Freq. Ranking	PL Freq. Ranking	EC Freq. Ranking	SE Freq. Ranking	ST Freq. Ranking	BM Freq. Ranking	IF Freq. Ranking
アドレス	adoresu	address	2,071	3,272	4,084	3,038	11,248	14,273	10,974	24,622	17,977	7,054	4,406	6,660	650	25,415	448
監査	kansa	auditing, inspection	2,335	4,294	5,901	4,071	13,648	21,623	16,911	15,486	46,296	252	704	9,205	5,602	8,857	11,547
ID	aidi:	ID	2,445	7,182	5,514	4,416	15,272	21,623	38,534	19,633	26,415	6,742	7,932	14,324	5,212	11,877	429
譲渡	jouto	transfer, conveyance	2,855	4,340	5,123	4,109	10,365	24,024	16,911	6,079	19,965	847	499	3,709	16,704	8,857	6,059
社債	shasai	corporate bond	3,216	5,782	9,014	6,108	26,812	54,591	38,534	24,622	46,296	442	839	25,244	30,863	36,771	9,913
HP	eichipi:	homepage, web-site	3,343	6,890	6,297	5,157	18,847	54,591	38,534	9,621	32,436	19,095	4,796	13,162	5,212	10,384	637
振り込み	furikomi	direct deposit, transfer	3,450	8,908	7,099	5,764	17,383	19,690	14,826	50,344	32,436	8,345	6,153	14,324	9,381	17,688	643
濃度	noudo	density, concentration	3,487	5,325	6,172	5,153	10,311	21,623	11,954	19,633	9,359	19,095	8,372	7,106	477	798	4,694
入金	nyuukin	recept of money	3,533	10,037	7,508	6,336	19,226	20,583	16,911	34,869	19,965	9,582	5,946	35,801	11,182	36,771	654
送料	souryou	shipping charge	3,545	8,323	7,496	6,268	24,707	54,591	38,534	19,633	32,436	28,382	7,932	35,801	3,682	25,415	661

Table 3-20 Ranking Comparison of the Most Frequent 10 Words with U_{DP} Ranking Higher than SFI Ranking by 1,000 or More

Lexeme in Kanji & Romanization		English Translation	*F* Ranking	*U* Ranking	U_{DP} Ranking	*SFI* Ranking	Ave. Freq. Rank. in 10 Sub Corpora	LW Freq. Ranking	LP Freq. Ranking	HE Freq. Ranking	AH Freq. Ranking	PL Freq. Ranking	EC Freq. Ranking	SE Freq. Ranking	ST Freq. Ranking	BM Freq. Ranking	IF Freq. Ranking
大匙	oosaji	tablespoon	5,101	14,772	9,009	10,685	21,942	41,621	38,534	12,259	16,316	28,382	28,593	17,819	30,863	633	4,398
国債	kokusai	government bonds	5,749	10,923	7,540	8,600	15,869	8,397	20,059	21,804	16,316	5,745	1,053	12,235	30,863	36,771	5,445
銀河	ginga	the Milky Way	5,932	14,892	8,566	10,653	16,035	7,985	7,915	16,577	8,990	19,095	28,593	25,244	627	36,771	8,552
HDD	eichidi:di:	hard disk drive	6,145	43,581	14,462	16,073	34,235	54,591	38,534	50,344	46,296	28,382	19,525	35,801	30,863	36,771	1,239
信心	shinjin	devotion	6,159	12,647	8,543	9,737	15,616	8,827	826	8,969	7,700	28,382	19,525	8,823	30,863	20,616	21,630
オブジェクト	obujekuto	object [computing etc.]	6,558	25,862	14,879	18,228	29,250	54,591	38,534	50,344	10,841	28,382	28,593	35,801	533	36,771	8,114
編む	amu	knit [v.]	6,970	8,070	6,331	7,370	8,927	6,478	5,928	6,341	6,440	11,493	19,525	8,436	1,752	14,082	8,790
小匙	kosaji	teaspoon	6,987	23,538	14,249	16,479	27,284	41,621	38,534	24,622	46,296	28,382	28,593	35,801	21,003	904	7,086
膣	chitsu	vagina	7,097	10,080	7,772	8,846	16,706	7,758	16,911	50,344	13,905	28,382	19,525	7,106	16,704	1,681	4,747
質量	shitsuryou	mass [physics]	7,278	12,914	9,270	10,308	12,635	24,024	9,494	17,992	22,670	9,582	11,085	17,819	970	7,181	5,533

In both tables, there is no notable feature with the distribution of sub-frequencies as *U* has. The most frequent ten words with U_{DP} ranking lower than SFI ranking by 1,000 or more (Table 3-19) are in the range of relatively high *F* rankings between 2,000 and 3,600. On the other hand, the most frequent ten words with U_{DP} ranking lower than SFI ranking by 1,000 or more (Table 3-20) are in the range of relatively low *F* rankings between 5,100 and 7,300. This suggests that U_{DP} and SFI will return different types of rankings and that SFI tends not to penalize unevenly distributed words as a whole.

4) The correlation coefficients between skewness (absolute value)/kurtosis and other indices are from Table 3-21 to 3-25

Table 3-21 Spearman's Rank Correlations Coefficients (Rho) between Skewness, Kurtosis, Frequency, Dispersion and Adjusted Frequency for the Words in VDRJ

(F Ranking 0001-20,180, F≧48, Range≦10) N = 20,180

	Skew_Abs	Kurtosis	*F*	*Range*	*D*	(1-*DP*)	*D2*	*U*	U_{DP}	*SFI*	ASFR
Skew_Abs	1.000	.984	-.089	-.261	-.415	-.411	-.387	-.208	-.191	-.167	.279
Kurtosis	.984	1.000	-.076	-.227	-.352	-.359	-.326	-.184	-.167	-.146	.244
F	-.089	-.076	1.000	.645	.343	.315	.405	.955	.965	.975	-.898
Range	-.261	-.227	.645	1.000	.601	.558	.683	.744	.734	.731	-.838
D	-.415	-.352	.343	.601	1.000	.913	.987	.547	.517	.488	-.625
(1-*DP*)	-.411	-.359	.315	.558	.913	1.000	.922	.504	.509	.453	-.597
D2	-.387	-.326	.405	.683	.987	.922	1.000	.600	.576	.546	-.691
U	-.208	-.184	.955	.744	.547	.504	.600	1.000	.991	.994	-.965
U_{DP}	-.191	-.167	.965	.734	.517	.509	.576	.991	1.000	.994	-.964
SFI	-.167	-.146	.975	.731	.488	.453	.546	.994	.994	1.000	-.956
ASFR	.279	.244	-.898	-.838	-.625	-.597	-.691	-.965	-.964	-.956	1.000

Skew Abs.: Absolute value of skewness ASFR: Average sub-frequency ranking

$p < .001$ for all correlation coefficients

Table 3-22 Spearman's Rank Correlations Coefficients (Rho) between Skewness, Kurtosis, Frequency, Dispersion and Adjusted Frequency for the Words with *Range* 8 or less in VDRJ

(F Ranking 0001-20,180, F≧48, Range≦8) N = 5,216

	Skew_Abs	Kurtosis	*F*	*Range*	*D*	*1-DP*	*D2*	*U*	*UDP*	*Um*	ASFR
Skew_Abs	1.000	.994	.080	-.310	-.531	-.527	-.494	-.322	-.251	-.195	.435
Kurtosis	.994	1.000	.070	-.286	-.492	-.488	-.452	-.309	-.238	-.187	.403
F	.080	.070	1.000	.203	-.119	-.108	-.103	.675	.756	.797	-.449
Range	-.310	-.286	.203	1.000	.561	.538	.653	.557	.522	.531	-.795
D	-.531	-.492	-.119	.561	1.000	.891	.984	.556	.420	.398	-.648
1-DP	-.527	-.488	-.108	.538	.891	1.000	.911	.494	.489	.377	-.682
D2	-.494	-.452	-.103	.653	.984	.911	1.000	.559	.447	.419	-.701
U	-.322	-.309	.675	.557	.556	.494	.559	1.000	.928	.961	-.824
UDP	-.251	-.238	.756	.522	.420	.489	.447	.928	1.000	.962	-.825
Um	-.195	-.187	.797	.531	.398	.377	.419	.961	.962	1.000	-.784
ASFR	.435	.403	-.449	-.795	-.648	-.682	-.701	-.824	-.825	-.784	1.000

Skew Abs.: Absolute value of skewness ASFR: Average sub-frequency ranking

$p < .001$ for all correlation coefficients

Table 3-23 Spearman's Rank Correlations Coefficients (Rho) between Skewness, Kurtosis, Frequency, Dispersion and Adjusted Frequency for the Words with *Range* 6 or less in VDRJ

(F Ranking 0001-20,180, F≧48, Range≦6) N = 1,700

	Skew_Abs	Kurtosis	*F*	*Range*	*D*	*1-DP*	*D2*	*U*	*UDP*	*Um*	ASFR
Skew_Abs	1.000	.998	.071	-.368	-.668	-.605	-.628	-.545	-.400	-.364	.587
Kurtosis	.998	1.000	.066	-.352	-.648	-.584	-.606	-.535	-.390	-.356	.565
F	.071	.066	1.000	.074	-.108	-.112	-.110	.441	.604	.626	-.213
Range	-.368	-.352	.074	1.000	.560	.555	.646	.518	.471	.487	-.773
D	-.668	-.648	-.108	.560	1.000	.874	.985	.777	.565	.605	-.767
1-DP	-.605	-.584	-.112	.555	.874	1.000	.905	.662	.646	.557	-.822
D2	-.628	-.606	-.110	.646	.985	.905	1.000	.760	.583	.609	-.817
U	-.545	-.535	.441	.518	.777	.662	.760	1.000	.868	.934	-.781
UDP	-.400	-.390	.604	.471	.565	.646	.583	.868	1.000	.943	-.776
Um	-.364	-.356	.626	.487	.605	.557	.609	.934	.943	1.000	-.724
ASFR	.587	.565	-.213	-.773	-.767	-.822	-.817	-.781	-.776	-.724	1.000

Skew Abs.: Absolute value of skewness ASFR: Average sub-frequency ranking

$p < .001$ for all correlation coefficients

Table 3-24 Spearman's Rank Correlations Coefficients (Rho) between Skewness, Kurtosis, Frequency, Dispersion and Adjusted Frequency for the Words with *Range* 4 or less in VDRJ

(F Ranking 0001-20,180, F≧48, Range≦4) N = 437

	Skew_Abs	Kurtosis	*F*	*Range*	*D*	*DP2*	*D2*	*U*	*UDP*	*Um*	ASFR
Skew_Abs	1.000	1.000	.075	-.440	-.827	-.636	-.799	-.737	-.455	-.483	.668
Kurtosis	1.000	1.000	.075	-.435	-.823	-.630	-.794	-.734	-.451	-.480	.661
F	.075	.075	1.000	.097	-.054	-.074	-.057	.285	.544	.522	-.124
Range	-.440	-.435	.097	1.000	.530	.490	.608	.550	.463	.494	-.780
D	-.827	-.823	-.054	.530	1.000	.809	.989	.902	.601	.708	-.769
DP2	-.636	-.630	-.074	.490	.809	1.000	.836	.722	.729	.677	-.809
D2	-.799	-.794	-.057	.608	.989	.836	1.000	.891	.617	.713	-.819
U	-.737	-.734	.285	.550	.902	.722	.891	1.000	.793	.892	-.781
UDP	-.455	-.451	.544	.463	.601	.729	.617	.793	1.000	.937	-.731
Um	-.483	-.480	.522	.494	.708	.677	.713	.892	.937	1.000	-.710
ASFR	.668	.661	-.124	-.780	-.769	-.809	-.819	-.781	-.731	-.710	1.000

Skew Abs.: Absolute value of skewness ASFR: Average sub-frequency ranking

$p < .001$ for all correlation coefficients

Table 3-25 Spearman's Rank Correlations Coefficients (Rho) between Skewness, Kurtosis, Frequency, Dispersion and Adjusted Frequency for the Words with *Range* 2 or less in VDRJ

(F Ranking 0001-20,180, F≧48, Range≦2) N = 99

	Skew_Abs	Kurtosis	*F*	*Range*	*D*	*DP2*	*D2*	*U*	*UDP*	*Um*	ASFR
Skew_Abs	1.000	1.000	.053	-.854	-.945	-.646	-.945	-.852	-.508	-.581	.834
Kurtosis	1.000	1.000	.053	-.854	-.945	-.646	-.945	-.852	-.508	-.581	.834
F	.053	.053	1.000	.044	.000	.116	.000	.120	.516	.418	-.157
Range	-.854	-.854	.044	1.000	.854	.619	.854	.686	.529	.583	-.783
D	-.945	-.945	.000	.854	1.000	.728	1.000	.934	.598	.714	-.872
DP2	-.646	-.646	.116	.619	.728	1.000	.728	.725	.867	.833	-.883
D2	-.945	-.945	.000	.854	1.000	.728	1.000	.934	.598	.714	-.872
U	-.852	-.852	.120	.686	.934	.725	.934	1.000	.665	.774	-.846
UDP	-.508	-.508	.516	.529	.598	.867	.598	.665	1.000	.935	-.792
Um	-.581	-.581	.418	.583	.714	.833	.714	.774	.935	1.000	-.802
ASFR	.834	.834	-.157	-.783	-.872	-.883	-.872	-.846	-.792	-.802	1.000

Skew Abs.: Absolute value of skewness ASFR: Average sub-frequency ranking

$p < .001$ for all correlation coefficients

Skewness (a measure of the asymmetry of the distribution) and kurtosis (a measure of flatness of the distribution) correlate very highly. These two show similar features of the distribution patterns for this data set at least.

Now, let's look at the correlation coefficient between skewness/kurtosis and dispersion/adjusted frequency indices. When the coefficients are calculated for all the top 20,180 words which occur 48 times or more in the whole corpus, there is no significant difference between the indices; however, when narrowing down the *Range* from 8 to 2, it

comes clearer that *D* and *U* have the highest reverse correlation with skewness and kurtosis among the dispersion and adjusted frequency indices respectively. The gap between the correlation coefficient between skewness and *D* and the correlation coefficient between skewness and *DP* is not significant for the words with a Range of 8 or less (n = 5216, skewness M = 2.18, SD = .77, r_D = -.531, r_{DP} = -.527, $p > .754$, n.s.). For the words with a Range of 6 or less, however, there is a significant difference between the two (n = 1700, skewness M = 2.49, SD = .66, r_D = -.668, r_{DP} = -.605, $p < .01$), and the gap becomes greater for the words with a Range of 4 or less (n = 437, skewness M = 2.79, SD = .51, r_D = -.827, r_{DP} = -.636, $p < .001$) and a Range of 2 or less (n = 99, skewness M = 3.03, SD = .32, r_D = -.945, r_{DP} = -.646, $p < .001$) [60].

The dispersion figure will be smaller for the more unevenly distributed words (*DP* will increase in number; however, the figure will decrease in the same way as *D* or D_2 as (1-*DP*) is used here). Therefore, the reverse correlation here means that the more the skewness and kurtosis, the more unevenly the word is distributed. Here Spearman's rank correlation is used, which means that *D* tends to penalise the ranking with the words with high skewness and kurtosis more severely. The result is consistent with the results in 2) and 3). It also agrees with Lyne (1985) who claims that *D* is more sensitive to skewness than D_2 (p.129). In addition, *D* is more sensitive to skewness and kurtosis than *DP* as well. Compared to *D*, *DP* tends to be more sensitive to the unevenness as a whole. Contrary to that, *D* will react more strongly to the uneven distribution caused by a single sub-section.

Taking all of these results into account, for the case where there is no significant difference as a whole, and only highly unevenly distributed words are to be evaluated, *D*

[60] The following equation was used for examining the gap between the two correlation coefficients (Institute of JUSE, 2010).

$z = \frac{z_1 - z_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}}$ where $z_1 = \frac{1}{2} ln\left(\frac{1+r_1}{1-r_1}\right)$ and $z_2 = \frac{1}{2} ln\left(\frac{1+r_2}{1-r_2}\right)$.

n_1, n_2: number of data, r_1, r_2: correlation coefficient.

will be the most suitable index as a dispersion measure. As a consequence, *U* will be the most suitable index as an adjusted frequency measure.

As Gries (2008) points out, there will be a problem with the words with a range of 1 where the words are at the same ranking regardless of the frequency if *D* or *U* is adopted (p.412). Nevertheless, these words are very low-frequency at the tens of thousands ranking level where little importance is found for ranking words for educational purposes. As is the case with this study, for ordering words in the most frequent twenty thousand for practical purposes, the weakness with *D*, which Gries points out, will be of little consequence.

As Gries points out, D_2, which is the dispersion measure used for computing SFI, will generally return a similar figure to *D*, but is not as sensitive to skewness as *D* is. As a consequence, SFI will not greatly penalise words unevenly distributed in one or two domains.

3.3.5.3 Conclusion for 3.3.5

U is adopted to order the words for the Vocabulary Database for Reading Japanese as it seems best fit to this study for the reasons given below.

1) Salience of frequency of a single domain can be due to occasional frequent use of the word in one or a limited number of texts. To fix this kind of sampling bias, it is better to strongly penalize words which are distinctively more frequently used in one single domain than the other domains. In particular, for the high-frequency range where there are more learners' needs, it is better to use an index by which the distinctively unevenly distributed words will be excluded. (As shown in Tables 3-10 to 3-16, correlations between the adjusted frequency measures are very high overall, and less than 20% of the most frequent 20,000 words have a ranking gap of 1,000 or more.)
2) The whole corpus is a monitor version of a balanced corpus where texts are sampled in a strict manner (in the book corpus and the internet Q & A forum site corpus

respectively). That means the total frequency can reflect the degree of language users' contact with different genres. Therefore, dispersion should reflect more about aspects which the total frequency will not show, i.e. how many different types of genres and media the word is used frequently in. U_{DP} and SFI do not seem to have enough power to do this.

3.3.6 Criteria for ordering words (2): Weighting sub-frequencies depending on purposes

The research question here is: SRQ 3) Are the most appropriate word ranking criteria different depending on target learners such as general learners or international students? If yes, what are the more suitable criteria for those different learner groups?

Nation (2004) shows that the adult, formal, British nature of the British National Corpus (BNC). For example, in the first 1,000 words, the BNC list has words such as *commission* and *labour* while the orally very common words such as *goodbye* and *damn* are in the fourth 1,000 list. Therefore, only the 10-million-word spoken part of the BNC was used to rank words in the first and the second 1,000 lists in Nation's lists (Nation & Webb, 2011, p 141).

This study also has the same problem. For example, words such as さようなら 'sayounara' (goodbye) and あさって'asatte' (the day after tomorrow) are at Level 4 (the most basic level) of the former Japanese Language Proficiency Test (F-JLPT) word lists and usually appear in elementary text books; However, the words are ranked at 6,338 and 16,912 respectively by the adjusted frequency (Juilland's *U*) ranking in VDRJ, the database developed for this study. Contrary to that, words such as 行為 'koui'(behaviour) and システム 'shisutemu'(system) are at Level 1 (the most advanced level) of the F-JLPT word lists but are ranked at 608 and 705 respectively in VDRJ.

As shown in Table 3-26, the top 2,000 words in VDRJ (W_01K and 02K) contain a considerable number of words which are generally thought to be intermediate or advanced. 43% of the VDRJ top 1,000 words are at Level 2 or above of the F-JLPT ((374+42+25)/1024= .43), where many formal or academic words are listed.

Some people may think that the VDRJ word list should have a formal written nature, because this study explores the word list for reading. Nevertheless, elementary learners will rarely acquire the written language in natural settings outside the classroom so that the settings outside the classroom can account for the acquisition of written language only after the intermediate level. In particular, for reading comprehension of authentic texts, a certain degree of text coverage by known words will be required. Therefore, text books will have a stronger impact on the acquisition of written language in general (See footnote 18 for some criteria other than frequency suggested in previous studies for selecting basic words).

3.3.6.1 Reasons for weighting sub-frequencies to create different word rankings

Assumed users of VDRJ and the word lists are 1) researchers, 2) academic learners such as international students, 3) non-academic "general" learners, and 4) the teachers and course designers for the learners mentioned above. For their convenience, in consideration of the issues with ranking basic words, this study proposes three types of word rankings shown below.

1) The Word Ranking for Written Japanese (WWJ)
2) The Word Ranking for International Students (WIS)
3) The Word Ranking for General Learners (WGL)

Table 3-26 Number of Words by VDRJ Word Level (Ranked by Juilland's *U*) and the Former Japanese Language Proficiency Test (F-JLPT) Word Level

Word Level (*1)	F-JLPT Level 4	F-JLPT Level 3	F-JLPT Level 2	F-JLPT Level 1	Not-in-the-Lists	Total
W_01K (*2)	355	228	374	42	25	1,024
W_02K	138	133	518	144	67	1,000
W_03K	72	89	448	239	152	1,000
W_04K	32	63	367	263	275	1,000
W_05K	27	19	311	259	384	1,000
W_06K	20	17	222	257	484	1,000
W_07K	9	17	180	219	575	1,000
W_08K	9	15	147	192	637	1,000
W_09K	9	7	131	167	686	1,000
W_10K	4	6	96	163	731	1,000
W_11K	7	2	81	135	775	1,000
W_12K	5	3	57	75	860	1,000
W_13K	3	3	49	92	853	1,000
W_14K		1	53	81	865	1,000
W_15K	3	2	29	55	911	1,000
W_16K	1	2	39	60	898	1,000
W_17K	1		22	46	931	1,000
W_18K	1		22	39	938	1,000
W_19K		1	19	48	932	1,000
W_20K	1		11	28	960	1,000
W_21K+	7	2	94	194	90,803	91,100
W_AKW (*3)	1		4	1	30,819	30,825
Total	705	610	3,274	2,799	134,340	141,949

*1 Among the four levels of the F-JLPT, Level 4 & 3 are thought to be elementary, Level 2 is intermediate, and Level 1 and Not-in-the-List are advanced. The word levels from W_01K to W_21K+ and W_AKW are the levels defined by Juilland's *U* in VDRJ

*2 W_01K' includes 24 words of 'W_01K+' which is the list for compound numerals.

*3 AKW stands for Assumed Known Words which are mostly proper nouns.

These are created based on different ranking criteria. In WWJ, the words are genuinely ranked by *U* where the ten sub-sections are equally weighted. WIS primarily serves for international students studying in Japanese universities, since the corpus used for making the lists is composed of texts collected in Japan. This ranking is made by weighting the sub-sections which have a relatively strong academic orientation. WGL is the word ranking for learners who study Japanese mainly for non-academic purposes. It is the word

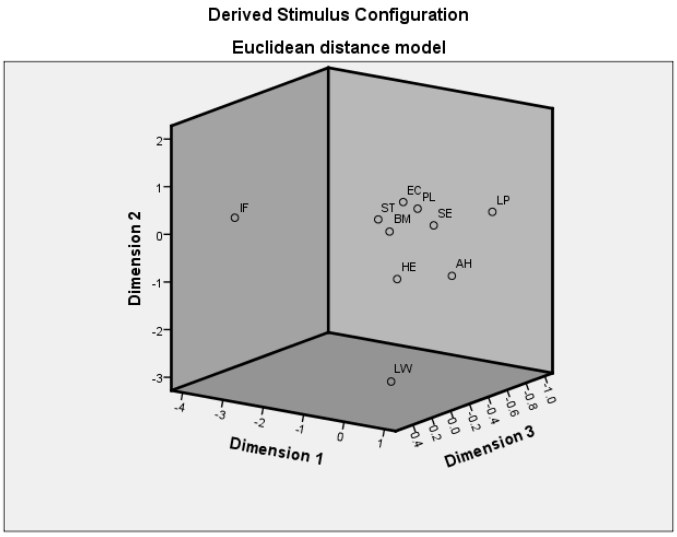
ranking list more for daily life, and is made by weighting sub-sections which have a non-academic orientation.

Then, how should the sub-sections be weighted on to create these different types of word lists? To create Nation's list, only the spoken section of the British National Corpus was used for selecting the first and second 1,000 word lists (Nation & Webb, 2011). However, not only is there no balanced spoken corpus suitable for measuring word frequency in general, but also using a spoken corpus is not a suitable way to make a word list for reading. In addition, it is hard to define the target domain at the basic level as many elementary learners do not have clear purposes for learning the language. Given these, to include the elementary course book vocabulary in the basic word list seems a practical solution as these words will more or less reflect the daily life needs, and the importance of written language for the second language learners at the elementary level is assumed for the preparation for reading authentic texts at the intermediate level or above.

Taking these factors into account, all the sub-frequencies are standardized as frequency per million first. This is a necessary step for weighting differently on different sub-sections depending on different purposes. The standardized frequencies can be used for calculating *F* (frequency) by weighting sub-frequencies differently where *U* is the product of *F* and *D* (dispersion).

To clarify the features of the sub-corpora for deciding the amount of weighting on them, multidimensional scaling (MDS) was conducted to examine how the frequency distributions of sub-sections are related to each other. MDS is a statistical technique to explore the similarities in data and visualize them on an N (generally two or three) - dimensional image.

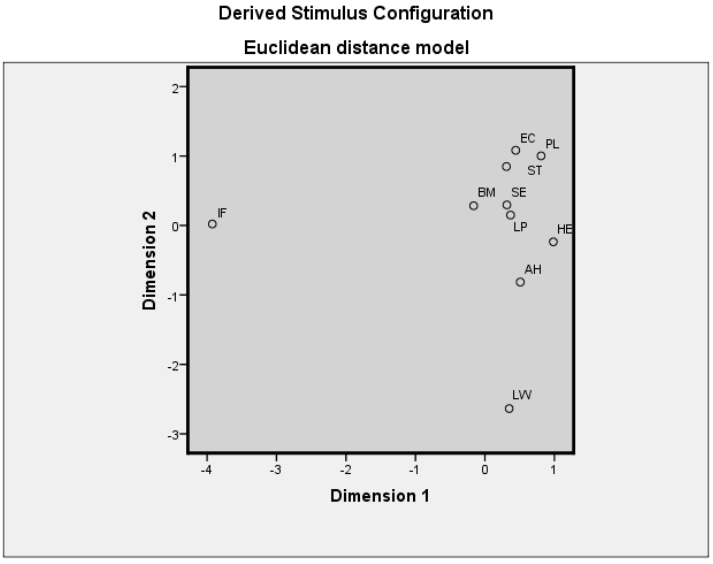
Figure 3-1 Multidimensional Scaling for Frequency Distribution of the Ten Sub-Sections in VDRJ (Three-dimensional)



The codes for the ten sub-sections for Figures 3-1 to 3-2 (See also Table 3-4)

LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.

Figure 3-2 Multidimensional Scaling for Frequency Distribution of the Ten Sub-Sections in VDRJ (Two-dimensional)



Figures 3-1 and 3-2 clearly show that the distribution patterns of the ten sub-sections can be divided into three categories of IF (internet Q &A forum sites), LW (literary works) and the other eight sections (henceforth AD: academic domains). As mentioned in 3.3.2, AD contains technical texts which have '3' at the thousands digit of C-code, and it is classified into the eight domains based on academic disciplines.

The classification into the three categories also corresponds to three of the four categories (fiction, academic prose, conversation (≒IF), newspaper) of Biber's (1995) classification of register variation. (Newspaper texts and magazine texts are not included in the corpus used for this study. As a register, the book text is expected to be ranked between newspaper and magazine texts. See 4.2 in Chapter 4.)

To explore more features of the three sections of IF, LW and AD, the following three issues are examined. 1) The number of words shared by the most frequent 1000 words of the three sections, 2) The distribution of the former Japanese Language Proficiency Test (F-JLPT) vocabulary (Level 1 to 4) across the most frequent 2000 words of the three sections, and 3) The different patterns of text coverage of IF, LW or AD.

Table 3-27 Words Listed in the Top 1,000 in the Word Frequency Lists of Internet Forum (IF), Literary Works (LW) and the Eight Academic Domains (AD) of VDRJ

	All (IF, LW & AD)	IF & LW only	IF & AD only	LW & AD only	IF only	LW only	AD only
Number of Words	475	118	134	103	273	304	288
Example (*)	恐らく 器 他人	消す はずす ごめん	最高 請求 負担	東 以来 建物	機種 送信 バイト	ねえ 瞳 不意	競争 債務 概念
English Translation of the Examples	probably container other people	put out remove sorry	supreme claim burden	east since then building	model of a machine transmission part-time job /byte	hey eye (lit.) unexpected(ly)	competition debt concept

* Examples are selected from the bottom (least frequent) of each category according to the total frequency ranking in VDRJ.

* Add up number of words belonging to each category of IF/LW/AD together,that comes 1,000.

As shown in Table 3-27, only less than half of words listed in the top 1000 in each of the three sections are overlapping. This means these three sections have considerably different lexical features. IF contains more colloquial words such as ごめん 'gomen' (sorry). IF vocabulary tends to reflect more daily needs than the other two, except some words specific in the internet community such as カテ 'kate' (category for a forum topic) and 送信 'soushin' (transmission). LW seems to contain more written vocabulary than IF, but is less formal than AD. It covers a wide range of basic vocabulary as well as IF, except some words specific in literary works such as 瞳 'hitomi' (eye (lit.)). AD contains more formal and academic words such as 概念 'gainen' (concept) than the other two.

As shown in Table 3-28 and Figure 3-3, LW covers more basic words (i.e. the F-JLPT Level 4 & 3 vocabulary) and IF comes to the second. AD contains more intermediate and advanced vocabulary (i.e. the F-JLPT Level 2 & 1 vocabulary) in the top 2000; however, AD seems to contain less low frequency or domain-specific words than IF and LW in the top 2000 as it has less words other than Level 4 to 1 vocabulary of the F-JLPT.

Table 3-28 Number of Words in the Word Frequency Lists of Internet Forum (IF), Literary Works (LW) and the Eight Academic Domains (AD) of VDRJ by the Former JLPT (F-JLPT) Word Level

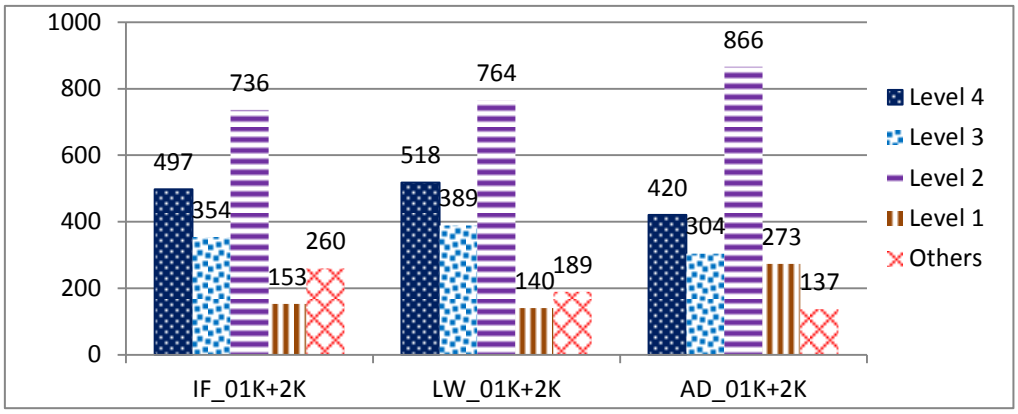
The F-JLPT Word Level	Level 4	Level 3	Level 2	Level 1	Others	Total
IF_01K	368	219	285	51	77	1,000
LW_01K	391	230	295	41	43	1,000
AD_01K	304	179	401	77	39	1,000
IF_02K	129	135	451	102	183	1,000
LW_02K	127	159	469	99	146	1,000
AD_02K	116	125	465	196	98	1,000
IF_03K	72	81	407	147	293	1,000
LW_03K	56	64	421	155	304	1,000
AD_03K	89	89	394	234	194	1,000

* F-JLPT: The former Japanese Language Proficiency Test, K: 1,000 words (e.g. 01K: the first 1,000 words)
* Among the four levels of the former JLPT, Level 4 & 3 are thought to be elementary, Level 2 is intermediate, and Level 1 and Not-in-the-List are

Tables 3-29 to 3-31 show that the three frequency rankings are different in text coverage to a large extent. This proves that if a learner or a teacher does not follow an appropriate order of vocabulary learning/teaching, it would be very inefficient. Table 3-29 shows that LW is closer to IF than AD up to the 70-80% coverage level (around the 1,000 word level in LW and AD); however, beyond that, AD is closer to IF. Table 3-30 shows that IF ranking covers LW texts better than AD up to the 60-70% coverage level (between 100 and 450 word levels in IF and AD); however, beyond that, AD ranking covers LW texts better than IF. As shown in Table 3-31, AD texts have higher lexical diversity at all levels than IF and LW. Interestingly, IF ranking covers AD texts better than LW up to 95-98% coverage level (between 20,000 and 50,000 word levels in IF and LW) and LW

overtakes IF beyond 95-98% coverage level. LW is expected to share the nature of written language with AD; however, IF will probably share some genres with AD while LW is totally different from AD in genre.

Graph 3-1 Number of Words out of the Most Frequent 2000 Words in the Three Sections of Internet Forum (IF), Literary Works (LW) and the Eight Academic Domains (AD) of VDRJ in the Former JLPT (F-JLPT) Word Levels



*F-JLPT: The former Japanese Language Proficiency Test, K: 1000 words (e.g. 01K: the first 1000 words)

Table 3-29 Number of Words Needed to Gain Different Levels of Coverage of the Internet Forum Texts by the Word Lists of Internet Forum (IF), Literary Works (LW) and the Eight Academic Domains (AD) of VDRJ

Text Coverage	60%	70%	80%	90%	95%	98%
IF	65	198	680	2,610	6,288	14,437
LW	76	282	1,267	6,508	19,248	46,109
AD	93	299	1,174	4,508	11,897	29,516

Table 3-30 Number of Words Needed to Gain Different Levels of Coverage of the Literary Texts by the Word Lists of Internet Forum (IF), Literary Works (LW) and the Eight Academic Domains (AD) of VDRJ

Text Coverage	60%	70%	80%	90%	95%	98%
IF	100	453	2,019	9,552	25,259	58,812
LW	84	287	1,119	4,812	11,519	22,820
AD	114	443	1,940	7,911	19,739	43,075

Table 3-31 Number of Words Needed to Gain Different Levels of Coverage of the Eight Academic Domain Texts by the Word Lists of Internet Forum (IF), Literary Works (LW) and the Eight Academic Domains (AD) of VDRJ

Text Coverage	60%	70%	80%	90%	95%	98%
IF	171	712	2,387	8,343	21,471	55,578
LW	187	804	2,650	10,030	23,318	49,104
AD	131	461	1,470	5,279	12,621	27,657

To sum up, IF vocabulary is more basic and less diverse than the other two, as it gains higher text coverage with fewer words. LW vocabulary has the nature of written language; however, it is not academic or formal but contains a wide range of general basic vocabulary as well as literary words. AD vocabulary tends to be more academic or formal, and includes more intermediate and advanced vocabulary than the other two. Nevertheless, no sub-section seems to reflect the ordinary elementary learner's basic daily-life needs which are expected to be reflected in Japanese language text books. It would be, as a compromise, the best way to put the F-JLPT levels 4 and 3 vocabularies at the top of the rankings for learners.

Based on the results mentioned above, besides the genuine usage coefficient (*Uw*) ranking for written Japanese by *U* with no weighting on any sub-sections, this study proposes word rankings for international students and general learners in the following ways. (See Table 3-32 for weights on the sections.)

1) Compute the mean standardized frequency for AD (the eight sub-sections other than IF and LW).
2) Compute the mean total standardized frequency for IF, LW and AD (*Fr1*) by weighting the same amount for these three sections. In other words, the eight sections of AD are only weighted one third. (In the genuine usage coefficient (*Uw*) ranking, AD accounts for 59%. See also Table 3-5 and 3-32)

3) Compute the mean total standardized frequency only for IF and LW (*Fr2*) by weighting 50% to each. (AD accounts for zero.)

4) Compute the adjusted frequencies Ur1/Ur2 by multiplying Fr1/Fr2 and *D*.

Table 3-32 Weights (percentages) on the Sections of Internet Forum (IF), Literary Works (LW) and the Eight Academic Domains (AD) of VDRJ for the Different Word Ranking indices

Usage Coefficient Type	Frequency Type	IF	LW	AD
*Uw = F*D*	*F*	15.9	25.1	59.0
*Ur1 = Fr1*D*	*Fr1*	33.3	33.3	33.3
*Ur2 = Fr2*D*	*Fr2*	50.0	50.0	0.0

F : Standardized frequency per million in VDRJ
Fr1 : Standardized frequency per million in VDRJ by weighting one third on each of the three genres of IF, LW and AD
Fr2 : Standardized frequency per million in VDRJ by weighting only on IF and LW with the same weight i.e. 50% for each

5) Besides the Word Ranking for Written Japanese (WWJ) where words are ordered only by *Uw*, the Word Ranking for International Students (WIS) and the Word Ranking for General Learners (WGL) are also created based on the ordering criteria shown in Table 3-33. (All the rankings are included in VDRJ.)

For WIS and WGL, basic vocabulary i.e. the F-JLPT Level 4 and 3 vocabulary is ordered by *Ur2* which only takes IF and LW into account, since AD is too formal for the level. Also, for WIS and WGL, the words at Level 2 or above were all sorted only by the second key up to the 20 K level, because, with the F-JLPT level criteria for character and vocabulary, there is a clear distinction between Level 3 or lower and Level 2 or above while there seems no clear distinction between Level 2 and beyond[61]. Beyond the 20 K level, only

[61] According to the F-JLPT level criteria for character and vocabulary, Level 4 and 3 aim at the daily life, Level 2 aims at "ordinary things", and Level 1 aims at the "social life" and "comprehensive Japanese"; however, before the introduction of Examination for Japanese University Admission for International Students

approximately 300 words are listed in the F-JLPT word lists. These words are ranked at the 20,001 and beyond in order of the levels of the F-JLPT.

Table 3-33 Methods for the Word Ranking for Written Japanese (WWJ), International Students (WIS) and General Learners (WGL)

Word Ranking	*WWJ*	*WIS*		*WGL*	
	1st Key	1st Key	2nd Key	1st Key	2nd Key
1-681	*Uw*	F-JLPT4	*Ur2*	F-JLPT4	*Ur2*
682-1,291	*Uw*	F-JLPT3	*Ur2*	F-JLPT3	*Ur2*
1,292-2,000	*Uw*	F-JLPT2-0	*Uw*	F-JLPT2-0	*Ur2*
2,001-20,000	*Uw*	F-JLPT2-0	*Uw*	F-JLPT2-0	*Ur1*
20,001+	*Uw*	F-JLPT2/F-JLPT1/F-JLPT0	*Uw*	F-JLPT2/F-JLPT1/F-JLPT0	*Ur1*

* WIS is priamarily assumed to be served for international students studying at Japanese universities as the texts in the corpus is mainly collected in Japan.
* WGL is assumed to be served for learners with non-academic purposes.
* F-JLPT: The former Japanese Language Proficiency Test word list level. 4 is the most basic, 1 is the highest and 0 is out of the levels (beyond 1).
* *Ur1* : Usage coefficient revised version 1 = *Fr1* *D
  Fr1 : (AD+LW+OC)/3
  AD: Standardized frequency per million of the 8 academic domains of LP, HE, AH, PL, EC, SE, ST and BM
* *Ur2* : Usage coefficient revised version 2 = *Fr2* *D
  Fr2 : (LW+OC)/2
  LW/OC: Standardized frequency per million in LW/OC
* Words are sorted by descending order with the indices.

For the words ranked at F-JLPT Level 2 or above, i.e. those ranked at 1,292 or above, different criteria were adopted for WGL and WIS. For WGL, up to the top 2,000, the words are ordered by *Ur2* which is a daily-life-oriented criterion, and the words ranked at 2,001 or above are ordered by *Ur1* which partly takes AD into account. The border between the basic and the intermediate is set at the 2,000 word level, because in teaching English as a second language, the General Service List (West, 1953) contains 2,000 words serving as basic words, and in teaching Japanese, it is also said that 2,000 words are required to complete the basic or elementary level (NLRI, 1984).

(EJU) in 2002, there was no public examination used for university admission, therefore, Level 2 and 1 vocabulary lists apparently include academic vocabulary frequently used in Japanese universities.

For WIS, the words ranked at F-JLPT Level 2 or above are all ordered by *Uw*, because more F-JLPT words ranked higher in *Uw* ranking than in *Ur1* or *Ur2* ranking. (The F-JLPT Level 2 and 1 vocabulary seem to be selected as essential words for the international students in Japan (See footnote 34).) The same criterion as WGL was adopted to order the words ranked at 20,001 or lower in WIS.

3.3.6.2 Conclusion for 3.3.6

The question here is SRQ 3) "Are the most appropriate word ranking criteria different depending on the target learners such as general learners or international students?" The simple answer will be yes. The text coverage data will prove this prediction in 3.5. The further question was "What are the more suitable criteria for those different learner groups?" The answers and the reasons are as follows.

1) The word ranking by Juilland's *U* (WWJ) shows that BCCWJ has a formal and written nature as the British National Corpus does. This is particularly problematic for ordering words at the basic level as learners will not generally learn the written language in natural settings.
2) The result of the multidimensional scaling shows that the ten sub-sections in BCCWJ can be divided into the three categories of the Internet Q&A forum sites (IF) and literary works (LW) and the other eight (AD). IF and LW vocabulary will fit the basic and daily-life needs better than AD while AD contains more academic and formal words than the other two.
3) In light of the conditions mentioned above, the words at the F-JLPT Level 4 and 3 are put at the top of the word rankings for international students (WIS) and general learners (WGL). For both word rankings, the weighted frequency measure in combination with IF and LW (*Ur2*) is used to order the words at the basic level. For the

words from the intermediate to 20,000 word levels, differently weighted frequency measures (*Uw* and *Ur1*) were used to order words for WIS and WGL.

3.4 The product: the Vocabulary Database for Reading Japanese (VDRJ)

For 141,950 lexemes, VDRJ provides information in the 84 fields shown in Table 3-34. As explained in 3.2.2, VDRJ was developed based on the Balanced Contemporary Corpus of Written Japanese (BCCWJ) 2009 monitor version (NINJAL, 2009) which contains approximately 33 million running words from books and internet forum sites.

The completed version of VDRJ is available from the accompanying CD or http://tatsuma2010.web.fc2.com/. The database was first published in 2010 under the name of TM Word List (from Version 1.0 to Version 3.3) (Matsushita, 2010), and changed the name to the Vocabulary Database for Reading Japanese (VDRJ) with more data added in 2011. The current VDRJ version is 1.1.

The five forms of database shown below are provided on the CD and the web-site.

1) The Vocabulary Database for Reading Japanese (VDRJ) for Research
2) The Vocabulary Database for Reading Japanese (VDRJ) for Teachers
3) The Vocabulary Database for Learners of Japanese (VDLJ): For International Students
4) The Vocabulary Database for Learners of Japanese (VDLJ): For General Learners
5) The Vocabulary Database for Learners of Japanese (VDLJ): Basic 2500

The first one 1) VDRJ for Research is the full version, and the others are created by reducing the information for users' convenience.

Table 3-34 Field Names of the Vocabulary Database for Reading Japanese (VDRJ) for Research (The term 'Specificity Level' in some columns is explained in 7.2.2.)

留学生用語彙レベル Word Level for International Students

留学生用語彙ランク Word Ranking for International Students
一般語彙レベル Word Level for General Learners
一般語彙ランク Word Ranking for General Learners
書きことば語彙レベル Word Level for Written Japanese
書きことば重要度ランク（想定既知語彙を除く） U Ranking for Written Japanese excluding Assumed Known Words
旧日本語能力試験出題基準レベル Former JLPT Level
人文・芸術領域特徴度レベル Specificity Level in Humanities and Arts (Ha)
社会科学領域特徴度レベル Specificity Level in Social Sciences (Ss)
自然科学（理学・工学系）領域特徴度レベル Specificity Level in Technological Natural Sciences (Ss)
自然科学（生物・医学系）領域特徴度レベル Specificity Level in Bio-Medical Natural Sciences (Bn)
文芸特徴語候補 Possible Literary Keywords
語彙階層ラベル Word Tier Label
見出し語彙素 Lexeme
標準的（新聞）表記 Standard (Newspaper) Orthography
標準的読み方（カタカナ） Standard Reading (Katakana)
品詞 Part of Speech
語種 Word-Origin Type
雑誌表記 Magazine Forms
使用度数 Frequency
修正済み使用度数（総延べ語数 32656221 語中） Corrected Frequency (Out of Total Token 32656221)
修正度数 Frequency for Correction
10 分野 100 万語あたり使用頻度(Fw) Standardized Freq/million in 10 Written Domains (Fw)
(Fw)累積テキストカバー率（想定既知語彙分を含む） Fw Cumulative Text Coverage including Assumed Known Words
8 分野 100 万語あたり使用頻度 Standardized Freq/million in 8 Domains
3 大分野 100 万語あたり使用頻度平均(Fr1) Freq revised ver 1/million in 3 big domains (Fr1)
修正(Fr1)累積テキストカバー率（想定既知語彙分を含む） Fr1 Cumulative Text Coverage including Assumed Known Words
LW、OC2 分野 100 万語あたり使用頻度平均(Fr2) Standardized Freq/million in LW+OC (Fr2)
分散度 D
書きことば使用度係数(Uw) Uw (Usage Coefficient) for Written Japanese
修正使用度係数(Ur1) Ur1 (Usage Coefficient revised ver 1)
修正使用度係数(Ur2) Ur2 (Usage Coefficient revised ver 2)
使用範囲 Range
書きことば重要度順位（想定既知語彙を含む） U Ranking for Written Japanese including Assumed Known Words
使用頻度順位 Freq Ranking
分散度順位 D Ranking
歪度 Skewness
歪度（絶対値） Skewness (Absolute Value)
尖度 Kurtosis

下位コーパス順位平均（想定既知語彙除く） Average Sub-frequency Ranking excluding Assumed Known Words
語彙素文字数 # of Characters
下位コーパス使用頻度（文芸創作） Sub-frequency in LW
100 万語あたり使用頻度（文芸創作） LW Freq per Million
使用頻度ランク（文芸創作）（想定既知語彙除く） LW Freq Ranking excluding Assumed Known Words
下位コーパス使用頻度（言語・哲学）Sub-frequency in LP
100 万語あたり使用頻度（言語・哲学） LP Freq per Million
使用頻度ランク（言語・哲学）（想定既知語彙除く） LP Freq Ranking excluding Assumed Known Words
下位コーパス使用頻度（歴史・民俗）Sub-frequency in HE
100 万語あたり使用頻度（歴史・民俗） HE Freq per Million
使用頻度ランク（歴史・民俗）（想定既知語彙除く） HE Freq Ranking excluding Assumed Known Words
下位コーパス使用頻度（芸術、その他の人文科学）Sub-frequency in AH
100 万語あたり使用頻度（芸術、その他の人文科学） AH Freq per Million
使用頻度ランク（芸術・その他の人文科学）（想定既知語彙除く） AH Freq Ranking excluding Assumed Known Words
下位コーパス使用頻度（政治・法律）Sub-frequency in PL
100 万語あたり使用頻度（政治・法律） PL Freq per Million
使用頻度ランク（政治・法律）（想定既知語彙除く） PL Freq Ranking excluding Assumed Known Words
下位コーパス使用頻度（経済・商業）Sub-frequency in EC
100 万語あたり使用頻度（経済・商業） EC Freq per Million
使用頻度ランク（経済・商業）（想定既知語彙除く） EC Freq Ranking excluding Assumed Known Words
下位コーパス使用頻度（社会・教育、その他の社会科学）Sub-frequency in SE
100 万語あたり使用頻度（社会・教育、その他の社会科学） SE Freq per Million
使用頻度ランク（社会・教育、その他の社会科学）（想定既知語彙除く） SE Freq Ranking excluding Assumed Known Words
下位コーパス使用頻度（科学・技術）Sub-frequency in ST
100 万語あたり使用頻度（科学・技術） ST Freq per Million
使用頻度ランク（科学・技術）（想定既知語彙除く） ST Freq Ranking excluding Assumed Known Words
下位コーパス使用頻度（生物・医学・生活科学）Sub-frequency in BM
100 万語あたり使用頻度（生物・医学・生活科学） BM Freq per Million
使用頻度ランク（生物・医学・生活科学）（想定既知語彙除く） BM Freq Ranking excluding Assumed Known Words
下位コーパス使用頻度（インターネット Q&A フォーラム）Sub-frequency in IF
100 万語あたり使用頻度（インターネット Q&A フォーラム） IF Freq per Million
使用頻度ランク（インターネット Q&A フォーラム）（想定既知語彙除く） IF Freq Ranking excluding Assumed Known Words
人文・芸術領域対数尤度比（平均以上、1.0 以上平均未満） LLR in Arts & Humanities (M or above, less than M and more than 1.0)
社会科学領域対数尤度比（平均以上、1.0 以上平均未満） LLR in Social Sciences (M or above, less than M and more than 1.0)
自然科学（理学・工学系）領域対数尤度比（平均以上、1.0 以上平均未満） LLR in Technological Natural Sciences (M or above, less than M and more than 1.0)

自然科学（生物・医学系）領域対数尤度比（平均以上、1.0 以上平均未満） LLR in Bio-Medical Natural Sciences (M or above, less than M and more than 1.0)
形態素解析・品詞に関するメモ Notes on Morphological Analysing & POS
書字形（例） Orthographic Form Example
発音形（例） Phonological Form Example
語彙素読み Reading of Lexeme
活用型 Conjugation Type
活用形（例） Conjugated Form Example
語形 Word Form
ID
ホームポジション並べ替え用 ID ID for Sorting by the Original Order

Assumed Known Words are placed at the top of the list as they should be counted as known when computing the cumulative text coverage. Within each category of the Assumed Known Words and the general words (words other than Assumed Known Words), all the listed words are sorted by the criteria shown below.

1) Word Level for International Students (Ascending)
2) The former Japanese Language Proficiency Test (F-JLPT) word level (Descending) and Usage Coefficient (*Uw*/*Ur1*/*Ur2*) as described in Table 3-33 (Descending)
3) Frequency (*Fw*/*Fr1*/*Fr2*) (Descending)
4) Dispersion (*D*) (Descending)
5) Lexeme (Ascending)

* Words in “IS/GL/W_01K” and “IS/GL/W_01K+” are sorted together by 2) - 5).

3.5 Validation of the word lists

3.5.1 Methods

There are mainly two types of methods for validating a word list. One is to check the text coverage (Coxhead, 2000; Coxhead & Hirsh, 2007; Nation, 2006, 2011) and the other is to check the reaction time on psychological experiments (Gries, 2010; New, Brysbaert,

Veronis, & Pallier, 2007). In this study, I use the former way, because the latter one will not be sensitive enough to detect the differences which this study needs to check, as the reaction time involves many factors other than frequency, such as visual and semantic complexity.

In Chapter 4, the general lexical features of written Japanese will be explored by analysing the database (VDRJ). If there are no unexplainable results there, it can also be the part of validation of the database. Besides that, the questions shown below are examined in this section.

1) Does the Word Ranking for International Students (WIS) and the Word Ranking for General Learners (WGL) provide higher text coverage than the existing word lists such as the former Japanese Language Proficiency Test (F-JLPT) word list (Japan Foundation & Association of International Education, Japan, 2002)?

It is also necessary to compare the word rankings (WWJ, WIS and WGL) which should provide different levels of text coverage depending on the genre or media, and to examine if the differences between them are as expected. Specifically, the questions here are as follows.

2) Does WIS provide higher text coverage for academic texts than WGL?
3) Does WGL provide higher text coverage for non-academic texts than WIS?
4) Does WGL provide higher text coverage for daily conversation texts than the Word Ranking for Written Japanese (WWJ) at all levels?
5) Does WIS provide higher text coverage for daily conversation texts than WWJ at the basic level?

The test corpora are shown below.

JS-NS: J-STAGE (Japan Science & Technology Information Aggregator) academic journal article texts in natural sciences (e.g. electricity, civil engineering, environmental studies, physical education, health and sports science). 2.18 million running words from seven types of academic journals downloaded from J-STAGE at http://www.jstage.jst.go.jp/browse/-char/ja.

MTT-NS: Meidai Technical Texts in Natural Sciences. 0.08 million running words from the six volumes of natural science model lecture texts out of the nine volumes of " Technical Lecture Japanese for International Students" edited by the members of Nagoya University (Meidai)[62].

TB: Text Bank in Social Sciences for Intermediate and Advanced Learners of Japanese. 0.19 million running words from the text bank.

UYN: Utiyama Yomiuri Newspaper Corpus. 5.68 million running words from the Yomiuri newspaper articles published from 1989 to 2001. The Japanese data from the Japanese-English News Article Alignment Data (JENAAD) (Utiyama & Isahara, 2003).

UPC: Utiyama Parallel Corpus. 2.30 million running words from literary works including novels, stories and essays. The Japanese data from the English-Japanese translation alignment data (Utiyama & Takahashi, 2003). Downloaded from http://www2.nict.go.jp/x/x161/members/mutiyama/align/index.html on 16 November 2010.

MC: Meidai Conversation Corpus: 1.13 million running words from various types of pair or group conversation at cafés, schools, homes or other places. Compiled by the

[62] Meidai Technical Texts are transcribed from spoken planned model lecture without onsite audience but for recording; therefore, the lectures seem to be given based on written texts as they contain few fillers and other features of spoken language. Therefore, these texts have the features of written texts, though they are lecture texts.

members of Nagoya University (Meidai). Downloaded from http://dbms.ninjal.ac.jp/nknet/ndata/nuc/ on 10 December 2010.

To check the text coverage, the software tool AntWordProfiler (Anthony, 2009) was used. To compare the coverage of the F-JLPT word lists and the other word rankings, the same number of words corresponding to each level of the F-JLPT are compiled into a baseword file (4,589 words from Level 4 to 2, and 7,388 words from Level 4 to 1). For example, the baseword file 'WIS_L2' are composed of the highest ranked words beyond the F-JLPT Level 4 & 3 (WIS, WGL share the F-JLPT Level 3 & 4 lists at the top of the lists), and has the same number of words as the F-JLPT Level 2. To make an accurate comparison, proper nouns and function words (particles and auxiliary verbs[63]) are excluded as most of them are excluded from F-JLPT word lists. For other purposes, baseword files each of which is made up of one thousand words are created up to the 20,000 word level (01K-20K) based on each word ranking of WWJ, WIS and WGL. Beyond the level, all the words are put in a baseword file named 21K+. All the Assumed Known Words such as proper nouns and hesitations are put in the separate list named AKW.

3.5.2 Results and Discussion

For the first question 1) "Does the Word Ranking for International Students (WIS) and the Word Ranking for General Learners (WGL) provide higher text coverage than existing word lists such as the former Japanese Language Proficiency Test (F-JLPT) word list (Japan Foundation & Association of International Education, Japan, 2002)?", the answer is yes as shown in Table 3-35.

In Table 3-35, the gap figures show the superiority of WIS and WGL to F-JLPT lists. The gaps are larger in newspapers and academic texts than in other types of texts on average.

[63] So-called 'joshi' 助詞 and 'jodoushi' 助動詞 in Japanese.

Both WIS and WGL outperform the F-JLPT word lists which have been the most widely used lists by learners and teachers of Japanese. F-JLPT is not designed for university admission but for 'general' Japanese; however, it has also been used for admission purposes for a long time as there was no other reliable exam at the time F-JLPT started. The experts who developed the word lists seemed to expect that F-JLPT would be used for admission. Consequently, the Level 1 and 2 lists, which serve for intermediate and advanced learners, seem to have an inclination towards academic vocabulary, while the Level 3 and 4 lists, which serve for elementary learners, contain basic vocabulary for daily conversation. Even so, interestingly, both WIS and WGL provide higher text coverage in all types of texts than F-JLPT. This tells us that the word rankings based on adjusted frequency data would provide better coverage than subjectively-selected word lists in general. Of course, the factors to order words are not only frequency; however, the gap is considerably large at 1.5% or more between WIS/WGL and F-JLPT (Table 3-35). More than one thousand words are needed to cover 1% beyond 92% coverage at the 5,000 word (05K) level or above in BCCWJ. The current JLPT word lists are not published; however, the WIS and WGL lists will be more similar to the current JLPT lists than the F-JLPT lists.

Table 3-35 Text Coverage (Percentage) in Different Genres by WIS, WGL and F-JLPT

Test Corpus Code	JS-NS	MTT-NS	TB	UYN	UPC	MC
Genre	Technical (Natural Sciences)	Academic (Natural Sciences)	Academic (Social Sciences)	Newspaper	Literary Works	Converation
Gap (WIS - 'F-JLPT')	**4.26**	**2.27**	**5.38**	**7.04**	**2.06**	**1.46**
Gap (WGL - 'F-JLPT')	**3.45**	**1.77**	**4.54**	**5.90**	**2.09**	**1.70**
WIS Level 4 to 2 (4,589 words)	79.61	83.75	91.14	87.40	88.34	91.04
WGL Level 4 to 2 (4,589 words)	78.80	83.25	90.30	86.26	88.37	91.28
F-JLPT Level 4 to 2 (4,589 words)	75.35	81.48	85.76	80.36	86.28	89.58
Gap (WIS - 'F-JLPT')	**2.63**	**0.65**	**2.16**	**3.29**	**1.57**	**1.58**
Gap (WGL - 'F-JLPT')	**1.74**	**0.25**	**1.62**	**2.60**	**1.60**	**1.76**
WIS Level 4 to 1 (7,388 words)	83.44	87.18	94.06	91.35	91.00	92.75
WGL Level 4 to 1 (7,388 words)	82.55	86.78	93.52	90.66	91.03	92.93
F-JLPT Level 4 to 1 (7,388 words)	**80.81**	**86.53**	**91.90**	**88.06**	**89.43**	**91.17**

* WIS: The Word Ranking for International Students
* WGL: The Word Ranking for General Learners
* F-JLPT: The former Japanese Language Proficiency Test word list (Level 4 is the most basic and Level 1 is the most advanced.)
* Bold figures are explained in the thesis.

Tables from 3-36 to 3-41 show that the word rankings are basically valid as text coverage for each 1,000 word level gradually decreases for the word frequency levels in all the cases shown in the tables.

For the question 2) "Does WIS provide higher text coverage for academic texts than WGL?", the answer is yes as shown in Table 3-36 and 3-38. The gap figures show which ranking performs better at the level. (i.e. Where *A*-*B* is positive, *A* performs better at the level.) The cumulative text coverage by WIS is higher than the one by WGL at all levels up to the 20,000 word (20K) level in both Table 3-36, 3-37 (natural science texts) and 3-38 (social science texts). As shown in Table 3-39, WIS and WWJ also outperform WGL in newspaper texts whose result is similar to academic texts. In Tables 3-36 to 3-39, at the 02K level, WIS provides much higher coverage than WGL by 4.08, 2.51, 4.60 and 4.52% respectively. As WIS and WGL share the same word rankings up to the middle of the 02K level, the gaps mean that some words are frequently used in science and newspaper texts beyond the shared words at the 01K-02K levels.

(From here down blank.)

Table 3-36 Text Coverage of JS-NS (Technical, Natural Sciences) at Each Word Level by WIS, WGL and WWJ

	WIS		WGL		WWJ		Gap (WIS-WGL)		Gap (WIS-WWJ)	
LEVEL LIST	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)
AKW	0.51	0.51	0.51	0.51	0.51	0.51	0.00	0.00	0.00	0.00
01K	55.07	55.58	55.07	55.58	66.22	66.73	0.00	0.00	**-11.15**	-11.15
02K	16.13	71.71	12.05	67.63	6.18	72.91	**4.08**	**4.08**	**9.95**	-1.20
03K	4.67	76.38	8.11	75.74	3.09	76.00	-3.44	**0.64**	1.58	0.38
04K	2.97	79.35	2.63	78.37	3.61	79.61	0.34	**0.98**	-0.64	-0.26
05K	1.67	81.02	1.85	80.23	1.60	81.21	-0.18	**0.79**	0.07	-0.19
06K	1.24	82.26	1.52	81.74	1.13	82.34	-0.27	**0.52**	0.11	-0.08
07K	1.38	83.64	1.21	82.95	1.34	83.68	0.17	**0.69**	0.04	-0.04
08K	1.07	84.71	0.91	83.86	1.07	84.75	0.17	**0.85**	0.00	-0.04
09K	0.83	85.54	1.25	85.11	0.80	85.55	-0.42	**0.44**	0.03	-0.01
10K	0.60	86.14	0.63	85.73	0.59	86.14	-0.03	**0.41**	0.01	0.00
11K	0.45	86.59	0.50	86.23	0.46	86.60	-0.05	**0.36**	-0.01	-0.01
12K	0.47	87.06	0.67	86.90	0.46	87.06	-0.21	**0.16**	0.01	0.00
13K	0.38	87.44	0.30	87.20	0.38	87.44	0.08	**0.24**	0.00	0.00
14K	0.54	87.98	0.35	87.55	0.54	87.98	0.19	**0.43**	0.00	0.00
15K	0.33	88.30	0.37	87.92	0.33	88.31	-0.04	**0.38**	0.00	-0.01
16K	0.27	88.57	0.32	88.23	0.27	88.58	-0.04	**0.34**	0.00	-0.01
17K	0.22	88.79	0.31	88.54	0.22	88.80	-0.09	**0.25**	0.00	-0.01
18K	0.19	88.99	0.17	88.71	0.19	88.99	0.02	**0.27**	0.00	0.00
19K	0.21	89.20	0.34	89.06	0.21	89.20	-0.13	**0.14**	0.00	0.00
20K	0.18	89.38	0.16	89.22	0.18	89.38	0.01	**0.15**	0.00	0.00
21K+	3.70	93.08	3.86	93.08	3.70	93.08	-0.15	0.00	0.00	0.00
Not in the Lists	6.92	100.00	6.92	100.00	6.92	100.00	0.00	0.00	0.00	0.00

* JS-NS: J-STAGE technical journal article texts in natural sciences (total token: 2,180,796)
* WIS: The Word Ranking for International Students
* WGL: The Word Ranking for General Learners
* WWJ: The Word Ranking in Written Japanese
* TC: Text coverage Cum. TC: Cumulative text coverage
* AKW: Assumed Known Words which are mostly proper nouns not requiring previous learning.
* Bold figures are explained in the thesis.

Table 3-37 Text Coverage of MTT-NS (Academic, Natural Sciences) at Each Word Level by WIS, WGL and WWJ

	WIS		WGL		WWJ		Gap (WIS-WGL)		Gap (WIS-WWJ)	
LEVEL LIST	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)
AKW	0.54	0.54	0.54	0.54	0.54	0.54	0.00	0.00	0.00	0.00
01K	64.62	65.17	64.62	65.17	73.05	73.59	0.00	0.00	**-8.43**	-8.43
02K	11.83	77.00	9.32	74.49	5.06	78.66	**2.51**	**2.51**	**6.77**	-1.66
03K	3.86	80.86	5.92	80.41	2.58	81.24	-2.06	**0.45**	1.28	-0.38
04K	2.64	83.49	2.35	82.76	2.49	83.73	0.28	**0.73**	0.14	-0.24
05K	1.60	85.09	1.68	84.44	1.57	85.30	-0.08	**0.65**	0.03	-0.21
06K	1.24	86.34	1.37	85.81	1.12	86.43	-0.13	**0.53**	0.12	-0.09
07K	1.15	87.49	1.27	87.07	1.11	87.54	-0.11	**0.42**	0.04	-0.05
08K	1.13	88.62	0.75	87.82	1.11	88.65	0.38	**0.79**	0.02	-0.03
09K	0.63	89.25	0.97	88.80	0.60	89.25	-0.34	**0.46**	0.03	0.00
10K	0.64	89.89	0.66	89.45	0.64	89.89	-0.02	**0.43**	0.00	0.00
11K	0.44	90.33	0.62	90.07	0.43	90.33	-0.18	**0.26**	0.00	0.00
12K	0.69	91.02	0.59	90.66	0.69	91.02	0.11	**0.36**	0.00	0.00
13K	0.34	91.36	0.39	91.05	0.37	91.39	-0.05	**0.31**	-0.03	-0.03
14K	0.29	91.65	0.32	91.36	0.26	91.64	-0.03	**0.28**	0.03	0.00
15K	0.33	91.98	0.25	91.61	0.33	91.97	0.08	**0.36**	0.00	0.00
16K	0.16	92.13	0.25	91.86	0.16	92.13	-0.09	**0.27**	0.00	0.00
17K	0.23	92.36	0.36	92.22	0.25	92.39	-0.13	**0.14**	-0.03	-0.02
18K	0.28	92.64	0.27	92.49	0.26	92.64	0.01	**0.16**	0.03	0.00
19K	0.16	92.80	0.10	92.59	0.16	92.80	0.05	**0.21**	0.00	0.00
20K	0.11	92.91	0.18	92.77	0.11	92.91	-0.07	**0.14**	0.00	0.00
21K+	3.76	96.67	3.90	96.67	3.76	96.67	-0.14	0.00	0.00	0.00
Not in the Lists	3.33	100.00	3.33	100.00	3.33	100.00	0.00	0.00	0.00	0.00

* MTT: Meidai Technical Texts (total token: 88,549)
* WIS: The Word Ranking for International Students
* WGL: The Word Ranking for General Learners
* WWJ: The Word Ranking in Written Japanese
* TC: Text coverage Cum. TC: Cumulative text coverage
* AKW: Assumed Known Words which are mostly proper nouns not requiring previous learning.
* Bold figures are explained in the thesis.

Table 3-38 Text Coverage of TB (Academic, Social Sciences) at Each Word Level by WIS, WGL and WWJ

	WIS		WGL		WWJ		Gap (WIS-WGL)		Gap (WIS-WWJ)	
LEVEL LIST	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)
AKW	0.43	0.43	0.43	0.43	0.44	0.44	0.00	0.00	-0.01	-0.01
01K	66.05	66.49	66.05	66.49	78.07	78.51	0.00	0.00	**-12.01**	-12.02
02K	16.63	83.12	12.03	78.52	6.51	85.02	**4.60**	**4.60**	**10.12**	-1.89
03K	4.98	88.10	8.48	87.00	3.61	88.63	-3.50	**1.10**	1.37	-0.53
04K	2.90	91.00	2.91	89.91	2.70	91.33	-0.01	**1.09**	0.20	-0.33
05K	1.45	92.45	1.83	91.74	1.33	92.66	-0.38	**0.71**	0.12	-0.20
06K	1.13	93.59	1.23	92.97	1.02	93.67	-0.10	**0.61**	0.12	-0.08
07K	0.97	94.55	1.05	94.02	0.93	94.60	-0.08	**0.53**	0.04	-0.05
08K	0.67	95.23	0.66	94.68	0.65	95.25	0.01	**0.54**	0.02	-0.02
09K	0.62	95.85	0.54	95.23	0.61	95.86	0.08	**0.62**	0.01	-0.01
10K	0.41	96.26	0.54	95.76	0.41	96.27	-0.13	**0.50**	0.00	-0.01
11K	0.42	96.68	0.41	96.17	0.41	96.69	0.01	**0.50**	0.00	-0.01
12K	0.28	96.96	0.40	96.58	0.28	96.96	-0.13	**0.38**	0.00	-0.01
13K	0.29	97.25	0.28	96.86	0.29	97.25	0.01	**0.39**	0.00	0.00
14K	0.17	97.42	0.29	97.15	0.17	97.42	-0.12	**0.27**	0.00	-0.01
15K	0.20	97.62	0.16	97.31	0.20	97.62	0.04	**0.31**	0.00	0.00
16K	0.16	97.78	0.20	97.50	0.16	97.79	-0.03	**0.28**	0.00	0.00
17K	0.15	97.93	0.17	97.68	0.15	97.93	-0.03	**0.26**	0.00	0.00
18K	0.11	98.04	0.15	97.82	0.11	98.04	-0.04	**0.22**	0.00	0.00
19K	0.11	98.15	0.10	97.92	0.11	98.15	0.01	**0.23**	0.00	0.00
20K	0.08	98.24	0.14	98.06	0.08	98.24	-0.06	**0.17**	0.00	0.00
21K+	1.06	99.29	1.23	99.29	1.06	99.29	-0.17	0.00	0.00	0.00
Not in the Lists	0.71	100.00	0.71	100.00	0.71	100.00	0.00	0.00	0.00	0.00

* TB: Text Bank in Social Sciences for Intermediate and Advanced Learners of Japanese (total token: 186,768)
* WIS: The Word Ranking for International Students
* WGL: The Word Ranking for General Learners
* WWJ: The Word Ranking in Written Japanese
* TC: Text coverage Cum. TC: Cumulative text coverage
* AKW: Assumed Known Words which are mostly proper nouns not requiring previous learning.
* Bold figures are explained in the thesis.

Table 3-39 Text Coverage of UYN (Newspapers) at Each Word Level by WIS, WGL and WWJ

	WIS		WGL		WWJ		Gap (WIS-WGL)		Gap (WIS-WWJ)	
LEVEL LIST	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)
AKW	1.29	1.29	1.29	1.29	1.30	1.30	0.00	0.00	0.00	0.00
01K	59.17	60.46	59.17	60.46	70.91	72.21	0.00	0.00	**-11.74**	-11.75
02K	16.82	77.28	12.30	72.76	8.45	80.66	**4.51**	**4.51**	**8.37**	-3.38
03K	7.08	84.36	10.21	82.98	4.79	85.45	-3.13	**1.39**	2.29	-1.08
04K	3.58	87.94	3.91	86.89	3.03	88.47	-0.33	**1.05**	0.55	-0.53
05K	2.43	90.37	2.30	89.19	2.18	90.65	0.12	**1.18**	0.24	-0.29
06K	1.67	92.03	1.97	91.16	1.52	92.18	-0.31	**0.87**	0.14	-0.14
07K	1.33	93.36	1.40	92.56	1.21	93.39	-0.07	**0.80**	0.12	-0.02
08K	0.78	94.15	0.94	93.50	0.75	94.13	-0.16	**0.64**	0.03	0.01
09K	0.75	94.89	0.78	94.28	0.76	94.89	-0.03	**0.61**	-0.01	0.01
10K	0.63	95.52	0.63	94.91	0.64	95.53	0.00	**0.62**	-0.01	-0.01
11K	0.53	96.06	0.54	95.44	0.52	96.06	0.00	**0.61**	0.01	0.00
12K	0.39	96.45	0.52	95.96	0.40	96.45	-0.12	**0.49**	-0.01	-0.01
13K	0.38	96.83	0.42	96.38	0.38	96.83	-0.04	**0.44**	0.00	-0.01
14K	0.30	97.12	0.35	96.74	0.30	97.13	-0.06	**0.39**	0.00	-0.01
15K	0.29	97.41	0.27	97.00	0.29	97.42	0.02	**0.41**	0.00	-0.01
16K	0.22	97.63	0.28	97.29	0.21	97.63	-0.07	**0.34**	0.01	0.00
17K	0.19	97.82	0.22	97.50	0.19	97.82	-0.02	**0.32**	0.00	0.00
18K	0.21	98.03	0.20	97.71	0.21	98.04	0.01	**0.33**	0.00	0.00
19K	0.15	98.18	0.17	97.88	0.15	98.18	-0.02	**0.31**	0.00	0.00
20K	0.13	98.31	0.19	98.07	0.13	98.32	-0.06	**0.25**	0.00	0.00
21K+	1.46	99.77	1.70	99.77	1.46	99.77	-0.25	0.00	0.00	0.00
Not in the Lists	0.23	100.00	0.23	100.00	0.23	100.00	0.00	0.00	0.00	0.00

* UYN: Utiyama Yomiuri Newspaper corpus (total token: 5,675,357)
* WIS: The Word Ranking for International Students
* WGL: The Word Ranking for General Learners
* WWJ: The Word Ranking in Written Japanese
* TC: Text coverage Cum. TC: Cumulative text coverage
* AKW: Assumed Known Words which are mostly proper nouns not requiring previous learning.
* Bold figures are explained in the thesis.

For the question 3) "Does WGL provide higher text coverage for non-academic texts than WIS?", the answer is basically yes but no in the 02K level as shown in Table 3-40. At the 02K level (from 1001 to 2000), WIS performs slightly better than WGL by 0.18% in the literary texts including essays, but WGL outperforms WIS at all the other levels in cumulative text coverage. (i.e. The negative figures in 'Gap (WIS-WGL)' mean that WGL provides higher text coverage than WIS.) For conversation texts, as shown in Table 3-41, WGL totally outperforms WIS.

Table 3-40 Text Coverage of UPC (Literary Works) at Each Word Level by WIS, WGL and WWJ

	WIS		WGL		WWJ		Gap (WIS-WGL)		**Gap (WWJ-WGL)**	
LEVEL LIST	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)
AKW	1.33	1.33	1.33	1.33	1.39	1.39	0.00	0.00	0.06	0.06
01K	74.89	76.22	74.89	76.22	78.58	79.96	0.00	0.00	**3.69**	3.74
02K	7.99	84.22	7.81	84.04	4.74	84.70	**0.18**	0.18	**-3.08**	0.67
03K	2.98	87.20	3.28	87.31	2.62	87.33	-0.30	**-0.12**	-0.65	0.01
04K	1.93	89.13	1.83	89.14	1.88	89.21	0.11	**-0.01**	0.06	0.07
05K	1.30	90.43	1.33	90.47	1.28	90.49	-0.03	**-0.04**	-0.04	0.03
06K	1.02	91.45	1.06	91.53	0.98	91.47	-0.04	**-0.08**	-0.08	-0.06
07K	0.89	92.34	0.81	92.33	0.89	92.36	0.08	**0.00**	0.09	0.03
08K	0.68	93.02	0.72	93.06	0.66	93.03	-0.04	**-0.04**	-0.06	**-0.03**
09K	0.54	93.56	0.54	93.60	0.54	93.57	0.00	**-0.04**	0.00	**-0.03**
10K	0.51	94.07	0.47	94.07	0.50	94.07	0.03	**-0.01**	0.02	**0.00**
11K	0.37	94.44	0.41	94.48	0.37	94.44	-0.03	**-0.04**	-0.03	**-0.04**
12K	0.33	94.77	0.32	94.80	0.33	94.77	0.01	**-0.03**	0.00	**-0.03**
13K	0.30	95.07	0.30	95.10	0.31	95.08	0.00	**-0.03**	0.01	**-0.02**
14K	0.26	95.33	0.28	95.38	0.26	95.33	-0.02	**-0.05**	-0.02	**-0.04**
15K	0.23	95.56	0.23	95.61	0.23	95.57	0.00	**-0.05**	0.00	**-0.04**
16K	0.23	95.79	0.23	95.84	0.22	95.79	0.00	**-0.05**	0.00	**-0.05**
17K	0.20	95.99	0.19	96.03	0.20	95.99	0.01	**-0.04**	0.01	**-0.04**
18K	0.19	96.18	0.20	96.22	0.19	96.18	-0.01	**-0.05**	-0.01	**-0.05**
19K	0.19	96.36	0.16	96.38	0.19	96.36	0.03	**-0.01**	0.03	**-0.01**
20K	0.13	96.50	0.14	96.52	0.13	96.50	0.00	**-0.02**	-0.01	**-0.02**
21K+	1.54	98.04	1.52	98.04	1.54	98.04	0.02	0.00	0.02	0.00
Not in the Lists	1.96	100.00	1.96	100.00	1.96	100.00	0.00	0.00	0.00	0.00

* UPC: Utiyama Parallel Corpus (total token: 2,102,178)
* WIS: The Word Ranking for International Students
* WGL: The Word Ranking for General Learners
* WWJ: The Word Ranking in Written Japanese
* TC: Text coverage Cum. TC: Cumulative text coverage
* AKW: Assumed Known Words which are mostly proper nouns not requiring previous learning.
* Bold figures are explained in the thesis.

For the questions 4) “Does WGL provide higher text coverage for daily conversation texts than the Word Ranking for Written Japanese (WWJ) at all levels?” and 5) “Does WIS provide higher text coverage for daily conversation texts than WWJ at the basic level?”, the answers are all yes as shown in Table 3-41. WGL provides higher cumulative text coverage than WWJ at all levels in the conversation corpus. WIS also performs better than WWJ at least up to the mid-frequency (beyond the top 2,000 words) level.

WWJ outperforms WIS and WGL in the other written test corpora, mainly because WWJ provides much higher text coverage at the 01K level (See ‘Gap (WIS-WWJ)’ at the

01K level in Tables 3-36 to 3-39). This means that the 01K level in WWJ contains some words which are much more frequently used in (formal) written texts than in (informal) oral texts.

Table 3-41 Text Coverage of MC (Conversation) at Each Word Level by WIS, WGL and WWJ

	WIS		WGL		WWJ		Gap (WIS-WGL)		**Gap (WWJ-WGL)**		**Gap (WIS-WWJ)**	
LEVEL LIST	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)
AKW	1.72	1.72	1.72	1.72	2.06	2.06	0.00	0.00	0.34	0.34	-0.34	-0.34
01K	81.72	83.44	81.72	83.44	79.31	81.37	0.00	0.00	-2.41	**-2.08**	2.41	**2.08**
02K	6.95	90.39	7.56	91.00	8.53	89.90	-0.62	**-0.62**	0.97	**-1.11**	-1.58	**0.49**
03K	1.74	92.13	1.54	92.55	2.01	91.90	0.20	**-0.42**	0.46	**-0.64**	-0.27	**0.22**
04K	1.36	93.49	1.24	93.79	1.44	93.34	0.12	**-0.30**	0.20	**-0.45**	-0.07	**0.15**
05K	0.77	94.27	0.90	94.69	0.84	94.18	-0.13	**-0.42**	-0.06	**-0.51**	-0.06	**0.09**
06K	0.65	94.92	1.04	95.73	0.70	94.88	-0.39	**-0.81**	-0.33	**-0.84**	-0.05	**0.03**
07K	0.96	95.88	0.33	96.06	0.95	95.84	0.63	**-0.18**	0.62	**-0.22**	0.01	**0.04**
08K	0.29	96.17	0.29	96.34	0.29	96.13	0.00	**-0.18**	0.01	**-0.22**	0.00	**0.04**
09K	0.25	96.41	0.25	96.59	0.28	96.40	0.00	**-0.18**	0.03	**-0.19**	-0.03	**0.01**
10K	0.21	96.63	0.19	96.79	0.21	96.61	0.02	**-0.16**	0.01	**-0.17**	0.01	**0.01**
11K	0.15	96.78	0.17	96.95	0.15	96.76	-0.01	**-0.17**	-0.01	**-0.19**	0.00	**0.02**
12K	0.17	96.95	0.17	97.13	0.17	96.94	0.00	**-0.18**	0.00	**-0.19**	-0.01	**0.01**
13K	0.20	97.14	0.13	97.26	0.20	97.14	0.06	**-0.11**	0.07	**-0.12**	0.00	**0.01**
14K	0.13	97.27	0.12	97.38	0.13	97.26	0.01	**-0.11**	0.01	**-0.11**	0.00	**0.01**
15K	0.10	97.37	0.11	97.49	0.10	97.37	-0.01	**-0.11**	-0.01	**-0.12**	0.00	**0.01**
16K	0.09	97.47	0.09	97.57	0.10	97.46	0.01	**-0.11**	0.01	**-0.11**	-0.01	**0.00**
17K	0.07	97.54	0.08	97.65	0.07	97.53	-0.01	**-0.11**	-0.01	**-0.12**	0.00	**0.00**
18K	0.08	97.62	0.06	97.71	0.08	97.61	0.02	**-0.10**	0.02	**-0.10**	0.00	**0.00**
19K	0.06	97.68	0.05	97.77	0.07	97.68	0.01	**-0.09**	0.01	**-0.09**	0.00	**0.00**
20K	0.08	97.76	0.05	97.82	0.08	97.76	0.02	**-0.06**	0.02	**-0.06**	0.00	**0.00**
21K+	0.81	98.56	0.74	98.56	0.81	98.56	0.06	0.00	0.06	0.00	0.00	0.00
Not in the Lists	1.44	100.00	1.44	100.00	1.44	100.00	0.00	0.00	0.00	0.00	0.00	0.00

* MC: Meidai Converation Corpus (total token: 1,129,538)
* WIS: The Word Ranking for International Students
* WGL: The Word Ranking for General Learners
* WWJ: The Word Ranking in Written Japanese
* TC: Text coverage Cum. TC: Cumulative text coverage
* AKW: Assumed Known Words which are mostly proper nouns not requiring previous learning.
* Bold figures are explained in the thesis.

The previous sub-research-question discussed in 3.3.6 was SRQ1-3) "Are the most appropriate word ranking criteria different depending on the target learners such as general learners or international students? If yes, what are the good criteria for those different learner groups?" As expected in 3.3.6, WIS and WGL perform differently for different types of texts. As intended, WIS fits academic texts and newspapers better than WGL, while WGL fits conversation better than WIS. This means *Uw* is better for written texts while F-JLPT Level 3 and 4 is better for conversation than *Uw*. *Ur2* and *Ur1* also work

better than *Uw* for conversation as well. If we assume daily conversation is more important than written texts for elementary general learners and elementary international students, WIS and WGL are better than WWJ.

Also as expected, WGL is better than WIS for non-academic texts (literary texts including essays) as WGL provides higher cumulative text coverage than WIS at most levels except for 02K. This means that, at the 02K level, *Uw* works better than *Ur2* where only literary works and internet forum-sites are counted, while *Ur1*, where literary works, internet-forum sites are more weighted than *Uw*, performs better than *Uw* at 03K or above. (See Table 3-32 for the percentages weighted on each of the domains.) This may be because the lexical feature of literary texts, of course, is closer to the one of literary works while considerably different from the one of the internet-forum texts.

Contrary to conversation, as shown in Tables 3-36 to 3-41, WWJ outperforms WIS and WGL for all types of written text at least from the elementary to intermediate level. (WGL works better for literary texts at 08K or above ('Gap (WWJ-WGL)' in Table 3-40). In particular, WWJ provides much higher text coverage at the 01K level by 8 to 12% for academic texts and newspapers and by 3.69% for literary texts. If a learner only needs to learn written Japanese but does not need to learn daily conversation (e.g. a researcher of Japanese studies outside Japan), it is good to follow the WWJ ranking.

Table 3-42 shows what kinds of words have a large ranking gap between WIS, WGL or WWJ at different word levels. Just because of these types of words, different word rankings make sense for different purposes of learning.

Table 3-42 Sample Words with a Large Ranking Gap between WIS, WGL or WWJ (from 01K, 03K and 05K WIS Word Level)

*Sorted by "Ranking Gap (WIS-WGL)" at each level

Lexeme	Reading	English Translation	WIS Word Level	WIS Word Ranking	WGL Word Ranking	WWJ Word Ranking	Ranking Gap (WIS-WGL)	Ranking Gap (WIS-WWJ)	Ranking Gap (WGL-WWJ)	Word Origin
社会	shakai	society	01K	872	872	159	0	**713**	**713**	Chinese
研究	kenkyuu	research		956	956	252	0	**704**	**704**	Chinese
ラジカセ	rajikase	radio-cassette recorder		675	675	26,724	0	**-26,049**	**-26,049**	English
字引き	jibiki	dictionary		680	680	31,276	0	**-30,596**	**-30,596**	Mixed
ＯＫ	o-ke-	OK	03K	2,876	1,727	2,505	**1,149**	371	**-778**	English
出産	shussan	childbirth		2,981	1,881	2,634	**1,100**	347	**-753**	Chinese
当該	tougai	said/concerned		2,688	3,504	2,270	-816	418	**1,234**	Chinese
筆者	hissha	the present writer		2,866	3,704	2,494	**-838**	372	1,210	Chinese
前述	zenjutsu	aforementioned		2,995	3,955	2,650	**-960**	345	**1,305**	Chinese
ＰＣ	pi-shi-	PC	05K	4,936	3,094	4,768	**1,842**	168	**-1,674**	English
初心	shoshin	initial enthusiasm		4,782	3,206	4,610	**1,576**	172	**-1,404**	Chinese
言及	genkyuu	to refer/mention		4,554	6,227	4,373	**-1,673**	181	**1,854**	Chinese
図表	zuhyou	chart/figure		4,667	6,497	4,490	**-1,830**	177	**2,007**	Chinese

At the 01K level, words with a higher ranking in WWJ are basic formal words (e.g. 社会 'shakai' (society)) which are placed at Level 2 (intermediate) in F-JLPT. These words are more important in written communication than in daily conversation. Words with a lower ranking in WWJ are outdated words (e.g. ラジカセ 'rajikase' (radio-cassette recorder) and 字引き 'jibiki' (dictionary (lit.)) which are placed at Level 4 (elementary) in F-JLPT. F-JLPT word lists contain some outdated words as the lists were selected in the 1980s.

At the 03K and 05K levels, words with higher rankings in WGL (e.g. OK 'o^ke^' (OK) and 初心 'shoshin' (initial enthusiasm)) than in WIS or WWJ are the words often used in daily domains. The other words (e.g. 前述 'zenjutsu' (aforementioned) and 言及 'genkyuu' (to refer/mention) have a higher frequency in formal, written texts, particularly in academic texts, than conversation or non-academic texts.

In summary, word rankings WIS/WGL made from VDRJ will work better for learners and teachers than the F-JLPT word lists since the former provide higher text

coverage than the latter. The best order of learning words will be different depending on the purpose. WIS will fit for students or academics better than WGL while WGL will work better than WIS for learners who mainly have daily life needs. WWJ will only fit learners who do not need to learn daily conversation but only need to read (and write) Japanese.

3.5.3 Usefulness of the VDRJ

As part of the validation of VDRJ, usefulness is the most important criterion. As discussed in 2.5, there are various uses of a vocabulary database and word lists derived from it. Some usages of VDRJ adopted in this thesis are described below.

First, we can make various baseword lists for lexical profiling[64]. In this chapter, text coverage is checked with baseword files created from the database. After the word-segmentation is done on the target text, the text coverage by the basewords can be checked. Baseword files of WIS, WGL and WWJ are already introduced in this chapter. These are to be used in Chapter 4 and 6.

Second, we can create domain-specific word lists and make them as baseword lists. These lists are to be created and used in Chapter 7 and 8.

Third, learning materials can be assessed from lexical perspectives by checking the lexical profiling and other features of the words used in the material. An example of this approach will be shown in Chapter 8.

3.6 Remaining issues

There are some remaining issues with VDRJ. First, word-segmentation cannot be perfect. We have to use a morphological analyser with an electric dictionary for word-segmentation as there is no space between words in Japanese. The combination of MeCab (analyser) and UniDic (dictionary) was adopted for this research as it currently returns the

[64] Lexical profiling is checking "the percentage of words at different vocabulary frequency levels" which is the same as the Lexical Frequency Profiling (Laufer, 1994, p 23). See footnote1 in this chapter.

highest accuracy rate; however, the job was far from perfect as there are many errors remaining. Major errors of counting frequencies were corrected mainly within the top 20,000 word level. However, we still need to wait for the improvement of the software technology.

Second, multiword units[65] are not included in this database. It would be useful if high-frequency multiword units are included in the database. This is left for a future study.

Third, we have to check the frequency and other features of each individual character appearing in the word lists. The learning burden of Kanji is very heavy. The most efficient order for learning Kanji will basically be the frequency order; however, there can be some Kanji which have high frequency but do not appear in the high-frequency words. Conversely, there can also be some Kanji used in the high-frequency words which do not have a high frequency overall. There may be some discrepancy between the character frequency and the word frequency. This will be examined in Chapter 6.

Fourth, as related to the previous point, meanings of some Kanji compounds are easily inferred correctly if the compound is semantically transparent. Japanese has relatively many (semi-)transparent compounds which do not require previous learning to understand the meaning. If a Kanji is able to occur in many transparent compounds, the Kanji should be learned first even if it does not appear in high frequency words. Therefore, the order of learning words and characters is not a straightforward issue. This is also an issue of the unit of analysis. A word must be a more important unit than character in general; however, taking account of the learning burden, the compounding power and transparency of words should also be considered. This issue will be further explored in Chapter 6.

Last but not least, identifying Assumed Known Words is also a problem. As mentioned, some common proper nouns have a similar semantic feature to general nouns

[65] Multiword units are defined as "items which are treated a single word token, even though they are spelt as a sequence of orthographic words" (Leech et al., 2001, p 8).

requiring previous learning. Nevertheless, it is difficult to set a border between words requiring previous learning and ones not requiring previous learning. Also, many Chinese cognates do not require intentional second language learning for Chinese-background learners to understand the meaning. As Chinese learners make up a considerably high proportion in many courses all over the world, this is a practical curriculum issue in teaching Japanese as a second language. This issue is left for a future study.

3.7 Conclusion of Chapter 3

In this chapter, I claimed the necessity of new word lists based on a new vocabulary database, and then described how I created the Vocabulary Database for Reading Japanese (VDRJ) and the word lists derived from the database. VDRJ is the first Japanese vocabulary database made from large corpora composed of books and the internet-forum sites, as contains approximately 33 million running words in total.

In the process of creating the database, there were some questions to be solved in terms of the index for ranking words and methods for weighting sub-frequencies. As for the index, *U* was adopted for VDRJ. To meet different learner needs, weighted sub-frequencies were used to compute *Ur1*and *Ur2* for ordering words in the Word Ranking for International Students (WIS) and the Word Ranking for General Learners (WGL). *Uw*, which is the original *U*, was also used for WIS as well as the Word Ranking for Written Japanese (WWJ).

After creating the database, its validity was examined, and some remaining issues were mentioned. The main findings in this chapter are as follows.

1) The adjusted frequency measures of *U*, U_{DP} and SFI do not make a significant difference on overall rankings of words. Spearman's rank correlation coefficients between the adjusted frequency measures are very high at .98 or above overall.
2) *U* is more sensitive to the salience of frequency of a single domain than U_{DP} and SFI.

This feature is suitable for fixing the sampling bias as well as for excluding unevenly distributed words from the high-frequency range.

3) The result of the multidimensional scaling shows that the ten sub-sections in BCCWJ can be divided into three categories of the Internet Q&A forum sites (IF) and literary works (LW) and the other eight (AD). IF and LW vocabulary will fit the basic and daily-life needs better than AD, while AD contains more academic and formal words than the other two.
4) The word ranking by Juilland's *U* (WWJ) shows that BCCWJ has a formal and written nature.
5) The word rankings WIS/WGL made from VDRJ will work better for learners and teachers than the former Japanese Language Proficiency Test (F-JLPT) word lists since the WIS/WGL provides higher text coverage than the F-JLPT lists.
6) The best order of learning words will be different depending on the purpose. WIS will fit for students or academics better than WGL, while WGL will work better for conversation than WIS. WWJ will only fit learners who do not need to learn daily conversation but only need to read (and write) Japanese.

Chapter 4 Statistical features of Japanese vocabulary

4.1 Introduction

These fifty years, there have been various statistical analyses of the Japanese language with large scale studies mainly done by the National Institute for Japanese Language and Linguistics (NINJAL, formerly the National Institute for Japanese Language, or NLRI, the National Language Research Institute). However, some of them are too old or too biased, or the corpus for the research is too small to reflect current general Japanese.

More importantly, there have only been word frequency lists based on magazine and newspaper corpora (i.e. Amano & Kondo, 2000; NLRI, 1962, 2006) but no large book corpus or an internet site corpus. The features of the corpus which the frequency is based on should be taken into account but have often been ignored when discussing the 'general' features of Japanese. For example, the corpus for NLRI (1962) contains many advertisements in magazines which are expected to have more loanwords from European languages than other media; however, little attention has been paid to this. Therefore, word origins and media should be analysed at the same time. Also, there are few studies about these aspects across the frequency levels.

As for part of speech (POS), UniDic (Den, Yamada, Ogura, Koiso, & Ogiso, 2009), the dictionary used for word segmentation for this study, can identify many more types of POS, which enables us to analyse the data from new aspects. For instance, UniDic can distinguish seven types of suffix such as adjectival suffixes, verbal suffixes and so on.

There are many studies on the proportion of lemmas[66] by word origins; however, this can be explored in combination with other aspects. Also, there are many studies about Chinese cognates whose orthographic forms are the same or similar in Japanese (Agency for Cultural Affairs, 1978; Arakawa, 1979; Araya, 1983; Hida & Ro, 1987; Kin, 1987,

[66] The lemma here is a similar unit to the lexeme adopted for this study.

1990; Lu, 2000; Saito, 1988); however, as discussed in 2.3.2 and 3.3.4.3.2, there is still some room to explore such cognates as previous studies are different from this study in purpose, method, corpus size and so on.

Most of the studies on the proportion of words by word origins or POS are based on counts of lemmas but not tokens. It is anticipated that the number of word lemmas or lexemes will have a certain degree of correlation with the amount of learning burden. Nevertheless, the proportion of words based on the count of tokens is also important as it directly relates to the text coverage which contributes to comprehension of text.

In sum, the distribution of words by word origins and POS should also be cross-checked by media and genre as well as the whole, at different (adjusted) frequency levels, based on the counts of both lexemes and tokens. In this chapter, taking advantage of new technology which has enabled us to deal with large language data individually, I analyse and present some new findings about statistical features of Japanese.

The database for this study is VDRJ (Vocabulary Database for Reading Japanese). As described in Chapter 3, it is based on the book corpus (28 million running words) and the internet forum site corpus (5 million running words) from the Balanced Corpus of Contemporary Written Japanese 2009 monitor version, which has approximately 33 million running words in total. It would have fewer words for current events than newspapers and magazines. Given these conditions, I will present some new findings about the lexical features of Japanese by analysing VDRJ. Specific viewpoints are as follows.

Firstly, to clarify the nature of the corpus on which VDRJ is based, I will compare the text coverage and proportion of word origins between different media: books, internet-forum, magazines and newspapers, then between different genres. Specific words in magazines, newspapers and VDRJ will also be extracted to show each domain's features.

Secondly, the distribution of POS at different frequency levels or in different genres

will be presented based on the counts of both lexemes and tokens. The distribution patterns of verbal nouns and affixes in Japanese will also be discussed.

Thirdly, based on the distribution of POS, indices for informality and formality for judging register variations in Japanese will be explored. The indexicality order of POS and the informality order of genres will be cross-checked.

Fourthly, distribution of Chinese-origin words at different frequency levels will be presented based on the counts of both lexemes and tokens. The distribution of Chinese cognates and related types of words is further explored. As is widely known, more than half of the learners of Japanese in Japan are Chinese-background learners (CBLs). To estimate the learning burden, the first language effect cannot be ignored. Before measuring the effect by tests, it must be useful to clarify the distribution and estimate the effect.

The main research questions (MRQs) are repeated below.

MRQs) In what order should learners of Japanese as a second language learn words and characters in order to be able to read Japanese? How will the order vary according to the purpose of learning?

In this chapter, the order of vocabulary learning will not be directly addressed. Instead, the variability of lexical features will be explored in terms of media and genres together with the consideration of word origins and POS. This is in order to gain insights into how the learning order of words will vary depending on the purpose of learning as well as to depict a broader picture of Japanese vocabulary. Specific sub-questions will be presented in each section.

4.2 Difference between media and genres in terms of text coverage and word origins

The goal of this section is to clarify lexical differences among the media and genres.

There has been no comprehensive research on this topic in Japanese as there was previously no large corpus available to individual researchers. As mentioned in 2.3.2, NLRI (1962) has been cited as data for 'general' Japanese for a long time; however, it is merely based on a set of magazine data. It shows 60.5% of the magazine texts are covered by the most frequent 1,000 words, 70.0% by the top 2,000, and 81.7% by the top 5,000. These figures, which are much lower than English and other languages, are often cited as the evidence that Japanese language has more diverse vocabulary than other languages (e.g. Tamamura, 1984, p 101). However, there are also text coverage data from newspaper texts (NLRI, 1970, p 30) which show that the most frequent 1,000 words provide much higher coverage at 73.5%, the top 2,000 words cover 79.9%, and 5,000 words cover 87.6%. These figures are at a similar level to coverage in other languages. How the characteristics of the corpus affect the text coverage should be adequately examined.

The main characteristics examined here are on the three aspects shown below.

1) Lexical homogeneity (diversity)

2) Informality (formality)

3) Colloquiality

Lexical homogeneity is examined by text coverage at different frequency levels. The higher the coverage, the more homogeneous the vocabulary use. In other words, the lower the coverage, the more diverse the vocabulary use.

Informality is examined by the proportion of Japanese-origin words. Concurrently, formality can be checked by the proportion of Chinese-origin words. As is widely known, Chinese-origin words are generally used more for formal or academic discourse in Japanese while Japanese-origin words are used more for daily discourse. The distribution of word lexemes by word origin is also checked at different frequency levels.

Colloquiality is examined by the use of indexical colloquial form or category. The more the use of the form or category, the more colloquial the texts. Nishimura (2010) has

already investigated the issue, which will be cited to compare the data with other aspects and discuss the differences between media and genres. Colloquiality is expected to be correlated with informality; however, there can be some genres which are colloquial but formal as well as genres which are literary but informal. (The measurement for colloquiality will be further explored in 4.4.)

Media-specific words will also be extracted in order to explore what kind of words typify the media. The media compared here are books, internet-forum sites, magazines and newspapers. Specific sub-research-questions (SRQs) are shown below. (The SRQ number follows the previous chapter.)

SRQ 5) How differently does text coverage increase depending on media and genres as the level of frequency gets lower?

SRQ 6) How high are the proportions for different word origins in different media and genres and how do the proportions relate to the use of colloquial forms or categories which represent colloquiality?

SRQ 7) What are the media-specific words in magazines and newspapers compared with VDRJ?

After answering these questions, the overall features of the media will be discussed. This is also to support that VDRJ represents more general Japanese than existing frequency lists.

4.2.1 Method

Media texts

The specific media and genres compared are as follows.

1) Literary books (LW): Imaginative texts from the Balanced Contemporary Corpus of

Written Japanese (BCCWJ) (NINJAL, 2009). These texts correspond to LW texts classified and introduced in 3.3.2. All original text files have the name starting with LB or PB, which are sampled texts but do not include the best seller book corpus. All the books were published between 1986 and 2005. They make up approximately 8 million running words in total.

2) Non-literary books (academic domains = AD): The book texts (the files of which the name starts with LB or PB which are sampled texts but do not include the best seller book corpus) except for LW from BCCWJ. These texts correspond to the eight sub-sections of LP, HE, AH, PL, EC, SE, ST, BM in Table 3-4 and 3-5 in 3.3.2.These are the genres excluding LW and IF from the ten genres in the tables shown above. The texts also correspond to AD classified and introduced in 3.3.6. All the books were published between 1986 and 2005. They make up approximately 19 million running words in total.
3) Internet-forum sites (IF): Yahoo Chiebukuro texts (the files of which the name starts with OC) of BCCWJ. These texts correspond to IF texts classified and introduced in 3.3.2. All the questions and answers in the forum were posted between October, 2004 and October, 2005. They make up approximately 5 million running words in total.
4) Magazines: Texts from 70 types of monthly magazines published in 1994 (NLRI, 2006). They make up approximately 1.07 million running words in total.
5) Newspapers: Texts from the Asahi published between 1985 and 1998 (Amano & Kondo, 2000).

The book corpora of literary works (LW) (1) and non-literary books (academic domains = AD) (2) and the internet-forum corpus (IF) (3) are the corpora used to create VDRJ. For comparing media (but not genres), LW and AD are added together as the ‘books’. For some genre analyses, the eight sub-genres of AD are separately analysed.

Data analysis

In order to compare the text coverage, the tables to show the text coverage in different media and genres by 1,000 word level are created from 1,000 to 10,000 word levels. The graph for the text coverage in different media is also created up to the 40,000 lexeme level. To examine the virtual learning burden of vocabulary, the required numbers of words to attain the different levels of text coverage in different media and genres are also shown by adding the coverage by assumed known words which are mostly proper nouns not requiring previous learning to understand the meaning. For all these statistics, function words such as particles (助詞 'joshi') auxiliary verbs (助動詞 'jodoushi') are included. In Japanese counts, they are often excluded (e.g. NLRI, 1962); however, it is not a current practice in English studies. It would be better to include them to discuss the statistical features of Japanese in comparison with other languages.

A different approach was taken to compare the proportions of word origins. The data from Mogi, Yamaguchi, Maruyama, & Tanaka (2005) is cited for the proportions in magazines and newspapers. For literary works ('LW'), internet-forum sites ('IF') and the other eight sub-genres ('AD') in BCCWJ that VDRJ is made from, the proportions are calculated using the filtering function of VDRJ. As Mogi et. al. (2005) exclude signs, function words (articles 助詞 'joshi' and auxiliary verbs 助動詞 'jodoushi' which are all Japanese-origin), proper nouns, numerals and unknown words (words not in the baseword lists), the analysis here all follows the way. The distribution of word lexemes by word origin is also counted at different frequency levels.

To extract the media-specific words, it would be the best to use log-likelihood ratio or another statistical index; however, there are no magazine and newspaper corpora at hand but frequency lists which are made from differently-segmented corpora; therefore, the ranking gap between media is used to extract media-specific words. The idea is that words which have a greatly higher ranking in a target corpus than in other corpora must be

specific to the target corpus. Specifically, for the target media, words are filtered by frequency ranking at 3,000 or higher (smaller in number), since the rankings are less reliable in the low frequency range. For the other media rankings, words are filtered by the ranking gap at 4,000 or lower (greater in number). All the words are ordered by the frequency ranking in the target media, and then the data from the top levels are chosen to explore the lexical features of the media compared with VDRJ (Books and the internet-forum sites).

4.2.2 Results and discussion

Lexical homogeneity

Text coverage by different numbers of words in different media is shown in Table 4-1 and Graph 4-1. Literary (LW) and non-literary (AD) texts are added together as books.

First of all, as shown in Table 4-1 and Graph 4-1, the text coverage by the top 1,000 words in Japanese is not as low as generally thought if function words are included. The top 1,000 words in magazines and VDRJ provide 75.3 and 79.0% coverage respectively. The magazine coverage is 3- 6% lower than English in which the top 1,000 coverage are between 78% and 81% (Nation, 2006, p 79); however, the VDRJ coverage is at the same level, or even 1% higher than the BNC list by Nation (2006). In addition, the text coverage in magazine texts, which is cited in Tamamura (1984) as the general Japanese coverage, is lower than other media. Magazines have higher lexical diversity than other media so that they cannot represent Japanese in general. Internet-forum sites are much more homogeneous in vocabulary. Books are not as lexically diverse as magazines and newspapers, either.

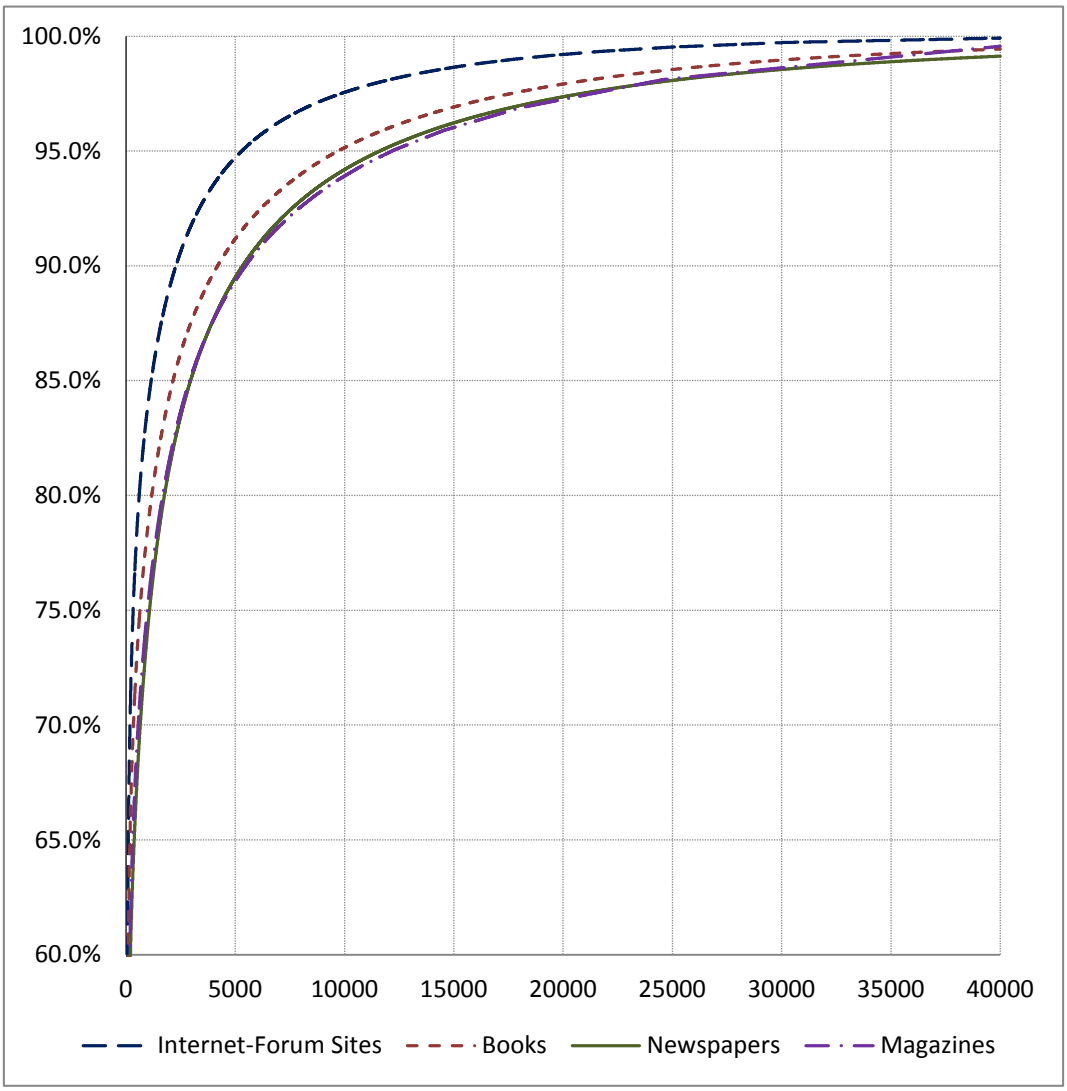
Table 4-1 Text Coverage (Percentage) by Different Numbers of Words in Different Media *Including function words. Rankings are all based on frequencies without any adjustment by dispersion.

Number of Words from the Top	AKW (*1)	1000	2000	3000	4000	**5000**	6000	7000	8000	9000	**10000**	11000	12000	13000	14000	**15000**	16000	17000	18000	19000	**20000**	**25000**	**30000**	**35000**	**40000**
Magazines (NLRI, 2006)	4.1	75.3	81.6	85.3	87.6	**89.3**	90.7	91.7	92.6	93.3	**93.9**	94.5	94.9	95.3	95.7	**96.0**	96.3	96.6	96.9	97.1	**97.3**	**98.2**	**98.6**	**99.1**	**99.6**
Newspapers (Amano & Kondo, 2000)	5.5	74.3	81.3	85.2	87.7	**89.5**	90.9	92.0	92.8	93.6	**94.2**	94.7	95.2	95.6	95.9	**96.2**	96.5	96.8	97.0	97.2	**97.4**	**98.1**	**98.6**	**98.9**	**99.1**
Books (NINJAL, 2009)	2.2	78.7	84.5	87.7	89.7	**91.2**	92.4	93.3	94.0	94.7	**95.2**	95.6	96.0	96.3	96.6	**96.9**	97.2	97.4	97.6	97.8	**97.9**	**98.5**	**99.0**	**99.2**	**99.4**
Internet-forum sites (IF) (NINJAL, 2009)	1.0	84.0	89.1	91.8	93.5	**94.7**	95.6	96.3	96.8	97.2	**97.6**	97.8	98.1	98.3	98.5	**98.6**	98.8	98.9	99.0	99.1	**99.2**	**99.5**	**99.7**	**99.8**	**99.9**
VDRJ (Books and IF) (NINJAL, 2009)	2.0	79.0	84.7	87.9	89.9	**91.4**	92.5	93.4	94.2	94.8	**95.3**	95.7	96.1	96.4	96.7	**97.0**	97.2	97.4	97.6	97.8	**98.0**	**98.6**	**99.0**	**99.2**	**99.4**

*1 AKW: Assumed Known Words, which include hesitations, proper names (excluding place names etc. with the ratio of 0.007% or more) and so on.

*2 Text coverage figures all include Assumed Known Words.

Graph 4-1 Text Coverage by Media * Including function words and Assumed Known Words



Required number of words to attain the different levels of text coverage in different media is shown in Table 4-2.

Table 4-2 Required Number of Words to Attain Different Levels of Text Coverage in Different Media (Assumed Known Words Included)

Number of Assumed Known Words/ Text Coverage	Assumed Known Words	60%	70%	80%	90%	91%	92%	93%	94%	**95%**	96%	97%	**98%**	99%
Magazines (NLRI, 2006)	14,728 (4.1%)	163	551	1,673	5,466	6,295	7,310	8,568	10,153	**12,164**	14,894	18,661	**23,989**	34,013
Newspapers (Amano & Kondo, 2000)	150,859 (5.5%)	195	648	1,752	5,332	6,103	7,038	8,194	9,666	**11,607**	14,257	18,112	**24,360**	37,112
Books (NINJAL, 2009)	28,307 (2.2%)	93	335	1,168	4,159	4,829	5,650	6,665	7,946	**9,625**	11,914	15,210	**20,399**	30,415
Internet-forum sites (IF) (NINJAL, 2009)	9,117 (1.0%)	59	177	599	2,279	2,646	3,091	3,642	4,351	**5,291**	6,578	8,483	**11,593**	17,777
VDRJ (Books and IF) (NINJAL, 2009)	30,683 (2.0%)	88	314	1,125	4,043	4,700	5,505	6,507	7,776	**9,446**	11,731	15,031	**20,256**	30,447

* Function words and Assumed Known Words (most proper names and hesitations etc.) are all included in the coverage.

* Assumed Known Words include hesitations, proper names (excluding place names etc. with the ratio of 0.007% or more) and so on.

* The coverage includes the Assumed Known Words, but the number of words does NOT include it. That is, the numbers shows the number of words which need to learn to attain the text coverage.

As shown in Table 4-2, the required number of words to attain certain levels of text coverage is also considerably different from media to media. Internet-forum sites only require 5,291 words to attain 95% coverage while books, newspapers and magazines require more than 9,000 words. Magazines require more words than books by 2,000 and more. This is probably because magazines have more technical words from a wide range of topics such as motor vehicles or classical music. Magazines and newspapers contain more than double the number of proper nouns which are not included in the required number of words. Magazines contain 4.1% assumed known words which are mostly proper nouns, newspapers contain 5.5%, while books only contain 2.2% and the internet-forum sites contain an even smaller number at 1.0%. Adding the proper nouns together, magazines and newspapers require more lexemes to gain text coverage than books and internet-forum sites.

As reviewed in 2.2.2, in studies about the relationship between the vocabulary coverage and the level of reading comprehension, required number of known words for 'adequate comprehension' vary between 95% and 98% (Hirsh & Nation, 1992; Hu & Nation, 2000; Komori et al., 2004; Laufer, 1989, 1992; Laufer & Ravenhorst-Kalovski, 2010; Schmitt et al., 2011). Setting 95% and 98% as tentative bench marks, Nation (2006) estimates that 4000-5000 word families (+proper nouns) are necessary to reach 95% coverage of a novel, and 8,000-9,000 word families (+proper nouns) are required to reach 98% coverage. Almost doubled the number of words (9,446 words for 95% coverage and 20,256 words for 98% coverage) is required to attain the same level of coverage in VDRJ. These numbers are surprisingly large; however, it cannot be instantly asserted that the learning burden of Japanese vocabulary is significantly heavier than that of English vocabulary, as the unit of counting for this study is different from English (See 3.3.3), and Japanese has more semantically transparent compounds whose meanings are easily inferred. This issue will be further investigated by computing the character frequencies in Chapter 5 and by matching with the character frequencies and word frequencies in Chapter 6.

If we compare the coverage of magazines and newspapers, magazines provide higher coverage up to the 4,000 word level in the high frequency band and approximately 20,000 word level and upwards in the low frequency band, while newspapers provide higher coverage in-between. Newspapers seem to require more lexemes of words in the basic expressions than magazines; however, they do not contain as many technical words as magazines. Besides, news articles have to be composed of generally understandable terms so that the mid-range vocabulary will be used more in news articles.

Tables 4-3 to 4-8 and Graph 4-2 are the comparisons between genres in cumulative text coverage and required number of words to attain the certain levels of text coverage. The abbreviations for genres used in this thesis are as follows.

AKW: Assumed Known Words, which include hesitations, proper names (excluding place names etc. with the proportion of 0.007% or more) and so on.
AD: Academic Domains which are the eight domains except for LW and IF in VDRJ.
LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.
Ha: Humanities and Arts, Ss: Social Sciences, Ns: Natural Sciences

Table 4-3 Cumulative Text Coverage (Percentage) in Different Genres in VDRJ

Genre Code	Number of Words from the Top	AKW	1,000	2,000	3,000	4,000	5,000	6,000	7,000	8,000	9,000	10,000
1-10	BCCWJ 2009	2.0	79.0	84.7	87.9	89.9	91.4	92.5	93.4	94.2	94.8	95.3
1-9	Books	2.2	78.7	84.5	87.7	89.7	91.2	92.4	93.3	94.0	94.7	95.2
2-9	AD	1.9	78.5	84.5	87.8	89.9	91.4	92.6	93.5	94.3	94.9	95.4
1	LW	3.0	82.1	87.0	89.7	91.4	92.7	93.7	94.5	95.1	95.7	96.1
2	Ah-LP	1.8	81.7	87.0	89.9	91.8	93.2	94.2	95.0	95.6	96.2	96.6
3	Ah-HE	3.4	78.2	84.0	87.2	89.4	91.0	92.3	93.2	94.0	94.7	95.3
4	Ah-AH	2.6	80.2	85.5	88.5	90.5	92.0	93.1	93.9	94.6	95.2	95.7
5	Ss-PL	1.5	82.0	88.5	91.7	93.7	95.0	95.9	96.6	97.2	97.6	98.0
6	Ss-EC	1.0	81.9	88.7	91.9	93.9	95.2	96.1	96.8	97.4	97.8	98.1
7	Ss-SE	1.1	81.6	87.7	90.9	92.8	94.2	95.2	95.9	96.5	97.0	97.4
8	Ns-ST	1.4	78.7	85.3	88.9	91.3	92.9	94.1	95.1	95.8	96.4	96.9
9	Ns-BM	1.2	79.0	85.3	88.8	91.1	92.7	93.8	94.8	95.5	96.1	96.6
10	IF	1.0	84.0	89.1	91.8	93.5	94.7	95.6	96.3	96.8	97.2	97.6

Table 4-4 Required Number of Words to Attain Different Levels of Text Coverage in Different Genres in VDRJ (Assumed Known Words Included)

Genre Code	Cumulative Text Coverage	AKW	60%	70%	80%	90%	91%	92%	93%	94%	**95%**	96%	97%	**98%**	99%
1-10	BCCWJ	2.0%	88	314	1,125	4,043	4,700	5,505	6,507	7,776	**9,446**	11,731	15,031	**20,256**	30,447
1-9	Books	2.2%	93	335	1,168	4,159	4,829	5,650	6,665	7,946	**9,625**	11,914	15,210	**20,399**	30,415
2-9	AD	1.9%	104	369	1,186	4,060	4,701	5,476	6,431	7,641	**9,222**	11,415	14,610	**19,722**	29,889
1	LW	3.0%	61	194	751	3,167	3,722	4,409	5,272	6,365	**7,799**	9,704	12,348	**16,352**	23,510
2	Ah-LP	1.8%	75	243	816	3,028	3,518	4,113	4,856	5,796	**7,016**	8,678	11,020	**14,643**	21,319
3	Ah-HE	3.4%	98	365	1,243	4,321	4,972	5,763	6,734	7,947	**9,502**	11,557	14,453	**18,873**	26,814
4	Ah-AH	2.6%	80	272	971	3,707	4,304	5,037	5,948	7,106	**8,603**	10,609	13,404	**17,586**	25,009
5	Ss-PL	1.5%	102	308	827	2,402	2,727	3,120	3,610	4,225	**5,021**	6,092	7,620	**10,037**	14,657
6	Ss-EC	1.0%	120	330	836	2,345	2,656	3,027	3,478	4,049	**4,795**	5,816	7,285	**9,605**	14,010
7	Ss-SE	1.1%	96	295	850	2,668	3,054	3,520	4,101	4,838	**5,797**	7,102	8,974	**11,916**	17,410
8	Ns-ST	1.4%	126	392	1,143	3,405	3,855	4,396	5,057	5,880	**6,924**	8,314	10,247	**13,204**	18,445
9	Ns-BM	1.2%	117	374	1,119	3,471	3,954	4,539	5,256	6,153	**7,301**	8,825	10,972	**14,256**	20,134
10	IF	1.0%	59	177	599	2,279	2,646	3,091	3,642	4,351	**5,291**	6,578	8,483	**11,593**	17,777

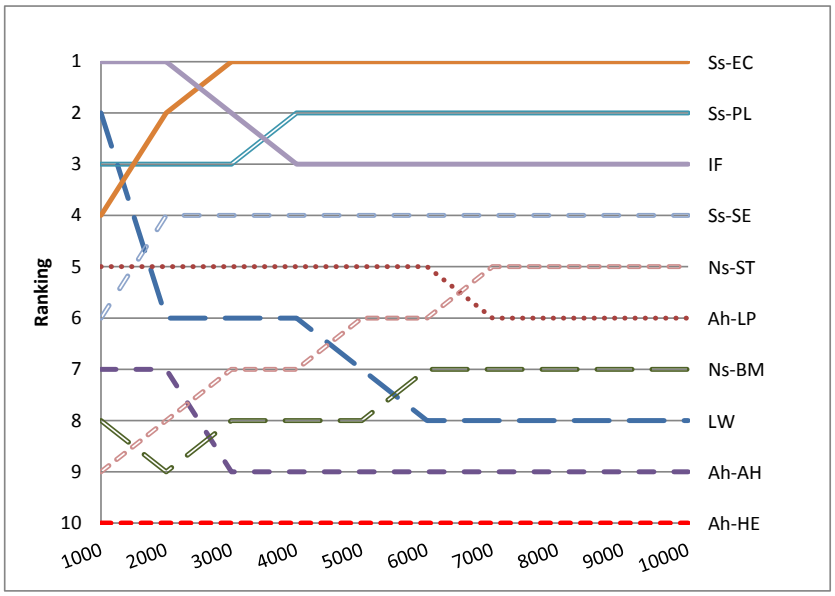
Table 4-5 Ranking in Required Number of Words to Attain Different Levels of Text Coverage out of the 10 Different Genres in VDRJ

Cumulative Text Coverage	60%	70%	80%	90%	91%	92%	93%	94%	95%	96%	97%	98%	99%
LW	2	2	2	6	6	7	8	8	8	8	8	8	8
Ah-LP	3	3	3	5	5	5	5	5	6	6	7	7	7
Ah-HE	6	8	10	10	10	10	10	10	10	10	10	10	10
Ah-AH	4	4	7	9	9	9	9	9	9	9	9	9	9
Ss-PL	7	6	4	3	3	3	2	2	2	2	2	2	2
Ss-EC	9	7	5	2	2	1	1	1	1	1	1	1	1
Ss-SE	5	5	6	4	4	4	4	4	4	4	4	4	3
Ns-ST	10	10	9	7	7	6	6	6	5	5	5	5	5
Ns-BM	8	9	8	8	8	8	7	7	7	7	6	6	6
IF	1	1	1	1	1	2	3	3	3	3	3	3	4

*Text coverage includes Assumed Known Words.

Graph 4-2 Ranking in Required Number of Words to Attain Different Levels of Text Coverage out of the 10 Different Genres in VDRJ

*The higher the ranking, the smaller the required number of words.



If we compare the coverage in different genres, as shown in Table 4-3 and 4-4, texts in social sciences such as economics and commerce (EC) and politics and law (PL) are lexically more homogeneous than other genres. Internet-forum sites (IF) and literary works (LW) provide higher coverage than social sciences at the 1,000-2,000 word levels and require fewer words to attain 60-80% coverage; however, both EC and PL overtakes LW at the 2,000 word level, and EC overtakes IF at 3,000, and PL overtake IF at 4,000. Both EC and PL keep higher coverage than IF and LW beyond the top 3,000 word level. Coverage in natural sciences (Ns: ST and BM) is lower than humanities and arts (Ha: languages, linguistics and philosophy (LP), history and ethnology (HE) and arts and other humanities (AH)) in the high frequency band, yet, both science and technology (ST) and biology and medicine (BM) keep up with the same levels as humanities and arts (Ha) at 3,000-10,000 word levels (Table 4-3), and then overtake Ha at 97% coverage and upwards (Table 4-4). Among all the ten genres, the largest gap exists between economy and commerce (EC) and history and ethnology (HE). EC requires only half number of words required in HE at both 95 and 98% coverage points. It is apparent that arts and humanities require more words

especially in low-frequency bands. HE requires 9,502 words for 95% coverage and 18,873 for 98% coverage. To cover the 3% increase, more than 9,000 words are required in HE. All in all, internet-forum sites (IF), literary works (LW) and humanities and arts (Ha) require fewer words in high frequency band; however, as the frequency level gets lower, social and natural sciences provide higher coverage and require fewer words overall. This is shown more clearly in Table 4-5 and Graph 4-2.

Another important point is the gap in required number of words shown in Table 4-4 between AD, the eight academic domains in VDRJ and each academic genre. AD generally requires a larger number of words than social and natural sciences. This means that vocabulary learning can be remarkably more efficient if learners decide their specialized fields early. For example, the gap between AD and EC/PL is more than 4,000 words at the 95% coverage point. To learn 4,000 words will generally require one or two years at least. Of course, it will not be always good to choose the major too early; however, considering the burden of vocabulary learning, it is worth being more conscious about the language use in the learner's own major field earlier.

In sum, it cannot be stated that Japanese vocabulary is more diverse than other languages. As for the lexical homogeneity of media, internet-forum sites are the most homogeneous among the four media, books comes second, and magazines and newspapers are lexically more diverse than the other two. Book texts contain a wide range of texts from casual novels to formal academic texts in different disciplines, which leads considerably different results of text coverage in different genres. Literature and humanities are lexically more homogenous in the high frequency band; however, social and natural science texts are lexically more homogeneous overall.

Informality and colloquiality

Proportions of types and tokens by word origin in different genres of VDRJ are

shown in Tables 4-6 to 4-8. The word type data was automatically given by UniDic (Den et al., 2009; the electronic dictionary used for morphological analysis) when the word-segmentation was done. Signs, function words, proper nouns, numerals and unknown words are all eliminated from the statistics so as to compare the results with Mogi et al. (2005). In the tables, "Western-origin & Others" are mostly Western-origin; however, words which are non-Japanese-origin and non-Chinese-origin are all included in this category.

Table 4-6 Proportion of Word Origins in Different Genres (Counted by Lexemes)

Genre in VDRJ	LW (%)	AD (%)	IF (%)	Whole=VDRJ (%)
Japanese-origin	37.5	30.1	35.2	31.8
Chinese-origin	47.3	50.3	43.1	48.2
Western-origin & Others (*)	10.8	15.5	17.6	15.7
Mixed-origin	4.3	4.0	4.0	4.3

* "Western-origin & Others" are overwhelmingly English-origin; however, words which are non-Japanese-origin and non-Chinese-origin are all included in this category.

Table 4-7 Proportion of Word Origins in the Three Large Genres of VDRJ (Counted by Tokens = Text Coverage)

Genre in VDRJ	LW (%)	AD (%)	IF (%)	Whole=VDRJ (%)
Japanese-origin	70.8	52.2	60.4	57.9
Chinese-origin	24.7	42.4	30.5	36.3
Western-origin & Others (*)	2.7	3.9	7.3	4.1
Mixed-origin	1.8	1.5	1.7	1.6

* "Western-origin & Others" are overwhelmingly English-origin; however, words which are non-Japanese-origin and non-Chinese-origin are all included in this category.

Table 4-8 Proportion of Word Origins in the Ten Sub-Sections of VDRJ (Counted by Tokens = Text Coverage)

Genre	VDRJ									
	LW	AD								IF
		Ah			Ss			Ns		
	LW	LP	HE	AH	PL	EC	SE	ST	BM	IF
Japanese-origin	**70.8**	**58.0**	53.5	**60.3**	*43.8*	*44.1*	51.3	*47.4*	54.5	**60.4**
Chinese-origin	*24.7*	*38.0*	42.3	*34.5*	**52.4**	**49.1**	43.7	*43.8*	38.8	*30.5*
Western-origin & Others (*)	2.7	2.5	2.7	3.6	*2.2*	**5.2**	3.6	**7.6**	**5.3**	**7.3**
Mixed-origin	1.8	1.5	1.5	1.6	1.6	1.6	1.5	1.2	1.4	1.7

* "Western-origin & Others" are overwhelmingly English-origin; however, words which are non-Japanese-origin

Table 4-6 and 4-7 show that internet-forum sites (IF) and literary works (LW) contain substantially more Japanese-origin words than the other eight academic domains (AD) in VDRJ both in lexeme and token. This result is in line to the result of multidimensional scaling and other analyses in 3.3.6. As mentioned in 4.2, Japanese-origin words are used more for daily topics and Chinese-origin words are used more for formal and academic discourse. Not only in Japanese, but also in many languages, borrowings are generally introduced in some special domain which the indigenous vocabulary does not cover. IF and LW are more informal than AD as they are more related to daily lives.

Nevertheless, in AD, the eight domains have considerably different proportions for word origins (In Table 4-8: bold and italic letters is used to show high and low figures). All the eight domains have fewer Japanese-origin words and more Chinese-origin words; however, arts and humanities such as languages, linguistics and philosophy (LP) and arts and other humanities (AH) tend to be high in the Japanese-origin but low in the Chinese-origin words while social sciences such as economics and commerce (EC) and politics and law (PL) have the opposite tendency. Western-origin words provide substantially higher proportions in natural sciences such as science and technology (ST) and biology and medicine (BM) as well as IF and EC. Comparing this result with the order of lexical homogeneity, it can be concluded that the more lexically homogeneous in the high frequency band the domain, the more informal the vocabulary use in the domain.

According to Mogi, Yamaguchi, Maruyama, & Tanaka (2005), the proportions of Japanese, Chinese and Western origin word tokens in magazines are 51.8%, 37.5% and 8.8% respectively, and in newspapers, the proportions are 39.4%, 54.1% and 5.0% respectively (p.343). Compared to the domains in VDRJ, magazines contain similar proportions of Japanese and Chinese origin words to AD in general but contain a remarkably higher proportion of Western-origin words. However, Mogi et al. (2005) also

reveal that only academic or technical journals (which are categorized as magazines here) contain a significantly high proportion of Chinese-origin words at 53.1%, which is much higher than the other six genres such as 'family' and 'hobbies'. Therefore, except for academic and technical journals, magazines will be more casual than AD in general. Newspapers provide a very similar pattern to social sciences, or are even more formal as they contain a higher proportion of Chinese-origin words at 54.1%, which is the highest among all the genres and media.

Table 4-9 and Graph 4-3 are the proportion of word lexemes by word origin at different frequency levels. The word level is based on the word ranking for the general written Japanese (WWJ).

As shown in Table 4-9 and Graph 4-3, the proportion of Japanese-origin words is the highest in the first 1,000 words (01K) in VDRJ, and drastically decreases at 02K, and keeps almost the same level at approximately one third up to 20K. Related to that, the proportion of Chinese-origin words increases sharply at 02K, and keeps the same level at approximately half up to 20K. Western-origin words only occupy 1.2% at 01K, but increase gradually to 05K at 10.6%, and then keep the same level up to 20K. At the very low frequency band at 21K+, Western-origin words sharply increase to 17.5% while Japanese and Chinese origin words decreases a little. (Please note that the proportion is for word lexemes but not tokens.) In light of the fact that LW and IF contain many Japanese-origin words, it can be postulated that high frequency words contain more everyday words in general.

In all, LW is the most informal, IF comes to the second, arts and humanities texts come third, magazines come fourth, and natural science texts come fifth. Social science texts are more formal, and newspapers are slightly more formal in vocabulary use overall. From the word-origin aspects, the first two thousand words (01K-02K) have a considerably different proportion from 03K or above. This high frequency band contains more informal

words.

Nishimura (2010) explores register variations by some indexical use of colloquial forms or categories. For example, novels, which should be a close category to LW here, contain more colloquial forms than IF but fewer than magazines and newspaper editorials. For example, the proportion for the sentence-final particles (終助詞 'shuujoshi'), which denote the modality of the language user's attitude in printed novels, is 5.3% among all the particle (助詞 'joshi') usages in printed novels while the proportions in IF, magazine and newspaper editorials are 8.3, 1.8 and 0.8% respectively. The proportions for the contraction てる '-teru' (← ている '-teiru') in printed (non-digitized) novels is at 2.0% among all the auxiliary verb (助動詞 'jodoushi') usages, while the proportions in the IF, magazines and newspaper editorials are 3.9%, 0.3% and 0.0% respectively (Table 3, p. 77). These data prove that IF is more colloquial than novels (i.e. LW). This order is opposite to the rank of informality. In other words, LW is more casual but less colloquial than IF, and vice versa.

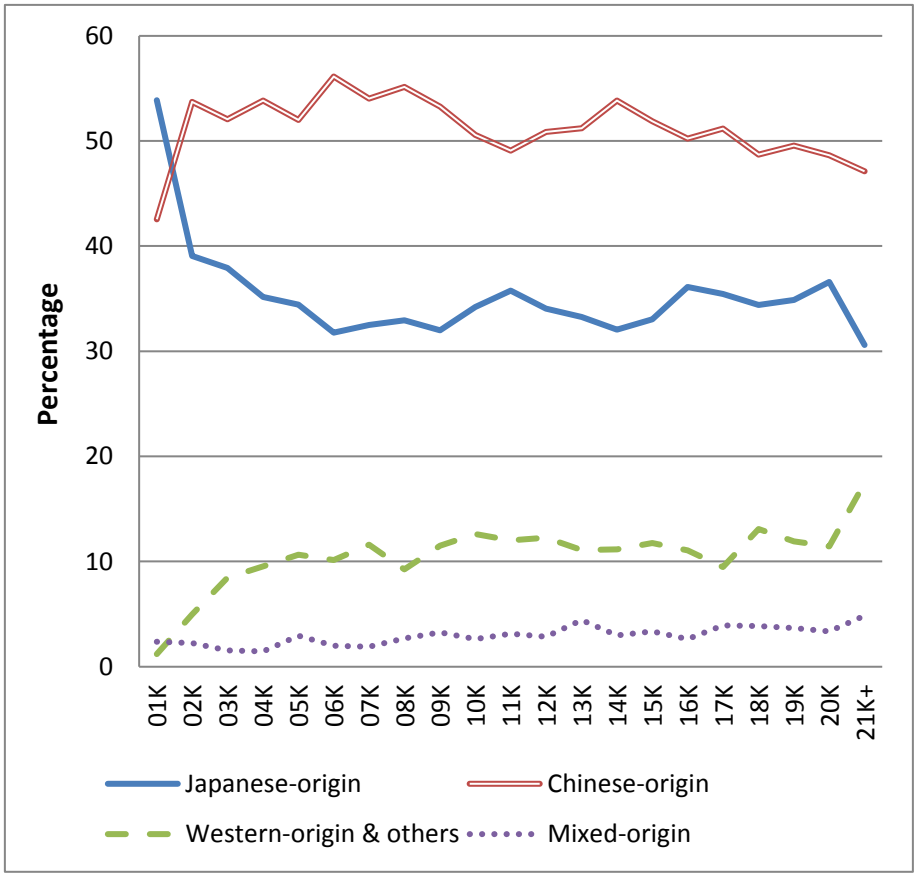
(From here down blank.)

Table 4-9 Proportion (Percentage) of Word Origins at Different Frequency Levels in VDRJ (Counted by Lexemes)

Word Level	01K	02K	03K	04K	05K	06K	07K	08K	09K	10K	11K	12K	13K	14K	15K	16K	17K	18K	19K	20K	21K+	Whole
Japanese-origin	53.9	39.1	37.9	35.2	34.4	31.8	32.5	32.9	32.0	34.2	35.8	34.0	33.3	32.0	33.0	36.1	35.4	34.4	34.9	36.6	30.6	31.8
Chinese-origin	42.6	53.7	52.1	53.8	52.0	56.1	54.0	55.1	53.3	50.6	49.1	50.9	51.2	53.8	51.9	50.2	51.2	48.7	49.5	48.6	47.1	48.2
Western-origin & others	1.2	5.0	8.5	9.5	10.6	10.1	11.6	9.3	11.5	12.6	12.0	12.3	11.1	11.2	11.8	11.1	9.5	13.1	11.9	11.4	17.5	15.7
Mixed-origin	2.4	2.2	1.5	1.5	2.9	2.0	1.9	2.7	3.3	2.6	3.1	2.8	4.4	3.0	3.3	2.6	3.9	3.9	3.7	3.3	4.8	4.3

* Proper nouns, signs and unknown words are eliminated from the statistics.
* The word level is based on the word ranking for the general written Japanese (WWJ).
* "Western-origin & Others" are overwhelmingly English-origin; however, words which are non-Japanese-origin and non-Chinese-origin are all included in this category.

Graph 4-3 Proportion (Percentage) of Word Origins at Different Frequency Levels
(Counted by Lexemes)



Specific words in magazines, newspapers and VDRJ

As shown above, magazines and newspapers are lexically more diverse than books and require more words to attain a certain level of text coverage. This will mean that magazines and newspapers will probably contain more specific words than books.

Media-specific words by comparing frequency rankings are shown in Tables from 4-10 to 4-12.

Specific words in magazines and newspapers have distinct dispositions (Table 4-10 and 4-11). Magazine-like words are terms for hobbies (e.g. モーター 'mo^ta^' (motor) and ジャズ 'jazu' (jazz)) and terms for advertisement (e.g. カタログ 'katarogu' (catalogue) and 当社 'tousha' (this company)). Place names (e.g. ジャパン 'japan' (Japan) and 金沢 'kanazawa' (Kanazawa, a historic city in Ishikawa prefecture) seem to be more frequent in magazines as well. These words account for the higher lexical diversity in magazines.

Table 4-10 Top 30 Magazine-specific Words Extracted by Comparing the Frequency Rankings

Magazine-specific words	Reading	English Translation
リットル	rittoru	liter
装備	soubi	equipment
ジャパン	japan	Japan
モーター	mo-ta-	motor
アルバム	arubamu	album
俳句	haiku	haiku
ジャズ	jazu	jazz
搭載	tousai	loading/equiped
カタログ	katarogu	catalog
模型	mokei	model/pattern
将棋	shougi	*shogi*/Japanese chess
タイヤ	taiya	tire
宝塚	takarazuka	Takarazuka (Revue)
録音	rokuon	recording
コレクション	korekushon	collection
バイオリン	baiorin	violin
ガイド	gaido	guide
当社	tousha	this company
競馬	keiba	horse race
本田	honda	Honda
イラスト	irasuto	illustration
出力	shutsuryoku	output
締め切り	shimekiri	deadline
車検	shaken	official vehical inspection
切手	kitte	postage stamp
金沢	kanazawa	Kanazawa (place name)
バスケット	basuketto	basket/basket ball
仕上げ	shiage	a finish
モーツァルト	mo-tsaruto	Mozart
スペシャル	supesharu	special

Table 4-11 Top 30 Newspaper-specific Words Extracted by Comparing the Frequency Rankings

Newspaper-specific words	Reading	English Translation
会談	kaidan	talks/conference
見通し	mitooshi	visibility/prospect/forcast
被告	hikoku	defendant/the accused
首脳	shunou	head/leader
政党	seitou	political party
論議	rongi	argument
見直し	minaoshi	revision/reworking
赤字	akaji	deficit/the red
打ち出す	uchidasu	work out/come out with
盛り込む	morikomu	incorporate
強まる	tsuyomaru	grow/strengthen
懸念	kenen	concern/fear
都内	tonai	within the Metropolitan area
疑惑	giwaku	suspicion/doubt
税制	zeisei	taxation system
加盟	kamei	joining/participation
通貨	tsuuka	currency
決着	ketchaku	settlement/decision
撤退	tettai	withdrawal
常務	joumu	managing director
提言	teigen	offering an opinion/offered opinion
合同	goudou	joint/combination/union
参入	sannyuu	entry (of a market)
新設	shinsetsu	establishment
冷戦	reisen	the Cold War
対日	tainichi	to Japan/toward Japan
買収	baishuu	buy out
急増	kyuuzou	rapid increase
シンポジウム	shimpojiumu	symposium
正午	shougo	noon

Newspaper-like words are terms for politics and economy (e.g. 政党 ‘seitou’ (political party) and 通貨 ‘tsuuka’ (currency)). There are also some other types of words for news such as terms for events (e.g. シンポジウム ‘shimpojiumu’ (symposium) and terms for time (e.g. 正午 ‘shougo’ (noon)). Terms for politics and economy will probably overlap

with the frequent words in PL and EC in VDRJ, and that is probably why the word origin proportions are also similar to each other.

Table 4-12 Top 30 VDRJ (Mostly Book)-specific Words Extracted by Comparing the Frequency Rankings

VDRJ (Mostly book)-specific words	Reading	English Translation
ああ	aa	ah
概念	gainen	concept
黙る	damaru	hold one's tongue
見なす	minasu	consider (... as)
恐ろしい	osoroshii	terrifying
定義	teigi	definition
属する	zokusuru	belong
性質	seishitsu	nature/disposition
そちら	sochira	your place
階級	kaikyuu	class/estate
有する	yuusuru	possess
検索	kensaku	retrieval/searching for
身分	mibun	status/position
見いだす	miidasu	find out/detect
不幸	hukou	unhappiness/misfortune
学問	gakumon	studies/scholarship
観念	kan'nen	sense/notion
久しい	hisashii	for a long time
引用	in'you	citation/quotation
著作	chosaku	writing/literary work
この世	konoyo	this world/the present life
ほめる	homeru	praise
塩	shio	salt
ギリシャ	girisha	Greece
次郎	jirou	Jiro (person's name)
このごろ	konogoro	lately/recently
道徳	doutoku	morality/morals
典型	tenkei	type/model
仏教	bukkyou	Buddhism
秀吉	hideyoshi	Hideyoshi (historic person's name)

VDRJ seems to contain more academic words (e.g. 概念 ‘gainen’ (concept) and 定

義 ‘teigi’ (definition)) than magazines and newspapers (Table 4-12). It also contains some literary words (e.g. ああ ‘aa’ (ah) and 黙る ‘damaru’ (hold one’s tongue)). However, compared to the specific words in magazines and newspapers, specific words in VDRJ are not so distinctive. This is appropriate for this study because the purpose of developing VDRJ is to reflect more general written vocabulary. VDRJ contains both casual (i.e. LW and IF) and formal domains (i.e. AD, especially social and natural sciences) as well as more general texts.

Domain-specific words in sub-genres in VDRJ will be extracted and discussed in Chapter 7.

4.2.3 Conclusion of 4.2

The overall comparison (ranking) in lexical homogeneity, informality and colloquiality in different genres and media is shown in Table 4-13.

Table 4-13 Ranking in Lexical Homogeneity, Informality and Colloquiality in Different Genres and Media

Aspect (Index)	Genre	LW	IF	Magazine	AD-Ah	AD-Ns	AD-Ss	Newspaper
	Media	Book	Internet	Magazine	Book	Book	Book	Newspaper
	VDRJ	✔	✔		✔	✔	✔	
Contemporariness (Words for Current Events)		Low	High	High	Low	Low	Low	High
Lexical homogeneity [opp. diversity] in high frequency band (text coverage)		2	1	6	3	5	4	7
Lexical homogeneity [opp. diversity] in low frequency band (text coverage)		4	2	7/6*	5	3	1	6/7*
Informality [opp. Formality] (Japanese-origin words/Chinese-origin words)		1	2	4	3	5	6	7
Colloquiality (sentence-final particles, contractions etc.)		2	1	3	4?	4?	4?	7

* Newspapers have higher lexical homogeneity in the middle frequency band, while magazines go higher in the low frequency band.

On the whole, books (especially AD) generally have the intermediate characteristics between magazines and newspapers. Books contain wide range of genres; however, book vocabulary is more stable as it does not contain many terms for current events. Therefore, the corpus that VDRJ was developed from is basically suitable for educational purposes

such as selecting and ordering words to learn. The weakness for informality and current terms is compensated by literary works (LW) and the internet-forum sites (IF).

4.3 Overall distribution of words by part of speech

The sub-research-question (SRQs) in this section is as follows. (The SRQ number follows the previous section.)

SRQ 8)How are the parts of speech (POS) distributed in the book corpus and internet-forum site corpus in BCCWJ 2009 monitor version which VDRJ is made from? What are the findings there?

This question is exploratory but not a question to test a specific hypothesis. There are some studies on POS distribution in Japanese (e.g. NLRI, 1964, 1971); however, considering the fact that the corpus for this study is the first large balanced Japanese corpus including books and that the dictionary UniDic used for morphological analysis can identify more types of POS categories than before, new findings would be expected by comparing the POS distribution in VDRJ.

4.3.1 Method

The number or proportion of word lexemes and tokens by POS are shown in tables or graphs by 1,000 word level and by sub-section of VDRJ. Computation of data can be done on VDRJ spread sheet using the filtering and the pivot table functions.

4.3.2 Results and discussion

The results are shown in Tables 4-14 to 4-18 and Graph 4-4. Discussions are made along with each table.

The absolute percentage figures in Table 4-14 must be compared to the ones from

equally-sized corpus because the number of lexemes of low-frequency words is substantially influenced by the corpus size. In consequence, it is not meaningful to compare the figures with previous studies such as NLRI (1964, 1971) of which the corpus sizes are approximately 0.54 million and 1.21 million tokens.

Table 4-14 Number and Ratio of Words in VDRJ by Part of Speech (Counted by Lexemes)

Part of Speech (Japanese)	Part of Speech	Word origin	All	All (%)	All (%)
助詞-格助詞	Case Particle	Particle	13	0.009%	0.058%
助詞-係助詞	Binding Particle		3	0.002%	
助詞-副助詞	Adverbial Particle		32	0.023%	
助詞-接続助詞	Conjunctive Particle		14	0.010%	
助詞-終助詞	Sentence-final Particle		21	0.015%	
助動詞	Auxiliary Verb	Auxiliary Verb	57	0.040%	0.042%
形状詞-助動詞語幹	Adjectival Noun: Auxiliary Verb Stem		2	0.001%	
名詞-普通名詞-一般	Common Noun	Noun	88,535	62.371%	84.337%
名詞-固有名詞-一般	Proper Noun: General		3,184	2.243%	
名詞-固有名詞-人名-一般	Proper Noun-General Person's Name		10,648	7.501%	
名詞-固有名詞-人名-姓	Proper Noun: Family Name		4,618	3.253%	
名詞-固有名詞-人名-名	Proper Noun: Given Name		5,238	3.690%	
名詞-固有名詞-地名-一般	Proper Noun: General Place-name		6,667	4.697%	
名詞-固有名詞-地名-国	Proper Noun: Country's Name		372	0.262%	
名詞-数詞	Noun: Numerals		71	0.050%	
名詞-普通名詞-副詞可能	Adverbial Noun		382	0.269%	
代名詞	Pronoun	Pronoun	80	0.056%	0.056%
名詞-普通名詞-サ変可能	Verbal Noun	Verbal Noun	8,590	6.051%	6.099%
名詞-普通名詞-サ変形状詞可能	Verbal Adjectival Noun		67	0.047%	
形状詞-タリ	*Tari* Nominal Adjective	Nominal Adjective	243	0.171%	1.844%
形状詞-一般	General Nominal Adjective		1,076	0.758%	
名詞-普通名詞-形状詞可能	Adjectival Noun		1,299	0.915%	
動詞-一般	Verb: General	Verb	7,242	5.102%	5.153%
動詞-非自立可能	Verb: Possibly Bound		72	0.051%	
形容詞-一般	Adjective: General	Adjective	643	0.453%	0.455%
形容詞-非自立可能	Adjective: Possibly Bound		3	0.002%	
連体詞	Prenoun Adjectival	Prenoun Adjectival	42	0.030%	0.030%
副詞	Adverb	Adverb	1,706	1.202%	1.202%
接続詞	Conjunction	Conjunction	19	0.013%	0.013%
感動詞-一般	Interjection: General	Interjection	179	0.126%	0.136%
感動詞-フィラー	Interjection: Filler		14	0.010%	
接頭辞	Prefix	Prefix	128	0.090%	0.090%
接尾辞-名詞的-一般	General Nominal Suffix	Suffix	338	0.238%	0.439%
接尾辞-名詞的-サ変可能	Verbal Nominal Suffix		3	0.002%	
接尾辞-名詞的-副詞可能	Adverbial Nominal Suffux		8	0.006%	
接尾辞-名詞的-助数詞	Suffix: Counter		252	0.178%	
接尾辞-動詞的	Verbal Suffix		6	0.004%	
接尾辞-形容詞的	Adjectival Suffix		9	0.006%	
接尾辞-形状詞的	Nominal Adjective Suffix		7	0.005%	
記号-一般	General Sign	Sign	57	0.040%	0.046%
補助記号-ＡＡ-一般	Auxiliary AA Sign		1	0.001%	
補助記号-一般	Auxiliary Sign		8	0.006%	
		Total	141,949	100.000%	100.000%

It is still worth comparing the relative proportions between categories. For example, verbal nouns are more than verbs in number of lexemes. The ratio is approximately 6:5.

Some findings with the detailed classification of part of speech (POS), which has been made possible by UniDic (Den et al., 2009) are as follows.

Table 4-15 Number of Words in VDRJ by Part of Speech and Word Origin (Counted by Lexemes)

Part of Speech (Japanese)	Part of Speech	Word origin	Japanese-origin	Chinese-origin	Western-origin & Others	Mixed-origin	Proper Noun	Sign	Unknown
助詞-格助詞	Case Particle	Particle	13						
助詞-係助詞	Binding Particle		3						
助詞-副助詞	Adverbial Particle		31	1					
助詞-接続助詞	Conjunctive Particle		14						
助詞-終助詞	Sentence-final Particle		21						
助動詞	Auxiliary Verb	Auxiliary Verb	55			2			
形状詞-助動詞語幹	Adjectival Noun: Auxiliary Verb Stem		1	1					
名詞-普通名詞-一般	Common Noun	Noun	13,360	26,543	10,721	2,559	115	958	34,279
名詞-固有名詞-一般	Proper Noun: General						3,184		
名詞-固有名詞-人名-一般	Proper Noun-General Person's Name						10,648		
名詞-固有名詞-人名-姓	Proper Noun: Family Name						4,618		
名詞-固有名詞-人名-名	Proper Noun: Given Name						5,238		
名詞-固有名詞-地名-一般	Proper Noun: General Place-name						6,667		
名詞-固有名詞-地名-国	Proper Noun: Country's Name						372		
名詞-数詞	Noun: Numerals		3	53	1	13			1
名詞-普通名詞-副詞可能	Adverbial Noun		119	241	1	21			
代名詞	Pronoun	Pronoun	59	14		7			
名詞-普通名詞-サ変可能	Verbal Noun	Verbal Noun	523	7,435	518	82	5	4	23
名詞-普通名詞-サ変形状詞可能	Verbal Adjectival Noun		7	53	5	1			1
形状詞-タリ	*Tari* Nominal Adjective	Nominal Adjective	2	241					
形状詞-一般	General Nominal Adjective		356	407	245	50	1		17
名詞-普通名詞-形状詞可能	Adjectival Noun		114	902	209	50			24
動詞-一般	Verb: General	Verb	6,814	2		401	2		23
動詞-非自立可能	Verb: Possibly Bound		71			1			
形容詞-一般	Adjective: General	Adjective	595			35			13
形容詞-非自立可能	Adjective: Possibly Bound		3						
連体詞	Prenoun Adjectival	Prenoun Adjectival	33			8			1
副詞	Adverb	Adverb	1,556	107	3	28	1		11
接続詞	Conjunction	Conjunction	18	1					
感動詞-一般	Interjection: General	Interjection	166	1		4			8
感動詞-フィラー	Interjection: Filler		14						
接頭辞	Prefix	Prefix	9	119					
接尾辞-名詞的-一般	General Nominal Suffix	Suffix	98	240					
接尾辞-名詞的-サ変可能	Verbal Nominal Suffix		1	2					
接尾辞-名詞的-副詞可能	Adverbial Nominal Suffux		7	1					
接尾辞-名詞的-助数詞	Suffix: Counter		13	83	142	4	2	8	
接尾辞-動詞的	Verbal Suffix		6						
接尾辞-形容詞的	Adjectival Suffix		9						
接尾辞-形状詞的	Nominal Adjective Suffix		5	1	1				
記号-一般	General Sign	Sign						57	
補助記号-ＡＡ-一般	Auxiliary AA Sign							1	
補助記号-一般	Auxiliary Sign							8	
		Total	24,099	36,448	11,846	3,266	30,853	1,036	34,401

Nouns

Table 4-14 shows that there is a considerable number of proper nouns (counted by

lexemes) in the corpus. Proper nouns occupy more than 20% in VDRJ (counted by lexemes); however, they only provide around 2% of the tokens (text coverage), as most proper nouns only occur once or twice in the corpus. The larger the corpus, the larger the number of low-frequency lexemes. Therefore, proper nouns, signs and unknown words are eliminated from the statistics in Table 4-16.

Table 4-16 Proportion (Percentage) of Word Origins in Each Part of Speech (Counted by Lexemes)

Part of Speech (Japanese)	Part of Speech	Word origin	Japanese-origin	Chinese-origin	Western-origin & Others	Mixed-origin
助詞-格助詞	Case Particle	Particle	100.0			
助詞-係助詞	Binding Particle		100.0			
助詞-副助詞	Adverbial Particle		96.9	3.1		
助詞-接続助詞	Conjunctive Particle		100.0			
助詞-終助詞	Sentence-final Particle		100.0			
助動詞	Auxiliary Verb	Auxiliary Verb	96.5			3.5
形状詞-助動詞語幹	Adjectival Noun: Auxiliary Verb Stem		50.0	50.0		
名詞-普通名詞-一般	Common Noun	Noun	25.1	49.9	20.2	4.8
名詞-数詞	Noun: Numerals		4.3	75.7	1.4	18.6
名詞-普通名詞-副詞可能	Adverbial Noun		31.2	63.1	0.3	5.5
代名詞	Pronoun	Pronoun	73.8	17.5		8.8
名詞-普通名詞-サ変可能	Verbal Noun	Verbal Noun	6.1	86.9	6.1	1.0
名詞-普通名詞-サ変形状詞可能	Verbal Adjectival Noun		10.6	80.3	7.6	1.5
形状詞-タリ	*Tari* Nominal Adjective	Nominal Adjective	0.8	99.2		
形状詞-一般	General Nominal Adjective		33.6	38.5	23.2	4.7
名詞-普通名詞-形状詞可能	Adjectival Noun		8.9	70.7	16.4	3.9
動詞-一般	Verb: General	Verb	94.4	0.0		5.6
動詞-非自立可能	Verb: Possibly Bound		98.6			1.4
形容詞-一般	Adjective: General	Adjective	94.4			5.6
形容詞-非自立可能	Adjective: Possibly Bound		100.0			
連体詞	Prenoun Adjectival	Prenoun Adjectival	80.5			19.5
副詞	Adverb	Adverb	91.9	6.3	0.2	1.7
接続詞	Conjunction	Conjunction	94.7	5.3		
感動詞-一般	Interjection: General	Interjection	97.1	0.6		2.3
感動詞-フィラー	Interjection: Filler		100.0			
接頭辞	Prefix	Prefix	7.0	93.0		
接尾辞-名詞的-一般	General Nominal Suffix	Suffix	29.0	71.0		
接尾辞-名詞的-サ変可能	Verbal Nominal Suffix		33.3	66.7		
接尾辞-名詞的-副詞可能	Adverbial Nominal Suffux		87.5	12.5		
接尾辞-名詞的-助数詞	Suffix: Counter		5.4	34.3	58.7	1.7
接尾辞-動詞的	Verbal Suffix		100.0			
接尾辞-形容詞的	Adjectival Suffix		100.0			
接尾辞-形状詞的	Nominal Adjective Suffix		71.4	14.3	14.3	
		Total	31.9	48.2	15.7	4.3

* Proper nouns, signs and unknown words are eliminated from the statistics because the number of these kinds of word types is substantially influenced by the corpus size.

Table 4-15 and 4-16 show that a substantial number and proportion of nouns, verbal nouns and nominal adjectives are of Chinese origin. As is widely known, loanwords in

Japanese are basically introduced as nouns which derive verbal noun or adjectival noun by adding -する '-suru' or -な '-na' to the noun (NLRI, 1971, p 23). These words are generally important for students and academics as they are largely used in formal or academic texts.

Affixes

Among identified numbers of lexemes in the sub-categories of particles and others, the proportion of Chinese-origin affixes is noticeable (Table 4-15 and 4-16). In Japanese, 751 affixes are identified by UniDic in the corpus. This is remarkably more than in English where only 91 affixes are identified (Level 1 to 6 in Bauer & Nation (1993)). The majority of suffixes are nominal and approximately 70% of them ((240+2+1)/(98+1+7+240+2+1) ≒.696) are of Chinese origin. In addition, the vast majority of prefixes (119/128 ≒.930) are of Chinese origin, too. These figures suggest that understanding word formation with Chinese-origin affixes is important for understanding Japanese, especially formal or academic texts.

The conclusion is also endorsed by the data from Table 4-17 and 4-18. The proportion of suffixes is at a high level at 2-3% of lexemes even from 03K to 07K. Adding the percentage of prefixes, the total percentage of affixes is around 3-4% between the 03K and 07K levels. The mid-frequency band from 03K to 07K is generally thought to contain intermediate vocabulary which contributes to formal or academic texts more than the basic vocabulary. (e.g. as shown in 4.2, the text coverage in the mid-frequency band in newspaper texts exceeds the coverage in magazine texts.)

Table 4-17 Proportion of Part of Speech at Each 1000 Word Level in VDRJ (Counted by Lexemes)

Word Level	01K	02K	03K	04K	05K	06K	07K	08K	09K	10K	11K	12K	13K	14K	15K	16K	17K	18K	19K	20K	01K–20K
Particle	4.1	1.0	0.4	0.1	0.1	0.1	0.1	0.2	0.3	0.0	0.1	0.2	0.1	0.1	0.2	0.0	0.1	0.1	0.2	0.0	0.4
Auxiliary Verb	1.7	0.4	0.4	0.4	0.1	0.6	0.2	0.0	0.0	0.1	0.1	0.2	0.1	0.0	0.1	0.0	0.4	0.2	0.0	0.0	0.3
Noun	38.2	43.4	48.1	47.7	52.4	54.2	53.9	53.3	54.9	56.6	53.1	56.7	54.7	56.7	56.8	60.0	59.4	61.2	60.5	60.6	54.1
Pronoun	2.2	0.2	0.4	0.2	0.1	0.1	0.1	0.2	0.0	0.1	0.1	0.4	0.1	0.2	0.1	0.1	0.1	0.0	0.2	0.1	0.3
Verbal Noun	10.0	19.3	19.9	20.9	19.9	20.0	20.2	19.4	20.4	18.5	18.5	19.2	18.6	18.9	19.5	15.7	17.6	16.1	16.7	15.1	18.2
Verb	19.0	17.7	15.9	15.6	15.7	14.6	12.3	14.3	14.0	13.0	15.0	12.0	14.6	11.8	11.1	11.8	12.3	11.6	12.1	10.0	13.7
Adjective	3.7	2.9	2.3	1.7	1.1	1.2	1.2	1.5	1.4	1.0	1.4	1.2	1.4	1.1	1.4	2.0	1.1	1.4	1.4	1.8	1.6
Nominal Adjective	3.9	4.7	5.5	5.8	5.4	4.4	5.7	6.3	4.4	5.4	6.8	5.6	6.0	6.2	5.8	6.0	5.0	4.7	4.8	6.3	5.4
Prenoun Adjectival	1.2	0.5	0.2	0.1	0.1	0.0	0.1	0.1	0.1	0.2	0.0	0.1	0.0	0.2	0.0	0.0	0.0	0.0	0.3	0.1	0.2
Adverb	5.7	3.9	3.0	2.6	2.1	1.5	2.4	2.4	3.0	2.8	2.7	2.4	2.8	2.5	2.6	3.2	2.7	3.4	2.6	3.9	2.9
Con-junction	1.0	0.3	0.1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.2	0.0	0.0	0.1	0.1	0.0	0.0	0.0	0.0	0.0	0.1
Inter-jection	0.2	0.5	0.2	0.8	0.6	0.1	0.7	0.1	0.5	0.1	0.2	0.7	0.2	0.1	0.6	0.3	0.2	0.3	0.2	0.5	0.4
Prefix	1.6	0.5	0.5	1.1	0.1	0.3	0.9	0.6	0.3	0.2	0.4	0.1	0.4	0.5	0.5	0.2	0.1	0.1	0.3	0.5	0.5
Suffix	7.5	4.7	3.0	2.9	2.3	2.9	2.1	1.5	0.7	1.8	1.4	1.2	1.0	1.4	1.2	0.7	1.0	0.8	0.7	1.0	2.0
Sign	0.0	0.0	0.1	0.1	0.0	0.0	0.1	0.1	0.0	0.2	0.0	0.0	0.0	0.2	0.0	0.0	0.0	0.1	0.0	0.1	0.1
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

* See Table 4-14 or 4-15 for the detailed classification of part of speech.

Graph 4-4 Number of Word Lexemes of Nouns, Verbal Nouns and Verbs at Different Word Levels in VDRJ

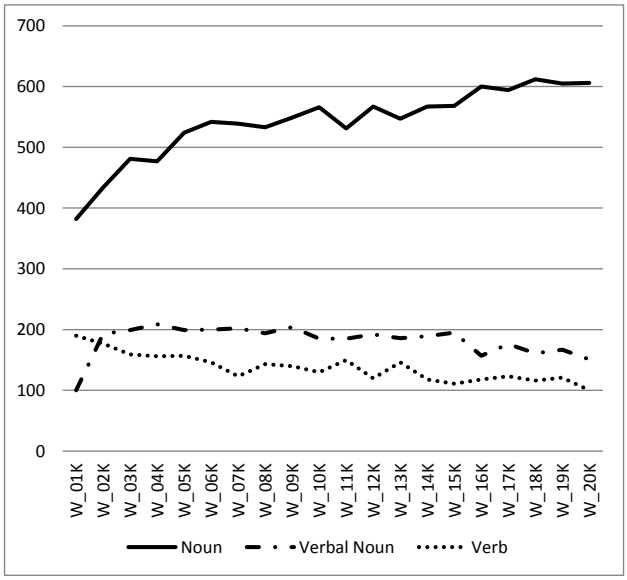


Table 4-18 Proportion of Part of Speech in Each Genre of VDRJ (Counted by Tokens)

Part of Speech (Japanese)	Genre / Part of Speech	VDRJ Whole	LW	LP	HE	AH	PL	EC	SE	ST	BM	IF
助詞	Particle	31.6	33.3	31.7	30.1	31.9	29.4	29.6	31.0	29.6	31.1	32.2
助動詞	Auxiliary Verb	11.0	12.6	10.2	9.2	10.4	8.6	8.4	9.5	8.9	9.5	14.3
名詞	Noun	24.8	21.9	25.2	30.1	25.6	27.8	27.4	25.1	28.5	26.5	21.4
代名詞	Pronoun	1.7	2.5	1.8	1.3	1.9	1.0	1.0	1.3	1.0	1.2	1.6
動名詞	Verbal Noun	5.7	3.2	5.5	5.6	4.7	9.3	9.5	7.8	8.2	6.5	5.5
動詞	Verb	14.3	15.5	14.9	13.2	14.3	13.1	13.2	14.2	13.3	14.3	14.3
形容詞	Adjective	1.7	2.0	1.5	1.2	1.7	1.0	1.2	1.4	1.3	1.7	2.2
名容詞	Nominal Adjective	1.5	1.4	1.6	1.3	1.5	1.6	1.7	1.7	1.6	1.6	1.6
連体詞	Prenoun Adjectival	1.0	1.1	1.2	1.1	1.2	1.0	1.0	1.0	1.0	1.0	0.7
副詞	Adverb	1.9	2.5	1.8	1.5	2.0	1.3	1.4	1.6	1.4	1.7	2.1
接続詞	Conjunction	0.4	0.3	0.6	0.5	0.4	0.6	0.6	0.5	0.5	0.5	0.2
感動詞	Interjection	0.2	0.5	0.1	0.1	0.2	0.1	0.0	0.1	0.1	0.1	0.1
接頭辞	Prefix	0.7	0.7	0.8	0.7	0.7	0.8	0.8	0.7	0.7	0.7	0.8
接尾辞	Suffix	3.4	2.7	3.1	4.1	3.5	4.4	4.2	4.0	3.9	3.7	2.8
記号・補助記号	Sign	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
総計	Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.

Table 4-18 shows that the text coverage by suffixes is between 3.7% and 4.4% in social and natural sciences (social sciences: PL, EC and SE, natural sciences: ST and BM), which are much higher than in the daily-life domains of literary works (LW) and internet-forum sites (IF) at 2.7% and 2.8% respectively. This suggests that the use of suffix may be an index of formality. This issue will be further explored in 4.4.

Verbal nouns

One noticeable thing shown in Table 4-17 is that the pattern of verbal nouns follows the pattern of verbs but not of nouns. Verbal nouns occur much less than verbs in 01K; however, from 02K and above, they occur more frequently and keep a parallel line to verbs (counted by lexemes). Both verbs and verbal nouns keep similar levels in the mid-frequency band and gradually decrease in the low-frequency band while nouns increase constantly (Graph 4-4). Considering together with the fact that verbal nouns are semantically more similar to verbs but function as nouns syntactically, the distribution

patterns of words in lexemes will follow the semantic demand but not the syntactical one.

More than 80% of verbal nouns, as shown in Table 4-15 and 4-16, are Chinese-origin words which are more frequently used in formal and academic texts. Nishimura (2010) claims that it is possible to interpret the use of nouns to mean that nouns have the role to transmit 'information' in the texts, in comparison with interjections which mainly transmit 'emotions'[67] (p.79). Verbal nouns will not generally transmit 'emotions' as they are often used in formal and academic texts even if they follow the distribution pattern of verbs. They will probably function as a conveyer of logic (e.g. 減少-する 'genshou-suru' (decrease)) or writer's stances (e.g. 主張-する 'shuchou-suru' (claim/contend)). These roles are neither conveying emotions nor conveying 'information'. They work to manage the information carried by general (i.e. non-verbal) nouns, which, in a sense, is common in the general function of verbs. The indexicality of verbal noun use for the formality of texts will be discussed together with the consideration of other POS in 4.4

4.3.3 Conclusion of 4.3

The main findings in this section are as follows.

1) Affixes occur more frequently in Japanese than in English, which inevitably means learning affixes is very important in learning Japanese. Especially, Chinese-origin suffixes occur often in Japanese.
2) The distribution of verbal nouns (動名詞／サ変動詞語幹／スル名詞) (counted by lexemes) at different frequency levels is much closer to the distribution of verbs but not nouns.

4.4 Orders of indexicality and informality

As the previous section reveals, some POS such as suffixes and verbal nouns can be

[67] She claims it to argue that the use of some parts of speech can be indices to measure the colloquiality on the continuum between the very colloquial register and the totally literary one.

used as an index for identifying register variation. The sub-research-questions (SRQs) here are as follows. (The SRQ number follows the previous section.)

SRQ 9) Is there any indexical pattern of POS distribution for identifying register variations in Japanese? If yes, what is it?

SRQ 10) What genres in VDRJ are more informal or formal depending on POS distribution?

SRQ 11) How is the informality order of genres based on POS distribution related to the lexical homogeneity order based on text coverage and the informality order based on word-origin distribution (discussed in 4.2)?

In this section, the total tokens of each POS are calculated in different category, and then the two sets of indexical POS for informality and the order of genres by the informality are proposed based on the distribution of the total tokens of (i.e. text coverage by) each POS in different genres.

As introduced in 2.3.4, Kabashima (1955, 1981) and Nishimura (2010) suggest that the distribution of POS has a clear pattern which distinguishes different registers. For example, Nishimura (2010) shows that the proportions of nouns, affixes and verbs are highest in printed written language use, the second highest in online use and the lowest in conversation. Contrary to that, the proportions of interjections, adjectives, adverbs and pronouns are the highest in conversation, the second highest in online language, and the lowest in printed written language. She claims the variations are a 'continuum' on the dimension I "Informational versus Involved Production" in the Multi-feature/multi-dimensional model proposed by Biber (1988). Based on Nishimura's idea, how the POS is distributed across different domains will be explored here.

4.4.1 Method

The method follows the procedure shown below.

1) Rank the proportions of each POS in each genre based on the proportion of POS shown in Table 4-18.
2) Reorder the POS based on the indexicality for informality.
3) Reorder the genres based on the informality, in order to detect a pattern.
4) Classify the POS as Index for informality, Index for formality and Non-indexical.
5) Create graphs to show the pattern of POS ranking in different genres.
6) Sum up the proportions of POS use for each type of index, and examine how the total proportions of POS use for the informality index and formality index correlates with each other.
7) Conduct the hierarchical cluster analysis to classify the genres based on the distribution of POS to examine if the detected pattern agrees with the result of the abovementioned analysis.

4.4.2 Results and discussion

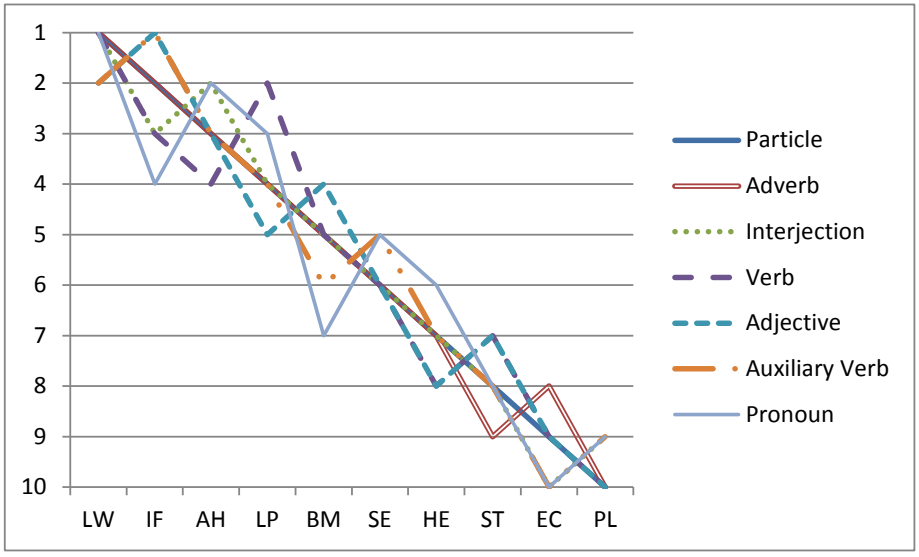
The results are shown in Table 4-19, 4-20, graphs from 4-5 to 4-7 and Figure 4-1. The results clearly demonstrate that particles, adverbs, interjections, verbs, adjectives, auxiliary verbs, and pronouns (indexicality order) can be the indices for informality or colloquiality (simply 'informality' tentatively), and suffixes, verbal nouns, conjunctions, and nouns (indexicality order) can be the indices for formality or literariness (simply 'formality' tentatively) (Table 4-19 and Graph 4-5 and 4-6). Prenoun adjectivals, signs, prefixes, and nominal adjectives do not show a clear pattern for indexicality (Graph 4-7).

Table 4-19 Ranking for the Use of Part of Speech in Each Genre in VDRJ (POS ordered by the indexicality for informality from the top, and genres ordered by informality from the left)

Part of Speech (Japanese)	Genre / Part of Speech	LW	IF	AH	LP	BM	SE	HE	ST	EC	PL
助詞	**Particle**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**
副詞	**Adverb**	**1**	**2**	**3**	**4**	**5**	**6**	**7**	**9**	**8**	**10**
感動詞	**Interjection**	**1**	**3**	**2**	**4**	**5**	**6**	**7**	**8**	**10**	**9**
動詞	**Verb**	**1**	**3**	**4**	**2**	**5**	**6**	**8**	**7**	**9**	**10**
形容詞	**Adjective**	**2**	**1**	**3**	**5**	**4**	**6**	**8**	**7**	**9**	**10**
助動詞	**Auxiliary Verb**	**2**	**1**	**3**	**4**	**6**	**5**	**7**	**8**	**10**	**9**
代名詞	**Pronoun**	**1**	**4**	**2**	**3**	**7**	**5**	**6**	**8**	**10**	**9**
連体詞	Prenoun Adjectival	3	10	2	1	9	5	4	6	8	7
記号・補助記号	Sign	7	1	5	3	8	9	2	4	6	10
接頭辞	Prefix	10	1	7	4	9	8	5	6	2	3
名容詞	Nominal Adjective	9	3	8	6	4	2	10	7	1	5
名詞	**Noun**	**9**	**10**	**6**	**7**	**5**	**8**	**1**	**2**	**4**	**3**
接続詞	**Conjunction**	**9**	**10**	**8**	**3**	**7**	**5**	**6**	**4**	**2**	**1**
動名詞	**Verbal Noun**	**10**	**7**	**9**	**8**	**5**	**4**	**6**	**3**	**1**	**2**
接尾辞	**Suffix**	**10**	**9**	**7**	**8**	**6**	**4**	**3**	**5**	**2**	**1**

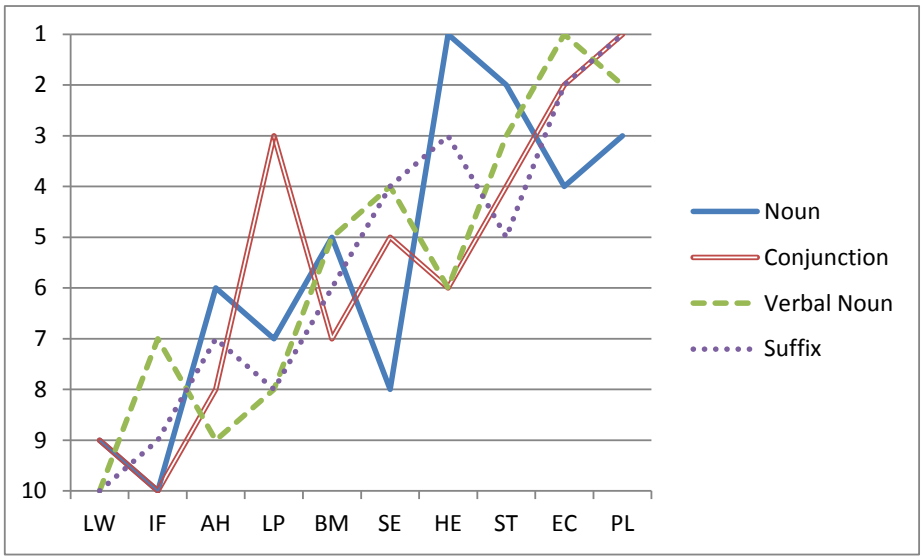
LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.

Graph 4-5 Ranking for the Use of the Indexical Part of Speech for Informality in Each Genre in VDRJ (POS ordered by the indexicality for informality from the top, and genres ordered by informality from the left)



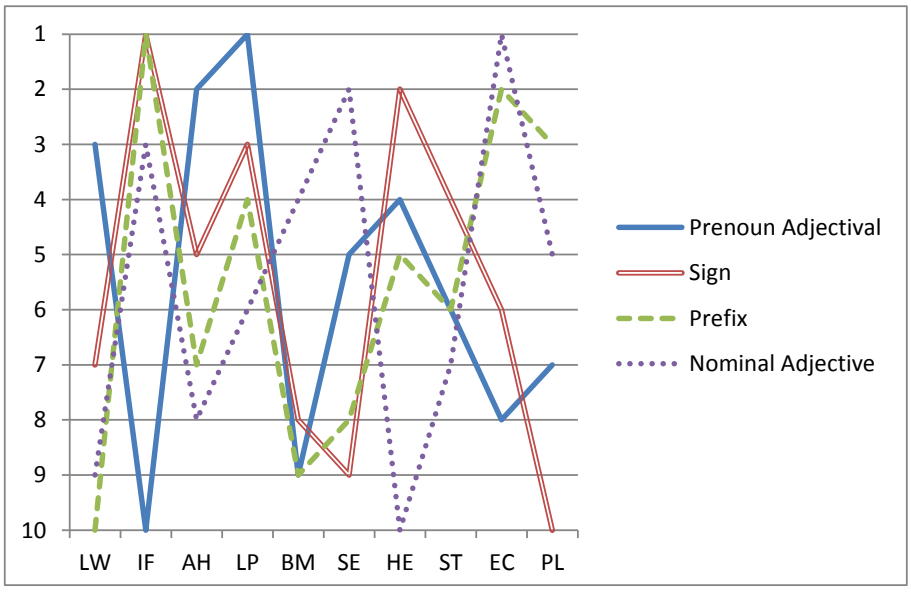
LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.

Graph 4-6 Ranking for the Use of the Indexical Part of Speech for Formality in Each Genre in VDRJ (POS ordered by the indexicality for informality from the top, and genres ordered by informality from the left)



LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.

Graph 4-7 Ranking for the Use of the Non-indexical Parts of Speech in Each Genre in VDRJ (POS ordered by the indexicality for informality from the top, and genres ordered by colloquiality from the left)



LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.

Among all POS, particles show the clearest indexicality for informality. Adverbs and interjections also show indexicality clearly even if they do not provide a high proportion. Indexicality for formality is not as clear as informality; however, suffixes and verbal nouns show a relatively clear disposition. As shown in Table 4-20, when we compare the rankings of Subtotal A of the seven POS for the informality index with Subtotal B of the four POS for the formality index, ascendant order of the former and descendant order of the latter totally agree with each other with no exceptions. The Subtotal A and the Subtotal B proportions show an extremely high reverse correlation at -.999 ($p < .001$) (Pearson's correlation coefficient).

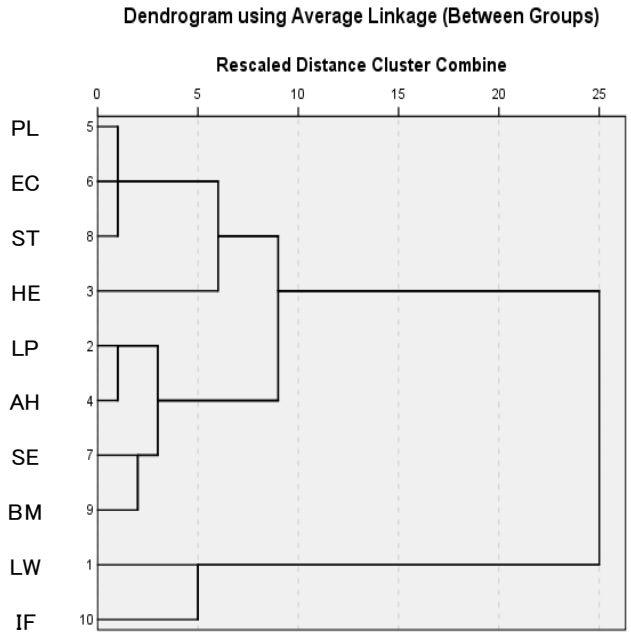
Table 4-20 Proportion of Indexical Sets of Parts of Speech at Each Genre in VDRJ (Counted by Tokens) (POS ordered by the indexicality for informality from the top, and genres ordered by informality from the left)

Part of Speech (Japanese)	Genre / Part of Speech	VDRJ Whole	LW	IF	AH	LP	BM	SE	HE	ST	EC	PL
助詞	Particle	31.6	33.3	32.2	31.9	31.7	31.1	31.0	30.1	29.6	29.6	29.4
副詞	Adverb	1.9	2.5	2.1	2.0	1.8	1.7	1.6	1.5	1.4	1.4	1.3
感動詞	Interjection	0.2	0.5	0.1	0.2	0.1	0.1	0.1	0.1	0.1	0.0	0.1
動詞	Verb	14.3	15.5	14.3	14.3	14.9	14.3	14.2	13.2	13.3	13.2	13.1
形容詞	Adjective	1.7	2.0	2.2	1.7	1.5	1.7	1.4	1.2	1.3	1.2	1.0
助動詞	Auxiliary Verb	11.0	12.6	14.3	10.4	10.2	9.5	9.5	9.2	8.9	8.4	8.6
代名詞	Pronoun	1.7	2.5	1.6	1.9	1.8	1.2	1.3	1.3	1.0	1.0	1.0
	Subtotal A	**62.3**	**68.8**	**66.9**	**62.4**	**61.9**	**59.5**	**59.1**	**56.6**	**55.5**	**54.8**	**54.4**
名詞	Noun	24.8	21.9	21.4	25.6	25.2	26.5	25.1	30.1	28.5	27.4	27.8
接続詞	Conjunction	0.4	0.3	0.2	0.4	0.6	0.5	0.5	0.5	0.5	0.6	0.6
動名詞	Verbal Noun	5.7	3.2	5.5	4.7	5.5	6.5	7.8	5.6	8.2	9.5	9.3
接尾辞	Suffix	3.4	2.7	2.8	3.5	3.1	3.7	4.0	4.1	3.9	4.2	4.4
	Subtotal B	**34.4**	**28.1**	**29.9**	**34.2**	**34.4**	**37.2**	**37.5**	**40.3**	**41.1**	**41.6**	**42.2**

Pearson's correlation of coefficient between the subtotal A and B is -.999 ($p < .001$).

Figure 4-1 Cluster Analysis of Proportion of Part of Speech in Genres in VDRJ

(Counted by Tokens) (Squared Euclidean distance, average linkage between groups)



LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.

The cluster analysis was also done by other linkage methods such as Ward's method or single linkage, but the classification patterns of the dendrograms basically appeared the same.

The results basically agree with Nishimura (2010); however, there is one point to consider. According to Nishimura (2010), verbs are thought to show features of the written language as they have a similar distribution pattern to nouns. Nevertheless, verbs have the opposite disposition in this study, but rather similar to Kabashima (1955, 1981) in terms of verbs. This issue is related the question: What do the indexical POS sets in this study represent? Do they represent informality or colloquiality? In Nishimura (2010), the proportion for verbs is the highest in written language, the second highest in online language, and the least in spoken language. However, looking closely at the proportions in the written language, for example, the order of the proportions for nouns and verbs in newspapers (nouns: 37.4%, verbs: 23.6%) is opposite to the order in printed novels (nouns: 12.8%, verbs: 23.6%). Both newspapers and novels are written language; however, newspapers will be more formal than novels. Besides the results with verbs, the other results with most other POS in this study are consistent with Nishimura (2010). Therefore,

verbs will be a tricky category in interpreting indexicality. The distinction and relationship between informality and colloquiality will be clearer by comparing the POS distribution of the formal spoken texts such as academic discussion.

Also, Nishimura (2010) points out that written texts have more case particles while the spoken texts have more sentence-final particles and adverbial particles. As mentioned above, the total proportion for all particles can be an index for informality; however, it can be a more powerful index if the case particles are excluded from the proportion.

Comparing the results with Kabashima's law (Kabashima, 1955, 1981), conjunctions show a different disposition. In Kabashima's law, conjunctions occur more in casual or literary texts while the results of this study show that conjunctions should be included in the index for formality. Kabashima analysed more genres such as conversation, Haiku and newspaper headlines while this study only covers books and internet texts; therefore, it is not appropriate to make an easy comparison. However, Kabashima combined conjunctions with interjections into a group. This grouping is worth reconsidering.

According to the results of the cluster analysis (Figure 4-1) and other results, it appears appropriate to classify the ten genres into four categories: 1) LW, 2) IF, 3) AH, LP, BM and SE, 4) HE, ST, PL and EC. The internal order within each cluster is a little different from the order of the proportions in this study; however, the result of the cluster analysis generally agrees with the order of the proportion rankings shown in Tables 4-19 and 4-20.

Comparing this result with the lexical homogeneity order and informality order in Table 4-5, 4-8, 4-13 and Graph 4-2 in 4.2, they largely agree with each other, i.e., the more informal, the more lexically diverse, and vice versa. However, HE shows very low lexical homogeneity but is more formal than more lexically homogeneous genres. IF also shows relatively high lexical homogeneity but is more informal than lexically more diverse genres.

Overall, the total proportions for the seven and the four POS show considerably

powerful indexicality to measure the informality/formality of genres.

4.4.3 Conclusion of 4.4

The main findings in this section are as follows.

1) The proportion(s) for the total tokens of particles, adverbs, interjections, verbs, adjectives, auxiliary verbs, and/or pronouns (indexicality order) can be the index for informality.
2) The proportion(s) for the total tokens of suffixes, verbal nouns, conjunctions, and/or nouns (indexicality order) can be the index for formality.
3) Based on the POS distribution and cluster analysis, the ten genres in VDRJ can be divided into four categories: 1) LW, 2) IF, 3) AH, LP, BM and SE, 4) HE, ST, PL and EC.
4) Generally, the more informal a genre is, the more lexically diverse it will be. Also, the more formal a genre is, the more lexically homogeneous will be.

These findings are not directly related to the main research question in this thesis; however, these also show lexical differences of genres in terms of formality and diversity. These suggest that different learning order of words will be efficient for different purposes. The more diverse the vocabulary in a genre, the heavier the learning burden in general.

4.5 Chinese-origin words and Chinese cognates

In the previous sections, register variations were explored by checking word origins and POS. In this section, Chinese cognates are checked in more details. Chinese origin words are largely Chinese cognates; yet, not all the Chinese-origin words are cognates. In addition, there are different types of cognates which will have different effect for Chinese-background learners on learning Japanese. These issues should be related to the amount of burden of learning vocabulary.

As discussed in 2.2.3 and 3.3.4.4 in Chapter 3, cognates or loanwords can be the Assumed Known Words for the users of the language which the cognate or loanword was derived or borrowed from. Therefore, it has been a long discussed issue what a curriculum should contain if Chinese-background learners (CBLs) and non-Chinese-background learners (non-CBLs) are mixed together in a Japanese language programme. As is widely known, CBLs have an advantage in lexical knowledge in reading Japanese as Japanese has many Chinese-origin words. Nevertheless, few researchers or teachers can accurately estimate or measure the gap between the two. What advantage in terms of number of words does a Chinese-background learner bring to the learning Japanese?

To get the clue for the answer to the above questions, the following sub-research-questions (SRQs) are set as the research questions in this section. (The SRQ number follows the previous section.)

SRQ 12) How many Chinese cognates are there in Japanese basic and intermediate vocabulary? How are they distributed at different frequency levels?

SRQ 13) Is the number and proportion of Chinese cognates in BCCWJ made from books and internet-forum sites similar to those from magazines or other types of texts?

The number and proportion of Chinese cognates at basic and intermediate levels in VDRJ will be calculated, and the gap in learning burden between CBLs and non-CBLs will be estimated and discussed at the end of this section.

4.5.1 Issues with Kanji vocabulary and Chinese cognates in Japanese

Firstly, the definitions of related terms must be clear. As discussed in 2.3 and 3.3.4.3.2 in Chapter 3, the lexical synchronic relationship between Chinese and Japanese is fairly complicated, mainly because the two languages share Kanji, so-called logographic

characters, which share the orthography but do not always share phonological information. In this section, the classification of the categories for related words follows inTable 4-21 (=Table 3-9).

Table 4-21 Categories for Interlingual Form-related Words between Chinese and Japanese (=Table 3-9)

Category	Phonological form	Orthographical form	Meaning	Examples (Japanese/Chinese)
Cognate	related	same/similar	same/related	"gakushuu"/"xue2 xi2" 学習/学习 (learning) "goudou"/"he2 tong2" 合同 (combined/contract)
Partial-cognate compound	related with each component	same/similar for each component	related/ different	"taisetsu"/"da4-qie4" 大切/大−切 （important/big-cut）
Interlingual-written cognate	different	same/similar	same/related/different	"kuda"/"guan3" 管 (tube) "baai"/"chang3 he2" 場合/场合 (case) "ugoku"/"dong4" 動く/动 (move)
Interlingual-written-partial-cognate compound	different	same/similar for each component	related/ different	"tokei"/"shi2-ji4" 時計/时−计 (clock/time-measure)

In contrastive studies between Chinese and Japanese, cognates are often called 同形語 ‘doukei-go’ which literally means ‘same-form word’ where ‘form’ only refers to the orthographical form. In the classification shown in Table 4-21, ‘doukei-go’ corresponds to ‘cognate’ or ‘interlingual-written cognate’. These two kinds of words orthographically ‘exist’ in both Chinese and Japanese. Among the four categories, the top two (‘cognate’ and ‘partial-cognate compound’) are Chinese-origin words. The bottom two (‘Interlingual-written’ cognate or partial-cognate compounds) are Japanese-origin words (at least phonologically) but share Kanji which may link sematic representations between Chinese and Japanese in the language user’s knowledge system. Partial-cognate compounds are not ‘doukei-go’. In other words, these types of words do not ‘exist’ in Chinese, but each individual Kanji exists in the both languages.

In any categories, the difference in character form (字体 ‘jitai’) between the two languages is not taken into account. In other words, whatever forms are used in the two corresponding Kanji forms in Chinese and Japanese, they are regarded as the same

characters if they share the same traditional form (so-called 康熙字典体 'Kouki-jiten-tai' (the Kangxi dictionary form)). In both Japan and the People's Republic of China (mainland China), the form of Kanji was simplified after World War II. Some of them were simplified in the same way in both countries, but some were not. In Chinese communities outside mainland China (including Hong Kong that became a part of the People's Republic of China in 1997), they still use the traditional form. In consequence, three types of Chinese character systems have five types of correspondence patterns of the character forms in three kinds of areas (Table 4-22). Approximately half of the 2,136 common Japanese Kanji have the same form as Chinese used in the Mainland China (MS) (Hishinuma, 1984, p 35) and the others have a different form.

Table 4-22 Example Characters for the Five Correspondence Patterns of Chinese Character Forms in the Three Areas

Area	Taiwan, Hong Kong, Macau, other Chinese communites in the world, Korea	Mainland China, Singapore	Japan
Code / Correspondence Pattern	TC (Traditional Chinese Communities)	MS (Mainland China and Singapore)	JP (Japan)
TC = MS = JP	**我**	**我**	**我**
TC = MS ≠ JP	**黑**	**黑**	黒
TC ≠ MS = JP	會	**会**	**会**
TC = JP ≠ MS	**書**	书	**書**
TC ≠ MS ≠ JP	發	发	発

Nevertheless, the orthographic difference will not have a marked effect on processing Kanji by CBLs of Japanese in general so that different simplified forms are linked in users' knowledge in general as sharing the same traditional forms (Kayamoto, 2000; Tamaoka & Matsushita, 1999). This is the reason why the difference in character form between the two languages is not taken into account for this section. There are some phonological effects on processing Kanji so that the pronunciation of some types of Kanji words are easy or difficult to learn; however, the effects are limited, and will generally vanish at the super-advanced level (Kayamoto, 2000; Tamaoka et al., 1999).

The most essential problem in processing Japanese Kanji by CBLs is the semantic effect (Kayamoto, 2002; Tamaoka et al., 1999). If the basic meaning and usage is the same as the corresponding Chinese word, it will be processed more quickly and correctly in general. There are various factors involving semantic processing of Kanji such as orthographical or phonological similarity (Kayamoto, 2002), frequency of usage (Chen, 2009) and prototypicality of the meaning (Kato, 2005). This study does not further discuss this issue but confirms here that the semantic effect is the most influential on processing Kanji by CBLs.

The research question here is: "How many Chinese cognates are there in Japanese basic and intermediate vocabulary?" There are studies on the quantitative status of Chinese cognates in both Chinese and Japanese. Araya (1983) used dictionaries to decide that approximately 50% of 3,800 common Chinese words are Chinese cognates in Japanese. This figure includes Japanese-origin words (i.e. 'interlingual-written cognates' in this study) whether the word has inflected suffixes or not. For example, 進む 'susumu' (progress) is identified as a cognate of the Chinese word〈进〉/jin4/ (enter). Sone (1988) used a Chinese word frequency list and first identified 6,112 words which are the remainder after excluding one-syllable words from the top 8,441 words. And then, he used a Japanese dictionary to identify 56% of the 6,112 words as Chinese cognates. Takano & Wang (2002) compared the Chinese word frequency list and the word list made from Japanese high-school textbooks. They identified 33% of the most frequent 3,000 Chinese words as Chinese cognates in Japanese. These studies are aimed at Japanese learners of Chinese.

Takano & Wang (2002) also tried to locate the Chinese cognates in Japanese vocabulary. They identified 41% of the most frequent 3,000 words in Japanese high-school textbooks as Chinese cognates. Matsushita (2009) identified 38% of the most frequent 5,022 words in magazines as Chinese cognates, and 41% of the most frequent 3,000 words in magazines as Chinese cognates, which is almost the same figure as Takano & Wang's

(2002).

Nevertheless, these results are still questionable as the corpus domains are textbooks or magazines, and the corpus sizes may not large enough, either. In this section, the number and proportion for Chinese cognates are calculated based on the frequency counts in books and internet-forum sites in BCCWJ 2009 monitor version which has 33 million tokens. And then, how many of them share the basic meaning and usage with correspondent Chinese word will be discussed based on previous studies. Lastly, the gap in learning burden between CBLs and non-CBLs will be estimated.

4.5.2 Method

Frequency lists

VDRJ is used for checking the distribution in books and internet texts. For magazine texts, the data is cited from Matsushita (2009) where the distribution is calculated from the Vocabulary Lists from the Language Survey of Contemporary Magazines with Two Million Running Characters (NLRI, 2006). This list is created from 1.06 million tokens (including 0.73 million tokens of content words) from 70 types of magazines published in 1994.

Identifying the standard orthography

Even for Chinese-origin words, if they are more frequently written in Kana (syllabic character) rather than in Kanji (e.g. たぶん（多分） 'tabun' (probably) or けんか（喧嘩） 'kenka' (quarrel/fight)), the Kana orthography is recognized as standard; the words are not identified as Chinese cognates.

Identifying Chinese-origin words

In VDRJ, Chinese-origin words are identified by the dictionary UniDic (Den et al.,

2009). In NLRI (2006), Chinese-origin words are identified by the tagged information on the list.

Identifying Chinese cognates

In VDRJ, Chinese cognates identified in Matsushita (2009) are all identified as Chinese cognates as well. For the other words, Chinese cognates are identified through discussion by two people. One of the two is the author of this thesis, and the other is a native Chinese Japanese-Chinese translator who has occupied the job for over ten years. For the words in NLRI (2006), identifying Chinese cognates basically follows Matsushita (2009); however, judgments for some words are modified by the two experts mentioned above. In Matsushita (2009), words adopted in A Word Frequency Dictionary for Modern Chinese (BLI, 1986) are all identified as Chinese cognates first, and then the other words are judged by three people. One of the three is the author of this thesis, and the other two are native Chinese postgraduate students majoring in Japanese in a Japanese university.

As mentioned in 4.5.1, different character forms are not taken into account, i.e. words which share the same traditional form (the Kangxi dictionary form) are identified as the same word. For example, 経済 (Japanese form) is identified as the cognate of 经济 (Chinese form in Mainland China and Singapore) because the two forms share the same traditional form 經濟.

Chinese cognates for this study are limited to Chinese-origin words. In other words, "Interlingual-logographic cognates" shown in Table 4-21 are not counted as Chinese cognates. For example, the word 場合 'baai' (case) is not counted as a Chinese cognate since it does not have any phonological relationship with the corresponding Chinese word 场合 /chang3he2/ (case).

Some other tricky cases were judged in the following ways. As mentioned above, words such as たぶん（多分） and 場合 are not identified as Chinese cognates because

たぶん is written in Kana more frequently than in Kanji, and 場合 is not a Chinese-origin word. The Japanese words 編集 and 種々 are counted as Chinese cognates as they are popular forms of 编辑 and 种种 in Chinese even if 編集 and 種々 are not canonical Chinese forms. The word 業者 is also counted as a Chinese cognate as it was introduced from Japanese and is currently used fairly frequently in China. The words 我慢 'gaman' (endurance) and 完了 'kanryou' (completion) are two words (i.e. 我慢 'wo3 man4' (I'm slow), 完了 'wan2 le' (finished)) in Chinese but one word in Japanese; however, they are all counted as Chinese cognates. Affixes such as -徒 are also counted as Chinese cognates if they are also used in Chinese as an affix or a word.

4.5.3 Results

The proportions of word origins (counted by lexemes) in VDRJ and magazine texts are shown in Table 4-23 and 4-24.

Table 4-23 Numbers and Proportions of Content Words by Word Origin at each 1000 Word Level of the Most Frequent 5000 Content Word in VDRJ (Book and Internet-Forum Texts) (Counted by Lexemes)

Word Level	Word Ranking	Whole	Chinese -origin	Western-origin and Others	Japanese -origin	Mixed-origin	Proper Nouns and Unknown	Whole	Chinese -origin	Western-origin and Others	Japanese -origin	Mixed-origin	Proper Nouns and Unknown
-1,000	0,001-1,000	1,000	449	13	497	25	16	100.0	**44.9**	1.3	49.7	2.5	1.6
-2,000	1,001-2,000	1,000	538	52	371	22	17	100.0	**53.8**	5.2	37.1	2.2	1.7
-3,000	2,001-3,000	1,000	505	83	363	17	32	100.0	**50.5**	8.3	36.3	1.7	3.2
-4,000	3,001-4,000	1,000	518	90	336	16	40	100.0	**51.8**	9.0	33.6	1.6	4.0
-5,000	4,001-5,000	1,000	501	104	322	25	48	100.0	**50.1**	10.4	32.2	2.5	4.8
Whole	0,001-5,000	5,000	2,511	342	1,889	105	153	100.0	**50.2**	6.8	37.8	2.1	3.1

* Function words are excluded. Words are ranked by U which is a product of frequency and dispersion (Juilland & Chang-Rodrigues1964).

Table 4-24 Numbers and Proportions of Content Words by Word Origin at each 1000 Word Level of the Most Frequent 5000 Content Word in Magazine Texts (NLRI, 2006) (Counted by Lexemes)

Level		Number of Words						Proportion (Percentage)					
Word Level	Word Ranking	Whole	Chinese -origin	Western-origin and Others	Japanese -origin	Mixed-origin	Proper Nouns and Unknown	Whole	Chinese -origin	Western-origin and Others	Japanese -origin	Mixed-origin	Proper Nouns and Unknown
-1,000	0,001-992	1,002	461	110	389	16	26	100.0	**46.0**	11.0	38.8	1.6	2.6
-2,000	1,003-1,964	999	452	150	339	14	44	100.0	**45.2**	15.0	33.9	1.4	4.4
-3,000	2,002-2,955	1,027	450	204	280	26	67	100.0	**43.8**	19.9	27.3	2.5	6.5
-4,000	3,029-3,903	1,034	416	245	270	24	79	100.0	**40.2**	23.7	26.1	2.3	7.6
-5,000	4,063-4,794	960	397	216	235	20	92	100.0	**41.4**	22.5	24.5	2.1	9.6
全体	0,001-4,794	5,022	2,176	925	1,513	100	308	100.0	**43.3**	18.4	30.1	2.0	6.1

* Ranking and number of words do not agree with each other as some words are at the same ranking.

Magazines contain many Western-origin words in advertisements so that the proportion of word origins is not normal. Magazine data such as NLRI (1964) are often cited as the general Japanese data; however, magazines cannot represent the general proportion of word origins.

Table 4-25 Ratios for Chinese-origin Words and Chinese Cognates at Each 1000 Word Level of the Most Frequent 5000 Content Words in VDRJ (Book and Internet-forum Texts) (Counted by Lexemes)

Level		Number of Words		Number/Ratio for Chinese Cognates		
Word Level	Word Ranking	Whole	Chinese-origin	Chinese Cognates	Ratio to Chinese-origin	Ratio to the Whole
-1,000	0,001-1,000	1,000	449	423	**94.2%**	**42.3%**
-2,000	1,001-2,000	1,000	538	495	**92.0%**	**49.5%**
-3,000	2,001-3,000	1,000	505	433	**85.7%**	**43.3%**
-4,000	3,001-4,000	1,000	518	428	**82.6%**	**42.8%**
-5,000	4,001-5,000	1,000	501	373	**74.5%**	**37.3%**
Whole	0,001-5,000	5,000	2,511	2,152	**85.7%**	**43.0%**

* Function words are excluded. Words are ranked by U which is a product of frequency and dispersion (Juilland & Chang-Rodrigues, 1964).

Table 4-26 Ratios for Chinese-origin Words and Chinese Cognates to the Most Frequent 5000 Content Words in Magazine Texts (NLRI, 2006) (Counted by Lexemes)

Level		Number of Words		Number/Ratio for Chinese Cognates		
Word Level	Word Ranking	Whole	Chinese-origin	Chinese Cognates	Ratio to Chinese-origin	Ratio to the Whole
-1,000	0,001-992	1,002	461	419	**90.9%**	**41.8%**
-2,000	1,003-1,964	999	452	414	**91.6%**	**41.4%**
-3,000	2,002-2,955	1,027	450	386	**85.8%**	**37.6%**
-4,000	3,029-3,903	1,034	415	343	**82.7%**	**33.2%**
-5,000	4,063-4,794	960	397	325	**81.9%**	**33.9%**
Whole	0,001-4,794	5,022	2,176	1,887	**86.7%**	**37.6%**

* Ranking and number of words do not agree with each other as some words are at the same ranking.

80-90% of Chinese-origin words are Chinese cognates in both VDRJ and magazines. Cognates are more at the top 2000 than at the lower level in both VDRJ and magazines, where over 90% of Chinese-origin words are cognates. The proportion for the cognates gets lower by degrees as the word level gets lower. Chinese cognates occupy approximately 40% of the all top 5000 words (43% in VDRJ and 38% in magazines, counted by lexemes). These figures are slightly more than the ones in Takano & Wang (2002) where a different type of corpus was used and the way of identifying cognates might also be different.

4.5.4 Discussion

Cognates will not always have the same meaning and usage as the original word. There have been some attempts to count how many Chinese cognates have the same meaning as the original. Takano & Wang (2002) identified 84% of the Chinese cognates as having the same meaning. Sone (1988) identified 73% of the most frequent 313 Chinese cognates as having the same meaning. Roughly three quarters of Chinese cognates at the basic level are estimated to have the same meaning as the original word.

Some words may have stylistic differences from the original word even if they have the same basic meaning; however, Matsushita (2009) claims that the stylistic gap is not

large because the frequencies of Chinese cognates and the original words have a correlation at .336 (Pearson, $p < .01$). Matsushita (2009) also identifies 67% of the character forms of Chinese cognates as the same and 23% as similar, and concludes that the differences of character forms are not problematic except for some tricky ones. Tamaoka et al. (1999) also claim that the orthographic difference has little impact on processing Japanese Kanji by advanced Chinese learners. Matsushita (2009) also claims that phonological similarity between Chinese cognates and the original words are not high in general, as the mean similarity point of the Chinese cognates is 2.60 out of 7 (SD = 1.14), which is calculated based on Kayamoto's (1995) seven-point-scale of phonological similarity judgement data. This suggests that Chinese-background learners (CBLs) will not have an advantage in learning pronunciation of Chinese cognates. CBLs will surely have the advantage in learning orthography (Kanji) even for non-cognates (Matsushita, Taft, & Tamaoka, 2004). According to these studies, the advantage will be primarily limited to understanding the meaning through reading, and learning to write Kanji.

Based on these data, now let us estimate the gap in learning burden between CBLs and non-CBLs. This is basically the same as answering to the question: "How large is the advantage for CBLs in learning Japanese vocabulary? Chinese-origin words occupy over 40% in the top 5,000 words in magazine texts and over 50% in the top 5,000 words in VDRJ. In both lists, Chinese cognates are 80 -90% in the Chinese-origin words, i.e. approximately 40% of the top 5,000 words. Three quarters of the Chinese cognates have the same meaning and orthography. That means around 30% of the top 5,000 words i.e. 1,500 words (counted by lexemes) can be the Assumed Known Words for CBLs, the words which they can understand by exploiting knowledge of the Chinese language. As the Chinese first language (L1) knowledge will automatically be activated when they see the cognates, the knowledge has to be inhibited if the meaning or usage is different.

How long a learning time can the gap be converted into? If a learner can learn 25-50

words per week, s/he can learn 1,000 to 2,000 words in 40 weeks which is the standard length of time for Japanese language institutes in Japan. Many of them have a curriculum where the learners are expected to finish the intermediate course within a year; however, this curriculum will only match with CBLs as the learners are expected to learn around 100 words per week to finish the intermediate course. This would be extremely hard for learners unless the learners have a certain level of previous knowledge as CBLs have. Non-CBLs will generally require one more year (i.e. two years in total) to finish the intermediate course even if they learn Japanese on a full-time basis.

As mentioned above, the gap mainly and primarily exists in reading and writing but not listening and speaking. In elementary courses, learners generally learn conversation mainly, therefore, a different curriculum will be more required at the intermediate level or above. The advantage in learning written vocabulary may also be a disadvantage in acquiring grammar and conversation (Hatasa, 1992). If a Japanese language program has both CBLs and non-CBLs, a double-tracked curriculum or selective modules should be introduced from the elementary level. This claim is merely based on a rough estimation from a quantitative contrastive analysis. This issue should be further explored with tests.

There are some remaining issues. First, the proportions for Chinese cognates should be further investigated up to super-advanced level over the top 5,000 words. Second, how Chinese knowledge has an impact on learning non-cognate Chinese-origin words (i.e. "cognatic compounds" in Table 4-21) and Japanese-origin words written in Kanji (i.e. "interlingual-logographic cognates" and "interlingual-logographic cognatic compounds" in Table 4-21) should also be explored. Third, how English knowledge has an impact on learning European-origin words should also be investigated. Daulton (2004) investigates how Japanese knowledge of Gairaigo (English-origin words) has an impact on learning high-frequency English vocabulary and concludes that they have a positive impact. Conversely, the knowledge of English will also be expected to have a positive impact on

learning English-origin words if the learner is able to read Katakana and process the phonological information from it to detect what the original English word is. Fourthly, Kanji vocabulary should be further classified based on different types of correspondence pattern of meanings and usages. The proportions of them should be calculated based on the classification. Fifthly, the learning burden should be calculated based on more detailed data. For example, how many words can be learned in a week should be examined in different contexts. Last but not least, a double-tracked curriculum and/or selective modules for learners with different language backgrounds should be developed as specific measures to deal with the gap in "built-in" knowledge.

4.5.5 Conclusion of 4.5

The main findings in this section include:

1) Chinese-origin words occupy half of the top 5,000 content words in VDRJ (counted by lexemes).
2) 80-90% of Chinese-origin words are Chinese cognates in both VDRJ and magazine texts.
3) Chinese cognates are approximately 40% of the top 5,000 content words (counted by lexemes).
4) Approximately 30% of the top 5,000 words (i.e. 1500 words) are expected to be Assumed Known Words, words which do not require previous second language learning) for Chinese-background learners (CBLs).
5) Non-CBLs will require approximately one more year learning on a full-time basis than CBLs to complete an intermediate course in a Japanese program. Therefore, a double-tracked curriculum or selective modules may be required from the elementary level to fill the gap in lexical knowledge between CBLs and non-CBLs.

4.6 Conclusion of Chapter 4

In this chapter, based on a new corpus, with a newly developed morphological

analyser and dictionary, statistical features of written Japanese were surveyed mainly from the viewpoints of lexical homogeneity (text coverage) and informality/formality (word origins, part of speech). This is a study to explore how different media and genres make differences in the efficient learning order of words as well as in understanding the features of Japanese vocabulary in general.

In 4.2, we have found that books are less biased compared to magazines and newspapers. In other words, lexical features of books are considerably diverse from genre to genre. As shown in Table 4-13, in the high-frequency band, the more homogeneous the vocabulary is, the more informal the genre will be; however, in the low-frequency band and on the whole, the relationship is reversed. In 4.3, the distribution of POS was surveyed from both lexeme and token counts. The significance of Chinese-origin suffixes and verbal nouns were pointed out as notable results. In 4.4, based on the results of 4.3, the POS distribution has been shown to have strong indexicality of informality/formality to identify register variations on a continuum. In every genre in VDRJ, the more the proportion for the seven POS including particles and adverbs, the less the proportion for the four POS including suffixes and verbal nouns, and vice versa. This relationship is evident and robust. In 4.5, the number and proportion of assumed known Chinese cognates for Chinese-background learners (CBLs) are estimated to be 30% of the most frequent 5,000 words (i.e. 1,500 words). This amount of vocabulary generally requires one-year or more learning in full-time mode. The amount will be larger in the domains which contain more Chinese-origin words. To read academic texts or newspapers, a more efficient order and methods would be particularly necessary for non-CBLs.

These results come from a newly developed vocabulary database. In the next chapter (Chapter 5), character database will be developed. The relationship between words and characters in Japanese will be explored in Chapter 6.

Chapter 5 Making and validating the Character Database of Japanese

5.1 Introduction

To think about the efficient learning order of words, one particular issue with Japanese vocabulary is the relationship between the characters and the words. It would be best to learn Kanji used in high-frequency words in order of the degree of the learner's need; however, many factors need to be considered regarding order of learning Kanji (Kano, 1994). Generally, learners first learn easily recognized Kanji with only a few strokes such as 山 'yama' (mountain) or 川 'kawa' (river), or Kanji which are components (e.g. 木 'ki' (tree)) of other Kanji (e.g. 森 'mori' (woods) , 板 'ita' (board)). Frequent Kanji are not always easier than less frequent ones. If the Kanji 特 'toku' (special) is more frequent than 牛 'ushi/gyuu' (cow/bull) or 寺 'tera/ji' (temple), either of which is a component of the Kanji 特, which Kanji should learners learn first?

One idea is that that learner should learn frequent ones first even if they are difficult. Even if we admit the claim, the word frequency order may not agree with the Kanji frequency order. In other words, some Kanji not used for high-frequency words may be an important Kanji if it is the component for many other low/middle-frequency words. Contrary to that, a Kanji used for a high-frequency word may not be very important if it is not used for any other words. In addition, some Kanji are considerably complicated in form, and many Kanji can form many words as a component. Therefore, the order of Kanji and the order of vocabulary may need to be separately considered. In order to investigate the issue, a good Kanji frequency list is essential.

In Chapter 3 and 4, the vocabulary database (VDRJ) was created and the statistical features of Japanese vocabulary were examined by exploiting the database. Nevertheless, no matter whether a word is written in Kanji or in Kana, the word will be counted as one lexeme in VDRJ. Given this, a character frequency list based on the orthographic form (書字形 'shoji-kei') is necessary. In this chapter, the issues with creating a character frequency

list and related problems are explored first. The abovementioned issue with the gap between word frequency and character frequency will be examined in Chapter 6 by exploiting the character database to be made in this chapter.

5.2 Significant research

5.2.1 Problems with existing Japanese character lists

Existing character frequency lists have similar problems to the word frequency list problems mentioned in 3.2.1., i.e. corpus size, sub-frequencies, age and representativeness. NLRI (1963, 1976), which are lists made from magazine and newspaper texts respectively, are outdated and not based on large corpora (Chikamatsu, Yokoyama, Nozaki, Long, & Fukuda, 2000). Their representativeness is also questionable as they do not contain any book and internet texts. The 4th edition database for the 1,945 basic Japanese kanji (Tamaoka, 2004) is a very informative and convenient Kanji database since it provides various types of information and is provided in Excel format on the web; however, this database only contains 1945 kanji listed in the former common Japanese Kanji list (常用漢字表 'jouyou-kanji-hyou') which was published in 1981, revised in 2010 and currently lists 2,136 Kanji (Agency for Cultural Affairs, 2010). Also, the frequency data in Tamaoka (2004) are all from newspapers (Amano & Kondo, 2000; NLRI, 1976; Yokoyama, Sasahara, Nozaki, & Long, 1998). Amano & Kondo (2000) and Chikamatsu et al. (2000) are based on corpora which are relatively new and large enough with over tens of millions of character/word tokens; however, both of them are based only on newspaper corpora. Besides, all of the abovementioned lists do not have appropriate sub-frequency data which enable us to compute dispersion and adjusted frequency measures. (NLRI (1963, 1976) have sub-frequencies; however, they are not provided a digitized form.)

The Kanji list for the former Japanese Language Proficiency Test (Japan Foundation & Association of International Education, Japan, 2002) is an influential list at educational institutes; however, it just shows the level of the Kanji out of the four levels but does not

show the frequency data itself. The levelling was done by so-called experts; however, the levelling process and criteria are not clear.

For these reasons, it is necessary to develop a new character list based on a large corpus containing a wide range of genres which provide sub-frequencies.

5.2.2 Research questions

A new Japanese character database entitled the Character Database of Japanese (CDJ) will be created through the process shown in the following sections. The main research questions (MRQs) are repeated below.

MRQs: In what order should learners of Japanese as a second language learn words and characters in order to be able to read Japanese? How will the order vary according to the purpose of learning?

The sub-research-questions (SRQs) in this chapter are as follows. (The SRQ number follows the previous section.)

SRQ 14) How can a Japanese character database and character lists be created to identify target characters for learners at different levels of proficiency?

SRQ 15) How well do the rankings for Kanji in CDJ correspond to or are correlated with the ones in other lists such as the former Japanese Language Proficiency Test (F-JLPT) Kanji lists, the Japanese primary school Kanji grades (学年配当 'gakunen-haitou') or the lists made from newspapers?

SRQ 16) Are the newly created word lists more valid than the existing ones?

When the vocabulary database for this study (VDRJ) was created in Chapter 3, the most

appropriate index for ranking words was investigated along with the appropriate sub-frequency weighting. For creating CDJ in this chapter, this is not posted as a research question since the same ranking index *U* (Juilland & Chang-Rodrigues, 1964) and the same weighting system as used for VDRJ are adopted for the same reasons as discussed in Chapter 3; however, a statistical check will be carried out in a similar way to Chapter 3 in order to confirm that the weighting is appropriate for providing the better character rankings for different types of learners.

5.3 Method

The method for creating the character database, as in Chapter 3, basically follows Nation & Webb's (2011) six steps (p. 135-144; Table 3-1 in this thesis); however, some of them should only be applied to making word lists but not to character lists. To summarize, Nation & Webb's steps are 1) research question or reason, 2) unit of counting, 3) corpus, 4) criteria for counting words and separate lists, 5) criteria for ordering words and 6) cross-checking the list.

1) The research question for this chapter is already stated in the previous section. The target users, which need to be clarified to identify the research question, are researchers, teachers and learners of Japanese, which are the same as VDRJ. The database (CDJ) is for researchers and teachers, and the character lists derived from the database are for learners including "general" learners and international students in Japanese universities[68].

2) The unit of counting for CDJ is the individual character including some signs such as 々 which is an indicator for repeating the previous Kanji. Before analysing the texts using AntWordProfiler (Anthony, 2009), a space is inserted between characters using a macro programme created on the text editor (Sakura editor). By doing so, each individual character can be treated as a unit in AntWordProfiler.

[68] For more details about the target users, see 3.3.1.

Table 5-1 Numbers of Types and Tokens of Characters by Field in CDJ *The corpus is made from books and internet forum-sites contained in NINJAL (2009).

Field	Code for the ten domains	G (General)		T (Technical)		Total	
		G Type	G Token	T Type	T Token	Type	Token
Literary Works/Imaginative Texts	LW	**5,304**	**13,507,821**	--	--	**5,304**	**13,507,821**
Humanities and Arts							
Languages and Linguistics	LP	3,438	666,901	2,081	164,031	3,600	830,932
Philosophy and Religion		4,166	2,441,115	2,321	205,203	4,254	2,646,318
History	HE	4,685	3,326,400	2,844	215,990	4,827	3,542,390
Ethnology		4,033	1,755,978	1,434	30,848	4,072	1,786,826
Fine Arts	AH	3,892	1,606,216	1,809	65,294	3,955	1,671,510
Literature (G=Literary works=Imaginative texts)		--	--	1,942	60,075	1,959	60,075
Other Humanities and Arts		4,658	3,210,243	568	5,483	4,685	3,215,726
The Whole of Humanities and Arts		**5,862**	**13,006,853**	**3,593**	**746,924**	**5,967**	**13,753,777**
Social Sciences							
Politics	PL	3,341	1,493,296	2,176	183,890	3,442	1,677,186
Law		2,785	803,086	2,252	511,590	2,982	1,314,676
Economics	EC	2,849	1,107,191	2,378	587,164	3,050	1,694,355
Commerce and Business		2,910	1,409,071	2,072	520,212	3,006	1,929,283
Sociology and Social Issues	SE	3,442	2,151,727	2,432	537,539	3,537	2,689,266
Education		2,922	1,019,728	2,200	424,441	3,036	1,444,169
Other Social Matters		2,919	688,367	1,520	59,071	2,962	747,438
The Whole of Social Sciences		**4,300**	**8,672,466**	**3,273**	**2,823,907**	**4,414**	**11,496,373**
Technological Natural Sciences							
Mathematics	ST	1,429	65,235	951	31,904	1,549	97,139
Physics		1,127	40,951	802	14,952	1,257	55,903
Astronomy, Earth and Planetary Science		2,170	164,043	1,303	33,365	2,285	197,408
Chemistry, Metal and Mine		1,787	61,754	1,121	38,012	1,916	99,766
Technology (Architecture, Civil Engineering)		2,689	499,353	2,045	176,911	2,837	676,264
Technology (Mechanics, Electricity, Marine Engineering)		2,356	328,477	1,562	120,951	2,476	449,428
Other Technological Natural Sciences		2,860	670,041	1,950	252,460	2,984	922,501
The Whole of Technological Natural Sciences		**3,481**	**1,829,854**	**2,566**	**668,555**	**3,592**	**2,498,409**
Biological Natural Science							
Biology	BM	2,677	434,890	1,611	66,511	3,600	501,401
Agriculture		2,598	392,516	1,368	46,480	2,653	438,996
Pharmacy		1,579	40,651	815	15,697	1,658	56,348
Medicine		2,743	798,212	1,754	136,905	2,813	935,117
Dentistry		1,006	19,286	679	6,326	1,162	25,612
Nursing		1,209	31,301	1,183	37,931	1,484	69,232
Other Biological Natural Sciences		3,233	1,585,283	2,004	121,128	3,320	1,706,411
The Whole of Biological Natural Science		**4,144**	**3,302,139**	**2,731**	**430,978**	**3,783**	**3,733,117**
Internet Q & A Forum (Yahoo Chiebukuro)	IF	**3,652**	**8,701,058**	--	--	**3,652**	**8,701,058**
The Whole of CDJ		**6,549**	**49,020,191**	**4,138**	**4,670,364**	**6,630**	**53,690,555**

Note 1: Published books and library books are added together.
Note 2: The figures contain number of signs. No additional processing was made for extracting noises.
Note 3: If the C-code of a text is 3,000-3,999, it is counted as a technical text.

3) The Corpus used for making CDJ is the BCCWJ 2009 monitor version, which is the same as the corpus used for making VDRJ. The sub-sections of the corpus are also the same as VDRJ. (For the details of the construction of the corpus, see 3.3.2.) The number of types and tokens of the characters in CDJ are shown in Table 5-1.

Comparing Table 5-1 with Table 3-4 which shows the number of types and tokens by field in VDRJ, the overall distributions seem to be similar. The average number of characters for a token in VDRJ can be calculated by dividing the total number of tokens by the total number of characters. The result is 1.64. This figure shows the average of the actually used words, that is, weighted more on high-frequency words. Calculating the mean length of all 141,950 lexemes from the column 'number of characters' in VDRJ, the result is 4.01 (SD = 2.34). When limiting the target to the top 20,000 words in WWJ, the Word Ranking for Written Japanese, the results are M=2.54, SD=1.28. When the target is limited to the top 5,000 words, the results are M=2.24 and SD=1.10. These figures mean, not surprisingly, that the higher the word frequency, the shorter the word length.

The numbers and proportions of character tokens in CDJ are shown in Table 5-2.

Table 5-2 Numbers and Proportion of Character Tokens by the Ten Domain Classification in CDJ *The corpus is made from books and internet-forum sites contained in NINJAL (2009).

Domain	Code for the ten domains	Number of Tokens	Proportion
Literary Works/Imaginative Texts	LW	13,507,821	25.2%
Languages, Linguistics and Philosophy	LP	3,477,250	6.5%
History and Ethnology	HE	5,329,216	9.9%
Arts and Other Humanities	AH	4,947,311	9.2%
Politics and Law	PL	2,991,862	5.6%
Economics and Commerce	EC	3,623,638	6.7%
Sociology, Education and Other Social Issues	SE	4,880,873	9.1%
Science and Technology	ST	2,498,409	4.7%
Biology and Medicine	BM	3,733,117	7.0%
Internet Q & A Forum	IF	8,701,058	16.2%
Total		53,690,555	100.0%

Comparing the proportion of character by domain (Table 5-2) with the proportion of tokens (Table 3-5), the proportions are considerably similar. This is not surprising because the average length of tokens will not be so different according to genres.

4) Criteria for counting words and separate lists for marginal words are particular issues with words. Unlike the unit of the word, one unit of a character can be clearly identified. All the character data can be in one file. For the user's convenience, the data for Kana, Roman alphabet, Kanji and others (e.g. θ, й, ゝ) are also separately created.

5) The criteria for ordering characters are the same as for VDRJ. The index used for ranking is *U* (Juilland & Chang-Rodrigues, 1964) and the same weighting system as VDRJ is adopted. (For more details, see 3.3.5 and 3.3.6.) For different types of learners, as done for word rankings in 3.3.6, the three types of Kanji rankings are made as follows.

1) The Ranking for Kanji in Written Japanese (KWJ)
2) The Ranking for Kanji for International Students (KIS)
3) The Ranking for Kanji for General Learners (KGL)

For the other types of characters, namely, Hiragana, Katakana, Roman alphabet, signs and others, no ranking is made but only usage coefficients and frequencies are shown as data. There are three reasons for this. First, there are not as many characters as Kanji for each type. Second, most of them should be learned regardless of their frequencies. Third, the order of learning should not depend on frequencies but on phonological order or on another order which takes account of cognitive considerations.

For making KWJ, KIS and KGL, *Fw* (Standardized frequency per million in VDRJ), *Fr1* (Standardized frequency per million in VDRJ by weighting one third on each of the three genres of IF, LW and AD), *Fr2* (Standardized frequency per million in VDRJ by

weighting only on IF and LW with the same weight i.e. 50% for each), *Uw* (*F*D*), *Ur1* (*Fr1*D*) and *Ur2* (*Fr2*D*) are computed as was done for word rankings. The weights on different domains in each usage coefficient type, which are the same as VDRJ, are shown in Table 5-3. The weights are used for creating KWJ, KIS and KGL as was done for the word rankings WWJ, WIS and WGL (For details, see Table 3-32 and its explanation).

Table 5-3 Weights (percentages) on the Sections of Internet Forum (IF), Literary Works (LW) and the Eight Academic Domains (AD) of CDJ for the Different Character Ranking Indices (=Table 3-32)

Usage Coefficient Type	Frequency Type	IF	LW	AD
*Uw = F*D*	*F*	15.9	25.1	59.0
*Ur1 = Fr1*D*	*Fr1*	33.3	33.3	33.3
*Ur2 = Fr2*D*	*Fr2*	50.0	50.0	0.0

F : Standardized frequency per million in VDRJ
Fr1 : Standardized frequency per million in VDRJ by weighting one third on each of the three genres of IF, LW and AD
Fr2 : Standardized frequency per million in VDRJ by weighting only on IF and LW with the same weight i.e. 50% for each

The adopted usage coefficient types for different Kanji rankings at different Kanji levels are shown in Table 5-4. The border between the ranges where words are ranked by *Ur1* and *Ur2*, in the second sorting key for KGL, is set at the ranking 400 because the top 400 Kanji cover the similar amount of text coverage as the top 2,000 words which is the border for WGL (the Word Ranking for General Learners) in VDRJ.

Table 5-4 Methods for the Kanji Rankings for Written Japanese (KWJ), International Students (KIS) and General Learners (KGL)

Kanji Ranking	*KWJ*	*KIS*		*KGL*	
	1st Key	1st Key	2nd Key	1st Key	2nd Key
1-103	*Uw*	F-JLPT4	*Ur2*	F-JLPT4	*Ur2*
104-284	*Uw*	F-JLPT3	*Ur2*	F-JLPT3	*Ur2*
285-400	*Uw*	F-JLPT2-0	*Uw*	F-JLPT2-0	*Ur2*
401+	*Uw*	F-JLPT2-0	*Uw*	F-JLPT2-0	*Ur1*

* KIS is priamarily assumed to be served for international students studying at Japanese universities as the texts in the corpus is mainly collected in Japan.
* KGL is assumed to be served for learners with non-academic purposes.
* F-JLPT: The former Japanese Language Proficiency Test word list level. 4 is the most basic, 1 is the highest and 0 is out of the levels (beyond 1).
* *Ur1* : Usage coefficient revised version 1 = *Fr1* *D
  Fr1 : (AD+LW+OC)/3
  AD: Standardized frequency per million of the 8 academic domains of LP, HE, AH, PL, EC, SE, ST and BM
* *Ur2* : Usage coefficient revised version 2 = *Fr2* *D
  Fr2 : (LW+OC)/2
  LW/OC: Standardized frequency per million in LW/OC
* Characters are sorted by descending order with the indices.

6) Cross-checking the list will be discussed in 5.5 and Chapter 6.

5.4 The product: The Character Database of Japanese (CDJ), Version 1

The completed database is available from the accompanying CD. For 6,522 characters, CDJ for Research provides information in the 53 fields shown in Table 5-5.

Table 5-5 Field Names of the Character Database of Japanese (CDJ)

留学生用漢字・記号レベル Level of Kanji for International Students
留学生用漢字・記号ランク Ranking for Kanji for International Students (KIS)
一般漢字・記号レベル Level of Kanji for General Learners
一般漢字・記号ランク Ranking for Kanji for General Learners (KGL)
書きことば漢字・記号レベル Level of Kanji in Written Japanese

書きことば漢字・記号ランク U Ranking for Kanji in Written Japanese (KWJ)
旧日本語能力試験出題基準レベル The Former JLPT Kanji Level
字種 Type of Character
文字 Item
使用度数 Frequency
字種混合頻度ランク Overall Freq Ranking
10 分野 100 万語あたり使用頻度(Fw) Standardized Freq/Million in 10 Written Domains (Fw)
(Fw)累積テキストカバー率 Fw Cumulative Text Coverage
8 分野 100 万語あたり使用頻度 Standardized Freq/Million in the 8 Domains
3 大分野 100 万語あたり使用頻度平均(Fr1) Freq revised ver 1/Million in the 3 Large Domains (Fr1)
(Fr1)累積テキストカバー率 Fr1 Cumulative Text Coverage
LW、OC2 分野 100 万語あたり使用頻度平均(Fr2) Standardized Freq/million in LW+OC (Fr2)
分散度 D
分散度順位 D Ranking
書きことば使用度係数(Uw) Uw (Usage Coefficient) for Written Japanese
修正使用度係数(Ur1) Ur1 (Usage Coefficient revised ver 1)
修正使用度係数(Ur2) Ur2 (Usage Coefficient revised ver 2)
使用範囲 Range
下位コーパス頻度（文芸創作） Sub-frequency in LW
100 万字あたり頻度（文芸創作） LW Freq per Million
使用頻度ランク（文芸創作） LW Freq Ranking

下位コーパス使用頻度（言語・哲学） Sub-frequency in LP
100 万字あたり使用頻度（言語・哲学） LP Freq per Million
使用頻度ランク（言語・哲学） LP Freq Ranking
下位コーパス使用頻度（歴史・民俗） Sub-frequency in HE
100 万字あたり使用頻度（歴史、民俗） HE Freq per Million
使用頻度ランク（歴史・民俗 HE Freq Ranking
下位コーパス使用頻度（芸術、その他の人文科学） Sub-frequency in AH
100 万語あたり使用頻度（芸術、その他の人文科学） AH Freq per Million
使用頻度ランク（芸術・その他の人文科学） AH Freq Ranking
下位コーパス使用頻度（政治・法律） Sub-frequency in PL
100 万語あたり使用頻度（政治・法律） PL Freq per Million
使用頻度ランク（政治・法律） PL Freq Ranking
下位コーパス使用頻度（経済・商業） Sub-frequency in EC
100 万語あたり使用頻度（経済・商業） EC Freq per Million
使用頻度ランク（経済・商業） EC Freq Ranking
下位コーパス使用頻度（社会・教育、その他の社会科学） Sub-frequency in SE
100 万語あたり使用頻度（社会・教育、その他の社会科学） SE Freq per Million
使用頻度ランク（社会・教育、その他の社会科学） SE Freq Ranking
下位コーパス使用頻度（科学・技術） Sub-frequency in ST
100 万語あたり使用頻度（科学・技術） ST Freq per Million
使用頻度ランク（科学・技術） ST Freq Ranking

下位コーパス使用頻度（生物・医学・生活科学） Sub-frequency in BM
100 万語あたり使用頻度（生物・医学・生活科学） BM Freq per Million
使用頻度ランク（生物・医学・生活科学） BM Freq Ranking
下位コーパス使用頻度（インターネット Q&A フォーラム） Q&A フォーラム）Sub-frequency in IF
100 万語あたり使用頻度（インターネット Q&A フォーラム） IF Freq per Million
使用頻度ランク（インターネット Q&A フォーラム） IF Freq Ranking

In the list, Roman alphabet, Hiragana, Katakana[69] are placed at the top because these types of characters should be included when calculating the text coverage as they are assumed to be known before a learner starts to learn Kanji.

All the other types of characters, namely, Kanji, signs and others, are sorted by the keys shown below.

1) The former Japanese Language Proficiency Test (F-JLPT) Kanji level (Descending) and Usage Coefficient (*Uw*/*Ur1*/*Ur2*) as described in Table 5-4 (Descending)
2) Frequency (*Fw*/*Fr1*/*Fr2*) (Descending)
3) Dispersion (*D*) (Descending)
4) Item (Ascending)

5.5 Validation of CDJ

5.5.1 Method

The validation method for CDJ is basically the same as the one used for VDRJ in

[69] Hiragana ゐ'wi' and ゑ 'we', and Katakana ヴ 'v', ヷ 'w', ヰ 'wi', ヸ 'wi', ヱ 'we', ヹ 'we' are not placed at the top of the list because they are not taught at the elementary level as they are not commonly used. These characters are classified as "S-Hiragana" or "S-Katakana" in the "Character Types" column.

Chapter 3. The questions for this section are as follows.

1) How well do the rankings for Kanji in CDJ correspond to or correlated with the ones in other lists such as the former Japanese Language Proficiency Test (F-JLPT) Kanji lists, the Japanese primary school Kanji grades (学年別漢字配当 'gakunen-haitou', MEXT, 1989) or the lists made from newspapers?

For this question, the data contained in the 4th edition database for the 1,945 basic Japanese kanji (Tamaoka, 2004) are exploited to compute the distribution and correlation[70]. The correlation between the frequency (*Fw*) and the adjusted frequency (*Uw*) is also checked as well as the correlations between different rankings.

2) Does the Ranking for Kanji for International Students (KIS) and the Ranking for Kanji for General Learners (KGL) provide higher text coverage than existing word lists such as the former Japanese Language Proficiency Test (F-JLPT) Kanji list (Japan Foundation & Association of International Education, Japan, 2002)?

As was done in Chapter 3, the Kanji rankings (KWJ, KIS and KGL), which should provide different levels of text coverage depending on the genre or media, are also compared to examine if the differences between them are as expected. Specifically, the questions here are as follows.

3) Does KIS provide higher text coverage for academic texts than KGL?
4) Does KGL provide higher text coverage for non-academic texts than KIS?
5) Does KGL provide higher text coverage for daily conversation texts than the Ranking

[70] The F-JLPT data in Tamaoka (2004) were updated by the author of this thesis as the data did not reflect the revision of the JLPT list made in 2002.

for Kanji in Written Japanese (KWJ) at all levels?

6) Does KIS provide higher text coverage for daily conversation texts than KWJ at the basic level?

The test corpora are the same as the ones used for testing the VDRJ word lists. The names of the test corpora are shown below. (For details, see 3.5.1.)

JS-NS: J-STAGE (Japan Science & Technology Information Aggregator) academic journal article texts in natural sciences.

MTT-NS: Meidai Technical Texts in Natural Sciences.

TB: Text Bank in Social Sciences for Intermediate and Advanced Learners of Japanese.

UYN: Utiyama Yomiuri Newspaper Corpus. (Utiyama & Isahara, 2003).

UPC: Utiyama Parallel Corpus. (Utiyama & Takahashi, 2003).

MC: Meidai Conversation Corpus.

As was done in Chapter 3, to check the text coverage, the software tool AntWordProfiler (Anthony, 2009) was used with baseword files. To compare the coverage of the F-JLPT Kanji lists and the other Kanji rankings, the same number of characters corresponding to each level of the F-JLPT are compiled into a baseword file. For example, the baseword file 'KIS_L2' is composed of the highest ranked Kanji beyond the F-JLPT Level 3 & 4 (KIS, KGL share the F-JLPT Level 3 & 4 lists at the top of the lists), and has the same number of Kanji as the F-JLPT Level 2. For comparing with other lists, each baseword file is made up of one hundred characters up to the 2,000 Kanji level (01C-20C) based on each Kanji ranking of KWJ, KIS and KGL. Beyond the level, all the words are put in a baseword file named 21C+. Roman alphabet, Hiragana, Katakana are put in separate lists.

As the methods are the same as the ones used for VDRJ in 3.5, the expected results

are also the same as the results in 3.5. Therefore, it is more important to check any different results from VDRJ. Any differences will be caused by the frequency gap between words and characters.

5.5.2 Results and discussion

The first question is the relationship between CDJ rankings and other lists. The correlation between the frequency (*Fw*) and adjusted frequency (*Uw*) is very high at 1.000 (Pearson, $p < .001$) and .987 (Spearman, $p < .001$) for all the characters. In the most frequent 2,000 Kanji, the correlation is still very high at .997 (Pearson, $p < .001$) and .993 (Spearman, $p < .001$). These results mean that adjusted frequency do not change the rankings for characters as much as the rankings for words. This is because many characters are used in a wide range of genres. In other words, the character distribution is not as uneven as words. The number of characters is limited and many of them are used for several different words. The ranking gap between *Fw* and *Uw* are also calculated. The ten Kanji with the largest ranking gap are 墳 (tumulus), 腎 (kidney), 倭 (the ancient name of Japan), 泌 (secretion), 胞 (for 細胞 'saibou' (cell)), 頷 (nod), 菩 (for 菩薩 'bosatsu' (*bodhisattva*)), 肪 (for 脂肪 'grease'), 患 (for 患者 'kanja' (patient)), 呟 (mutter) which are used only in a limited domain such as medicine or ancient Japanese history. However, such characters are not as many as words. Among the most frequent 2,000 Kanji, only 162 Kanji have the ranking gap which is more than one hundred.

Tables from 5-6 to 5-8 show how Japanese kanji are distributed at different levels of CDJ rankings, i.e. KWJ, KIS and KGL, and F-JLPT. As shown in Table 5-6, the Kanji in F-JLPT Level 3 and 4, which are elementary levels, are mostly at the levels between 01C and 04C in KWJ; however, some words occur in low-frequency levels beyond 10C. The Level 2 (intermediate) and the Level 1 (advanced) Kanji are distributed across a considerably wide range of levels. The Level 2 Kanji are spread out from 01C to 20C, and the Level 1 Kanji are spread from 03C to 21C+. The important criteria to rank Kanji are not

only frequency but also many factors such as utility as a component of other Kanji; however, there seem to be no clear criteria to distinguish the Level 2 and the Level 1 for the F-JLPT. The total number of Kanji at Level 4 to Level 2 is 1,000 which is close to the number of Kanji taught at primary schools in Japan (Grade 1 to 6 Kanji in Tables 5-9 to 5-11), yet, the selected Kanji are not totally the same. The current JLPT has new Kanji lists which are not available to the public; however, the KIS and KGL list will probably be similar to the current JLPT lists as the current JLPT takes account of newer frequency lists (Akimoto & Oshio, 2008).

Table 5-6 Distribution of Japanese Kanji by the KWJ Level and the F-JLPT Kanji Level

KWJ Level \ F-JLPT Kanji Level	4	3	2	1	None	Total
W_01C	***48***	**34**	18			100
W_02C	18	30	*51*		1	100
W_03C	14	31	*53*	2		100
W_04C	8	27	*58*	7		100
W_05C	4	7	***76***	12	1	100
W_06C	5	15	*62*	18		100
W_07C		9	*72*	19		100
W_08C	2	9	*61*	27	1	100
W_09C	3	5	*51*	40	1	100
W_10C	1	3	*57*	39		100
W_11C		8	39	*49*	4	100
W_12C		2	37	*59*	2	100
W_13C			29	*68*	3	100
W_14C			23	*74*	3	100
W_15C		1	18	*74*	7	100
W_16C			15	*75*	10	100
W_17C			10	***76***	14	100
W_18C			4	*75*	21	100
W_19C			4	*66*	30	100
W_20C			1	48	*51*	100
W_21C+				189	***4142***	4331
Total	103	181	739	1017	4291	6331

* KWJ: The Ranking for Kanji and Signs in Written Jananese
* Numbers in bold types are the greatest (=mode) at each F-JLPT Level.
* Italic numbers are the greatest (=mode) at each KWJ level.

The overall distributions across KIS/KGL and F-JLPT (Table 5-7 and 5-8) are similar to the distribution across KWJ and F-JLPT (Table 5-6) except for the 01C and 02C levels where all the F-JLPT Level 3 and 4 words are placed in KIS and KGL.

Table 5-7 Distribution of Japanese Kanji by the KIS Level and the F-JLPT Kanji Level

F-JLPT Kanji Level / KIS Level	4	3	2	1	None	Total
W_01C	***100***					100
W_02C	3	***97***				100
W_03C		*84*	16			100
W_04C			***97***	2	1	100
W_05C			*92*	8		100
W_06C			*79*	20	1	100
W_07C			*82*	18		100
W_08C			*72*	27	1	100
W_09C			*57*	42	1	100
W_10C			*58*	42		100
W_11C			43	*53*	4	100
W_12C			39	*59*	2	100
W_13C			29	*68*	3	100
W_14C			23	*74*	3	100
W_15C			18	*75*	7	100
W_16C			15	*75*	10	100
W_17C			10	***76***	14	100
W_18C			4	*75*	21	100
W_19C			4	*66*	30	100
W_20C			1	48	*51*	100
W_21C+				189	***4142***	4331
Total	103	181	739	1017	4291	6331

* KIS: The Ranking for Kanji and Signs for international
* Numbers in bold types are the greatest (=mode) at each F-JLPT Level.
* Italic numbers are the greatest (=mode) at each KIS level.

Table 5-8 Distribution of Japanese Kanji by the KGL Level and the F-JLPT Kanji Level

KGL Level \ F-JLPT Kanji Level	4	3	2	1	None	Total
W_01C	***100***					100
W_02C	3	***97***				100
W_03C		*84*	15		1	100
W_04C			***94***	6		100
W_05C			*93*	7		100
W_06C			*84*	15	1	100
W_07C			*80*	20		100
W_08C			*77*	22	1	100
W_09C			*53*	46	1	100
W_10C			*56*	43	1	100
W_11C			41	*56*	3	100
W_12C			47	*52*	1	100
W_13C			26	*72*	2	100
W_14C			25	*68*	7	100
W_15C			18	*74*	8	100
W_16C			9	***79***	12	100
W_17C			9	*72*	19	100
W_18C			7	*74*	19	100
W_19C			2	*60*	38	100
W_20C			3	*54*	43	100
W_21C+				197	***4134***	4331
Total	103	181	739	1017	4291	6331

* KGL: The Ranking for Kanji and Signs for General Learners
* Numbers in bold types are the greatest (=mode) at each F-JLPT Level.
* Italic numbers are the greatest (=mode) at each KGL level.

Tables from 5-9 to 5-11 show the distributions across the CDJ rankings, i.e. KWJ, KIS and KGL, and the Japanese primary school Kanji grades (MEXT, 1989). All the distributions show that the Grade 1 and 2 Kanji in the Japanese primary school Kanji grades are also ranked highly in all the CDJ rankings. As the primary school grade gets higher, the CDJ ranking also moves to the low-frequency range. These results show that the CDJ rankings are basically valid.

Table 5-9 Distribution of Japanese Kanji by the KWJ Level and the Japanese Primary School Kanji Grades

Grades / KWJ Level	1	2	3	4	5	6	7	None	Total
W_01C	**30**	***39***	20	7	2	1	1		100
W_02C	13	22	***33***	19	10		2	1	100
W_03C	10	24	*25*	17	18	5	1		100
W_04C	6	20	19	***23***	21	7	3	1	100
W_05C	6	7	18	21	***23***	12	10	3	100
W_06C	3	7	*21*	19	*21*	12	17		100
W_07C	1	7	15	20	15	***26***	16		100
W_08C	3	8	9	14	18	22	*24*	2	100
W_09C	4	6	7	9	13	17	*42*	2	100
W_10C		5	8	14	10	16	*46*	1	100
W_11C	1	8	8	7	9	13	*49*	5	100
W_12C	1	4	4	10	7	10	*61*	3	100
W_13C	1		4	8	6	11	*65*	5	100
W_14C			2	5	3	10	*70*	10	100
W_15C		1	3	1	2	5	***73***	15	100
W_16C			2	1	5	3	*69*	20	100
W_17C	1	1	1	4	1	2	*67*	23	100
W_18C				1		2	*62*	35	100
W_19C						4	*55*	41	100
W_20C		1	1			1	*37*	60	100
W_21C+					1	2	169	*4159*	4331
Total	80	160	200	200	185	181	939	4386	6331

* KWJ: The Ranking for Kanji and Signs in Written Jananese
* Numbers in bold types are the greatest at each grade.
* Italic numbers are the greatest at each KWJ level.

Table 5-10 Distribution of Japanese Kanji by the KIS Level and the Japanese Primary School Kanji Grades

Grades / KIS Level	1	2	3	4	5	6	7	None	Total
IS_01C	***56***	41	3						100
IS_02C	7	***45***	**39**	7	1	1			100
IS_03C	9	33	*34*	18	3	2	1		100
IS_04C		9	*32*	**27**	25	4	2	1	100
IS_05C	1	5	16	**27**	***32***	11	7	1	100
IS_06C	1	5	15	*25*	23	14	14	3	100
IS_07C	1	3	13	19	21	*22*	21		100
IS_08C		3	10	17	21	***27***	20	2	100
IS_09C	1	4	9	8	14	18	*44*	2	100
IS_10C		2	6	16	12	17	*46*	1	100
IS_11C	1	4	7	7	8	13	*55*	5	100
IS_12C	1	3	4	9	7	12	*61*	3	100
IS_13C	1		4	8	6	11	*65*	5	100
IS_14C			2	5	3	10	*70*	10	100
IS_15C		1	2	1	2	5	***74***	15	100
IS_16C			2	1	5	3	*69*	20	100
IS_17C	1	1	1	4	1	2	*67*	23	100
IS_18C				1		2	*62*	35	100
IS_19C						4	*55*	41	100
IS_20C		1	1			1	*37*	60	100
IS_21C+					1	2	169	*4159*	4331
総計	80	160	200	200	185	181	939	4386	6331

* KIS: The Ranking for Kanji and Signs for Interenational Students
* Numbers in bold types are the greatest at each grade.
* Italic numbers are the greatest at each KIS level.

Table 5-11 Distribution of Japanese Kanji by the KGL Level and the Japanese Primary School Kanji Grades

Grades / KGL Level	1	2	3	4	5	6	7	None	Total
GL_01C	***56***	41	3						100
GL_02C	7	***45***	**39**	7	1	1			100
GL_03C	9	33	*34*	17	2	2	2	1	100
GL_04C		8	30	***31***	18	5	7	1	100
GL_05C	1	7	22	24	***29***	10	6	1	100
GL_06C	1	3	13	21	***29***	16	15	2	100
GL_07C	1	2	10	22	*23*	**22**	20		100
GL_08C	1	7	13	17	17	*21*	*21*	3	100
GL_09C		2	5	12	13	17	*49*	2	100
GL_10C		3	9	10	15	17	*44*	2	100
GL_11C	1	6	5	7	10	16	*50*	5	100
GL_12C	1		3	12	8	12	*61*	3	100
GL_13C	1		4	5	6	11	*62*	11	100
GL_14C			3	7	5	8	*64*	13	100
GL_15C		1	5	1	3	9	*66*	15	100
GL_16C		1		3	1	2	***70***	23	100
GL_17C				2	2	1	*67*	28	100
GL_18C	1		1	2	1	2	*62*	31	100
GL_19C					1	5	*50*	44	100
GL_20C			1				*49*	50	100
GL_21C+		1			1	4	174	*4151*	4331
総計	80	160	200	200	185	181	939	4386	6331

* KGL: The Ranking for Kanji and Signs for General Learners
* Numbers in bold types are the greatest at each grade.
* Italic numbers are the greatest at each KGL level.

Correlations between the levels and rankings in CDJ and other lists are also computed (Table 5-12 and 5-13). The correlation between CDJ rankings (i.e. KWJ, KIS and KGL) and F-JLPT/Grades (the Japanese primary school Kanji grades) are not very high between *r* = .74 and .80; however, this is because the F-JLPT levels and the Japanese primary school Kanji grades only have five and seven levels respectively. The more important thing is that KWJ, which is the ranking purely depending on the adjusted frequencies based on the book and internet-forum texts, show higher correlations (Spearman's *Rho*) with F-JLPT at *r* = .742 and the Japanese primary school Kanji grades at *r* = .742 than the other frequency rankings of KF1976 (NLRI, 1976), KF1998 (Yokoyama et al., 1998) and KF2000 (Amano

& Kondo, 2000). This means that, for the levels/grades made by the expert committees, the adjusted frequency where dispersion is taken into account, work better than pure frequency counts, and that the book and internet-forum texts works better than newspaper texts in general. KIS and KGL, where the F-JLPT Level 3 and 4 words are placed at the top, have even higher correlations with F-JLPT lists than the other lists as expected. However, more interestingly, KIS and KGL also show higher correlations with the Japanese primary school Kanji grades at r = .776 and .778 respectively than the other lists which are between r = .59 and .76. This may be because F-JLPT took account of the grades when they made the lists. Or, more essentially, both F-JLPT and the Japanese primary school Kanji grades take account of the 'basicness' of Kanji which may include cognitive basicness and utility as a component of Kanji compounds. The CDJ rankings seem to reflect more of the basicness than the frequencies from newspaper texts.

The rankings in KWJ and the frequencies in KF1976 (NLRI, 1976) and KF1998 (Yokoyama et al., 1998) are highly correlated at .91 (Spearman's *Rho*) and .85 (Pearson) or higher[71]. The correlation between KWJ and KF2000 (Amano & Kondo, 2000) is a little lower at .82 (Spearman's *Rho*) and .75 (Pearson); however, all of these data prove that the overall rankings in CDJ correlate well with newspaper frequencies. The gap between 1.000 and the coefficient figures show that there are some Kanji which are ranked considerably differently in different lists.

[71] In CDJ rankings (i.e. KWJ, KIS and KGL) and the Japanese primary school Kanji grades, the smaller the number, the more basic the Kanji. Therefore, the correlation figures in Table 5-12 show negative between these rankings/grades and other frequencies.

Table 5-12 Correlation between the Kanji Levels and Rankings in CDJ and the Other Lists (Spearman's Rank Correlation)

	KWJ	KIS	KGL	F-JLPT	Grades	KF1976	KF1998	KF2000
n	1945	1945	1945	1926	1945	1945	1945	1945
KWJ	1.000	.978	.973	-.742	.742	-.913	-.913	-.823
KIS	.978	1.000	.996	-.793	.776	-.899	-.889	-.794
KGL	.973	.996	1.000	-.797	.768	-.886	-.871	-.771
F-JLPT	-.742	-.793	-.797	1.000	-.777	.729	.677	.599
Grades	.742	.776	.768	-.777	1.000	-.759	-.707	-.632
KF1976	-.913	-.899	-.886	.729	-.759	1.000	.944	.842
KF1998	-.913	-.889	-.871	.677	-.707	.944	1.000	.878
KF2000	-.823	-.794	-.771	.599	-.632	.842	.878	1.000

* All the coefficients are significant ($p < .001$)
* Data for Grades, KF1976, KF1998, KFCD1998, KF2000 are cited from Tamaoka (2004).
* KWJ: The Ranking for Kanji and Signs in Written Jananese
* KIS: The Ranking for Kanji and Signs for Interenational Students
* KGL: The Ranking for Kanji and Signs for General Learners
* F-JLPT: The former Japanese Language Proficiency Test
* Grades: The Japanese primary school Kanji grades
* KF1976: Kanji frequency data from NLRI (1976)
* KF1998: Kanji frequency data from Yokoyama et al. (1998)
* KF2000: Kanji Frequency data from Amano & Kondo (2000)

Table 5-13 Correlation between the Kanji Levels and Rankings in CDJ and the Other Lists (Pearson's Correlation Coefficient)

	Uw	*Ur1*	*Ur2*	KF1976	KF1998	KF2000
n	1945	1945	1945	1945	1945	1945
Uw	1.000	.992	.952	.836	.850	.750
Ur1	.992	1.000	.983	.800	.814	.726
Ur2	.952	.983	1.000	.729	.740	.666
KF1976	.836	.800	.729	1.000	.969	.721
KF1998	.850	.814	.740	.969	1.000	.799
KF2000	.750	.726	.666	.721	.799	1.000

* All the coefficients are significant ($p < .001$)
* Data for Grades, KF1976, KF1998, KFCD1998, KF2000 are cited from Tamaoka (2004).
* KWJ: The Ranking for Kanji and Signs in Written Jananese
* KIS: The Ranking for Kanji and Signs for Interenational Students
* KGL: The Ranking for Kanji and Signs for General Learners
* F-JLPT: The former Japanese Language
* Grades: The Japanese primary school Kanji
* KF1976: Kanji frequency data from NLRI
* KF1998: Kanji frequency data from Yokoyama et al. (1998)
* KF2000: Kanji Frequency data from Amano & Kondo (2000)

For the second question 2) Does the Ranking for Kanji for International Students (KIS) and the Ranking for Kanji for General Learners (KGL) provide higher text coverage than existing word lists such as the former Japanese Language Proficiency Test (F-JLPT) Kanji list (Japan Foundation & Association of International Education, Japan, 2002)?, the answer is yes as shown in Table 5-14.

Table 5-14 Text Coverage (Percentage) in Different Genres by KIS, KGL and F-JLPT

Test Corpus Code	JS-NS	MTT-NS	TB	UYN	UPC	MC
Genre	Technical (Natural Sciences)	Academic (Natural Sciences)	Academic (Social Sciences)	Newspaper	Literary Works	Converation
KIS Level 4 to 2 (Hiragana, Katakana, Roman alphabet +1023 Kanji)	96.51	97.63	97.74	96.68	97.60	98.60
KGL Level 4 to 2 (Hiragana, Katakana, Roman alphabet +1023 Kanji)	96.32	97.59	97.52	96.42	97.63	98.64
F-JLPT Level 4 to 2 (Hiragana, Katakana, Roman alphabet +1023 Kanji)	94.87	96.68	96.34	94.94	96.76	98.35
Gap (KIS - 'F-JLPT')	**1.64**	**0.95**	**1.40**	**1.74**	0.84	0.25
Gap (KGL - 'F-JLPT')	1.45	0.91	1.18	1.48	**0.87**	**0.29**
KIS Level 4 to 1 (Hiragana, Katakana, Roman alphabet +2040 Kanji)	99.52	99.83	99.87	99.86	99.68	99.31
KGL Level 4 to 1 (Hiragana, Katakana, Roman alphabet +2040 Kanji)	99.48	99.81	99.83	99.83	99.68	99.31
F-JLPT Level 4 to 1 (Hiragana, Katakana, Roman alphabet +2040 Kanji)	99.49	99.80	99.79	99.77	99.43	99.27
Gap (KIS - 'F-JLPT')	**0.03**	**0.03**	**0.08**	**0.09**	0.25	0.04
Gap (KGL - 'F-JLPT')	-0.01	0.01	0.04	0.06	0.25	0.04

*KIS: The Ranking for Kanji and Signs for International Students
*KGL: The Ranking for Kanji and Signs for General Learners
*F-JLPT: The former Japanese Language Proficiency Test word list
(Level 4 is the most basic and Level 1 is the most advanced.)
*Bold figures are explained in the body.

The gap (KIS/KGL - 'F-JLPT') figures show how much percent higher text coverage by KIS/KGL provides compared to F-JLPT. The figures are all positive, that is, KIS and KGL are superior to F-JLPT. As the figures in bold type show, for academic and newspaper texts, KIS performs better than KGL while KGL performs better than KIS for literary works and conversation. The gaps are larger when the rankings are compared at the Level 2 (Note that both KIS and KGL share the Level 3 and 4 vocabulary with F-JLPT) than at all levels including the Level 1 because Level 1 includes most of the common Kanji. In other words, the ranking gap mainly exists in the order of Kanji in the mid-frequency level which is at the rankings between 300 and 1,000.

The gap figures in text coverage shown in Table 5-14 are smaller than the ones shown in word frequencies (e.g. Table 3-35). This is inevitable because many words are composed of two or more characters. In other words, the gaps in character coverage will lead to greater gaps in word coverage.

Tables from 5-15 to 5-18 show that the Kanji rankings are basically valid as text

coverage for each 100 Kanji level gradually decreases from the more frequent words (01C) to the less frequent (20C) in most of the cases shown in the tables.

For the third question 3) Does KIS provide higher text coverage for academic texts than KGL?, as expected, the answer is yes (Table 5-15 and 5-16) as the 'Gap (KIS-KGL)' figures are all positive from the 03C level to the 20C level. (At the 01C and 02C levels, there cannot be gaps between KIS and KGL as the both rankings share all the characters at the levels.)

As shown in Table 5-15, natural science journal articles, compared to other types of texts, contain notably high proportions of Roman alphabet and Katakana at 5.53 and 8.41% respectively. In particular, the proportion for Roman alphabet is much higher than other types of texts because there are many technical terms described in English. The proportions for Katakana are considerably high in literary works and conversation texts as well as natural science texts. This may be because the average length of Katakana words is longer than Hiragana words and Kanji words.

(From here down blank.)

Table 5-15 Text Coverage of JS-NS (Technical, Natural Sciences) at Each Character Level by KIS, KGL and KWJ

	KIS		KGL		KWJ		Gap (KIS-KGL)		Gap (KIS-KWJ)	
LEVEL LIST	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)
Roman alphabet	**5.53**	5.53	**5.53**	5.53	**5.53**	5.53	0.00	0.00	0.00	0.00
Hiragana	**43.07**	48.60	**43.07**	48.60	**43.07**	48.60	0.00	0.00	0.00	0.00
Katakana	**8.41**	57.01	**8.41**	57.01	**8.41**	57.01	0.00	0.00	0.00	0.00
01C	6.99	64.00	6.99	64.00	13.59	70.60	**0.00**	0.00	-6.59	**-6.59**
02C	7.02	71.03	7.02	71.03	6.79	77.39	**0.00**	0.00	0.23	**-6.36**
03C	4.97	76.00	4.05	75.08	5.28	82.67	0.92	**0.92**	-0.31	**-6.67**
04C	**8.11**	84.11	**8.03**	83.11	3.66	86.32	0.08	**1.00**	4.46	**-2.21**
05C	4.13	88.24	4.28	87.39	3.04	89.37	-0.15	**0.85**	1.08	**-1.13**
06C	2.55	90.78	2.68	90.07	2.22	91.59	-0.14	**0.71**	0.32	**-0.80**
07C	1.96	92.75	2.27	92.34	1.62	93.20	-0.31	**0.40**	0.34	**-0.46**
08C	1.56	94.31	1.62	93.97	1.30	94.50	-0.06	**0.34**	0.27	**-0.19**
09C	1.13	95.44	1.09	95.06	1.04	95.54	0.04	**0.38**	0.10	**-0.10**
10C	0.95	96.39	0.90	95.96	0.89	96.42	0.05	**0.43**	0.06	**-0.04**
11C	0.73	97.12	0.91	96.87	0.71	97.13	-0.18	**0.25**	0.02	**-0.02**
12C	0.61	97.73	0.59	97.47	0.59	97.73	0.02	**0.26**	0.02	**0.00**
13C	0.30	98.03	0.49	97.96	0.31	98.03	-0.19	**0.07**	0.00	**-0.01**
14C	0.25	98.28	0.23	98.19	0.25	98.28	0.02	**0.09**	0.00	**0.00**
15C	0.39	98.67	0.39	98.58	0.39	98.67	0.00	**0.09**	0.00	**0.00**
16C	0.16	98.83	0.14	98.72	0.16	98.83	0.02	**0.11**	0.00	**0.00**
17C	0.29	99.12	0.28	99.00	0.29	99.12	0.01	**0.12**	0.00	**0.00**
18C	0.21	99.32	0.23	99.23	0.21	99.32	-0.02	**0.10**	0.00	**0.00**
19C	0.10	99.43	0.15	99.38	0.10	99.44	-0.05	**0.05**	0.00	**-0.01**
20C	0.05	99.48	0.08	99.44	0.05	99.48	-0.03	**0.04**	0.00	**0.00**
21C+	0.52	99.99	0.54	99.99	0.52	99.99	-0.02	0.00	0.00	0.00
Not in the Lists	0.00	100.00	0.00	100.00	0.00	100.00	0.00	0.00	0.00	0.00

* JS-NS: J-STAGE technical journal article texts in natural sciences (total character token: 3,322,109)
* KIS: The Ranking Kanji and Signs for International Students
* KGL: The Ranking for Kanji and Signs for General Learners
* KWJ: The Ranking for Kanji and Signs in Written Japanese
* TC: Text coverage Cum. TC: Cumulative text coverage
* AKW: Assumed Known Words which are mostly proper nouns not requiring previous learning.
* Bold figures are explained in the body.

At the 04C level in KIS and KGL, the figures are notably greater than other levels. The most frequent Kanji at this level in the texts are 定 (fix, constant), 化 (change, -ize, chemistry), 数 (number), 流 (flow), 対 (to, towards, against), 関 (relate, function), 法 (method, law), 結 (connect, tie), 成 (become, or for the word 成分 which means 'ingredient' or 'constituent'), 加 (add). All of these Kanji, which are all placed at the 04C level in KIS and KGL, are essential in natural sciences. This level also includes Kanji such as 面 (side, aspect), 解 (solution), 表 (table, surface), 形 (shape, form), 線 (line), 点 (point).

In social science texts, 04C also provides high text coverage (Table 5-16).

Table 5-16 Text Coverage of TB (Academic, Social Sciences) at Each Character Level by KIS, KGL and KWJ

	KIS		KGL		KWJ		Gap (KIS-KGL)		Gap (KIS-KWJ)	
LEVEL LIST	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)
Roman alphabet	0.67	0.67	0.67	0.67	0.67	0.67	0.00	0.00	0.00	0.00
Hiragana	**55.09**	55.76	55.09	55.76	55.09	55.76	0.00	0.00	0.00	0.00
Katakana	4.97	60.73	4.97	60.73	4.97	60.73	0.00	0.00	0.00	0.00
01C	8.39	69.12	8.39	69.12	14.92	75.65	0.00	0.00	-6.53	**-6.53**
02C	7.49	76.62	7.49	76.62	6.41	82.06	0.00	0.00	1.09	**-5.44**
03C	4.62	81.23	3.56	80.18	4.32	86.38	1.06	**1.06**	0.30	**-5.15**
04C	**6.36**	87.60	**6.12**	86.30	3.47	89.85	0.24	**1.30**	2.89	**-2.25**
05C	3.70	91.30	4.12	90.41	2.44	92.29	-0.42	**0.88**	1.26	**-0.99**
06C	2.15	93.44	2.50	92.91	1.84	94.13	-0.35	**0.53**	0.30	**-0.69**
07C	1.66	95.10	1.64	94.55	1.21	95.34	0.02	**0.55**	0.45	**-0.24**
08C	1.07	96.17	1.13	95.68	0.94	96.28	-0.05	**0.50**	0.13	**-0.11**
09C	0.82	97.00	0.91	96.59	0.75	97.03	-0.09	**0.41**	0.07	**-0.03**
10C	0.64	97.64	0.83	97.42	0.62	97.65	-0.19	**0.22**	0.02	**-0.01**
11C	0.58	98.22	0.47	97.90	0.58	98.23	0.11	**0.33**	0.01	**0.00**
12C	0.45	98.67	0.49	98.38	0.44	98.67	-0.04	**0.29**	0.00	**0.00**
13C	0.41	99.07	0.49	98.87	0.41	99.08	-0.08	**0.21**	0.00	**0.00**
14C	0.19	99.26	0.22	99.08	0.18	99.26	-0.03	**0.18**	0.00	0.00
15C	0.18	99.44	0.27	99.35	0.18	99.44	-0.09	**0.08**	0.00	0.00
16C	0.14	99.58	0.11	99.46	0.14	99.58	0.04	**0.12**	0.00	0.00
17C	0.08	99.66	0.14	99.60	0.08	99.66	-0.07	**0.05**	0.00	0.00
18C	0.08	99.74	0.10	99.70	0.08	99.74	-0.02	**0.04**	0.00	0.00
19C	0.08	99.81	0.06	99.76	0.08	99.81	0.01	**0.05**	0.00	0.00
20C	0.04	99.85	0.04	99.80	0.04	99.85	0.00	**0.05**	0.00	0.00
21C+	0.15	100.00	0.20	100.00	0.15	100.00	-0.05	0.00	0.00	0.00
Not in the Lists	0.00	100.00	0.00	100.00	0.00	100.00	0.00	0.00	0.00	0.00

* TB: Text Bank in Social Sciences for Intermediate and Advanced Learners of Japanese (total character token: 295,768)

* KIS: The Ranking Kanji and Signs for International Students
* KGL: The Ranking for Kanji and Signs for General Learners
* KWJ: The Ranking for Kanji and Signs in Written Japanese
* TC: Text coverage Cum. TC: Cumulative text coverage
* AKW: Assumed Known Words which are mostly proper nouns not requiring previous learning.
* Bold figures are explained in the body.

The frequent Kanji at the 04C level used in the social science texts are 化 (-ize), 制 (system, restriction), 経 (for 経済 'keizai' (economy)), 政 (for 政治 'seiji' (politics)),数 (number), 利 (benefit) and so on. These are also essential Kanji for social sciences. These Kanji, for academic purposes, should be learned right after the very basic Kanji at the 01 and 02C levels despite the frequency since they will have higher domain-specificity in academic or formal texts. This issue will further be explored in Chapter 7.

Table 5-17 Text Coverage of UPC (Literary Works) at Each Character Level by KIS, KGL and KWJ

	KIS		KGL		KWJ		Gap (KIS-KGL)		**Gap (KWJ-KGL)**	
LEVEL LIST	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)
Roman alphabet	1.72	1.72	1.72	1.72	1.72	1.72	0.00	0.00	0.00	0.00
Hiragana	**65.58**	67.30	65.58	67.30	65.58	67.30	0.00	0.00	0.00	0.00
Katakana	**8.43**	75.73	8.43	75.73	8.43	75.73	0.00	0.00	0.00	0.00
01C	6.35	82.08	6.35	82.08	**9.37**	85.09	0.00	0.00	**3.02**	3.02
02C	4.77	86.84	4.77	86.84	**3.24**	88.34	0.00	0.00	**-1.52**	1.50
03C	2.08	88.92	2.12	88.97	**2.31**	90.65	-0.05	**-0.05**	**0.19**	1.69
04C	**2.81**	91.73	**2.72**	91.69	**1.63**	92.29	0.08	**0.04**	**-1.09**	0.59
05C	1.56	93.29	1.72	93.41	**1.41**	93.70	-0.16	**-0.12**	**-0.30**	0.29
06C	1.38	94.67	1.39	94.80	**1.12**	94.82	-0.01	**-0.13**	**-0.27**	0.02
07C	0.93	95.60	0.85	95.66	**0.82**	95.65	0.08	**-0.06**	**-0.03**	**-0.01**
08C	0.71	96.31	0.73	96.39	**0.72**	96.37	-0.02	**-0.08**	**-0.01**	**-0.02**
09C	0.64	96.95	0.58	96.97	**0.61**	96.99	0.06	**-0.02**	**0.03**	**0.02**
10C	0.55	97.50	0.55	97.52	**0.52**	97.50	0.00	**-0.02**	**-0.03**	**-0.02**
11C	0.45	97.95	0.47	97.99	**0.46**	97.96	-0.02	**-0.03**	**-0.01**	**-0.02**
12C	0.36	98.32	0.36	98.35	**0.35**	98.32	0.00	**-0.03**	**-0.01**	**-0.03**
13C	0.26	98.58	0.24	98.59	**0.26**	98.58	0.03	**0.00**	0.03	**-0.01**
14C	0.23	98.81	0.23	98.82	**0.23**	98.81	0.00	**-0.01**	0.00	**-0.01**
15C	0.22	99.03	0.22	99.04	**0.22**	99.03	0.00	**-0.01**	0.00	**-0.01**
16C	0.16	99.19	0.18	99.21	**0.16**	99.19	-0.01	**-0.02**	-0.01	**-0.02**
17C	0.14	99.34	0.14	99.35	**0.14**	99.34	0.00	**-0.02**	0.00	**-0.02**
18C	0.12	99.46	0.12	99.48	**0.12**	99.46	0.00	**-0.02**	0.00	**-0.02**
19C	0.12	99.57	0.10	99.57	**0.12**	99.57	0.02	0.00	0.02	0.00
20C	0.07	99.65	0.07	99.65	**0.07**	99.65	0.00	0.00	0.00	0.00
21C+	0.35	100.00	0.35	100.00	0.35	100.00	0.00	0.00	0.00	0.00
Not in the Lists	0.00	100.00	0.00	100.00	0.00	100.00	0.00	0.00	0.00	0.00

* UPC: Utiyama Parallel Corpus (total character token: 3,508,356)
* KIS: The Ranking Kanji and Signs for International Students
* KGL: The Ranking for Kanji and Signs for General Learners
* KWJ: The Ranking for Kanji and Signs in Written Japanese
* TC: Text coverage Cum. TC: Cumulative text coverage
* AKW: Assumed Known Words which are mostly proper nouns not requiring previous learning.
* Bold figures are explained in the body.

In literary works (Table 5-17) and even in conversation texts (Table 5-18), the proportions for 04C in KIS and KGL are also slightly higher than 03C. Taking account of the fact that KIS and KGL share the F-JLPT Level 4 and 3 Kanji up to the ranking 284 at 03C, the rankings between 285 and 400 contain many important Kanji for written texts which must be placed at the level between 01C and 03C in KWJ.

Table 5-18 Text Coverage of MC (Conversation) at Each Character Level by KIS, KGL and KWJ

	KIS		KGL		KWJ		Gap (KIS-KGL)		**Gap (KWJ-KGL)**		**Gap (KIS-KWJ)**	
LEVEL LIST	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)	TC (%)	Cum. TC (%)		
Roman alphabet	0.19	0.19	0.19	0.19	0.19	0.19	0.00	0.00	0.00	0.00	0.00	0.00
Hiragana	**76.52**	76.72	76.52	76.72	76.52	76.72	0.00	0.00	0.00	0.00	0.00	0.00
Katakana	7.01	83.73	7.01	83.73	7.01	83.73	0.00	0.00	0.00	0.00	0.00	0.00
01C	6.21	89.94	6.21	89.94	7.53	91.26	0.00	0.00	**1.32**	1.32	**-1.32**	-1.32
02C	3.12	93.06	3.12	93.06	2.16	93.42	0.00	0.00	**-0.96**	0.36	**0.96**	-0.36
03C	1.31	94.37	1.43	94.49	1.42	94.84	-0.12	**-0.12**	**-0.01**	0.35	**-0.10**	-0.46
04C	**1.50**	95.87	**1.50**	95.99	0.98	95.82	-0.01	**-0.12**	**-0.52**	**-0.17**	0.52	0.05
05C	0.72	96.59	0.69	96.68	0.54	96.36	0.02	**-0.10**	**-0.15**	**-0.33**	0.18	0.23
06C	0.54	97.12	0.61	97.29	0.67	97.03	-0.07	**-0.17**	0.06	**-0.27**	-0.13	0.10
07C	0.47	97.59	0.42	97.71	0.50	97.52	0.05	**-0.12**	0.08	**-0.19**	-0.03	0.07
08C	0.42	98.01	0.42	98.12	0.38	97.90	0.01	**-0.11**	-0.03	**-0.22**	0.04	0.11
09C	0.25	98.26	0.28	98.41	0.29	98.19	-0.04	**-0.15**	0.00	**-0.22**	-0.04	0.07
10C	0.30	98.56	0.20	98.60	0.31	98.50	0.10	**-0.05**	0.11	**-0.11**	-0.01	0.06
11C	0.19	98.75	0.18	98.78	0.24	98.74	0.01	**-0.03**	0.06	**-0.05**	-0.05	0.01
12C	0.15	98.90	0.14	98.92	0.16	98.90	0.01	**-0.02**	0.02	**-0.02**	-0.01	0.00
13C	0.12	99.02	0.11	99.03	0.12	99.02	0.01	**-0.01**	0.01	**-0.01**	0.00	0.00
14C	0.09	99.11	0.08	99.11	0.09	99.11	0.01	**0.00**	0.01	**0.00**	0.00	0.00
15C	0.06	99.17	0.07	99.18	0.06	99.17	-0.01	**-0.01**	0.00	**-0.01**	0.00	0.00
16C	0.05	99.22	0.05	99.23	0.05	99.22	0.00	**-0.01**	0.00	**-0.01**	0.00	0.00
17C	0.04	99.26	0.03	99.26	0.04	99.26	0.01	0.00	0.01	0.00	0.00	0.00
18C	0.04	99.30	0.03	99.29	0.04	99.30	0.01	0.00	0.01	0.00	0.00	0.00
19C	0.02	99.31	0.02	99.31	0.02	99.31	-0.01	0.00	-0.01	0.00	0.00	0.00
20C	0.01	99.32	0.01	99.32	0.01	99.32	0.00	0.00	0.00	0.00	0.00	0.00
21C+	0.65	99.97	0.65	99.97	0.65	99.97	0.00	0.00	0.00	0.00	0.00	0.00
Not in the Lists	0.03	100.00	0.03	100.00	0.03	100.00	0.00	0.00	0.00	0.00	0.00	0.00

* MC: Meidai Converation Corpus (total character token: 1,936,658)
* KIS: The Ranking Kanji and Signs for International Students
* KGL: The Ranking for Kanji and Signs for General Learners
* KWJ: The Ranking for Kanji and Signs in Written Japanese
* TC: Text coverage Cum. TC: Cumulative text coverage
* AKW: Assumed Known Words which are mostly proper nouns not requiring previous learning.
* Bold figures are explained in the body.

For the fourth question 4) Does KGL provide higher text coverage for non-academic texts than KIS?, the answer is yes as the 'Gap (KIS-KGL)' figures are mostly negative in Table 5-17 and 5-18.

For the fifth and sixth questions 5) Does KGL provide higher text coverage for daily conversation texts than the Ranking for Kanji in Written Japanese (KWJ) at all levels? and 6) Does KIS provide higher text coverage for daily conversation texts than KWJ at the basic level?, the answers are not straightforward. The main reason is that KWJ unexpectedly provides higher text coverage than KIS and KGL by 1.32% at the 01C level (Table 5-18). By scrutinizing the result of the word profiling, seven Kanji 思 (think), 私 (I, private), 方 (direction, side, part, or a function word used for choosing one out of two choices), 自 (for 自分 (self)), 知 (know), 通 (through, pass, street), 持 (own, have) are identified as placed at 01C in KWJ but at 02C in KIS and KGL. Particularly, the first three

have a remarkably high frequency. However, 私 and 方 are often written in Hiragana instead of Kanji. Therefore, the coverage depends on how the conversation texts are transcribed. If these words are transcribed in Hiragana, the results will be as expected, i.e. KGL and KIS provide higher coverage than KWJ.

Conversely, similar to WWJ (the Word Ranking for Written Japanese) in VDRJ, KWJ will be a good ranking for learners who do not need everyday conversation but only want to read (and write) formal Japanese texts.

Lastly, I would like to discuss the proportion of tokens by character types. The proportion of Hiragana increases in the order of academic journal texts (= technical natural science texts) (43.1%), (general) social science texts (55.1%), literary works (65.6%) and conversation (76.5%) (Table 5-7 to 5-10). These results suggest that the proportion for Hiragana may possibly be an index for informality. Or the proportion for Kanji can be an index for formality. To examine this prediction, the proportion of characters tokens by type of character is computed (Table 5-19).

Table 5-19 Proportion of Characters Tokens by Type of Character in the Order of the Ratio for Hiragana

Genre	LW	IF	LP	AH	BM	SE	HE	PL	ST	EC
Roman alphabet	0.1	2.8	1.5	0.5	1.1	0.8	0.7	1.1	2.5	1.5
Hiragana	**66.1**	**60.7**	**59.7**	**58.6**	**56.2**	**56.2**	**52.8**	**50.9**	**50.6**	**50.3**
Katakana	5.3	9.6	5.6	8.1	9.7	6.5	7.2	5.4	11.7	8.8
Kanji	28.5	26.9	33.1	32.7	32.9	36.4	39.1	42.6	35.1	39.4
Others	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1	0.1
Total	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0	100.0

* The sign for long vowels '—' is included in 'Katakana'.

LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.

As expected, the order of the proportion for Hiragana is similar to the order of the proportion for Japanese-origin words shown in Table 4-8 in Chapter 4 (Table 5-20).

Table 5-20 Rankings of the Orders of Ratios for Hiragana and Japanese-origin Words by Genre

Genre	LW	IF	LP	AH	BM	SE	HE	PL	ST	EC
Hiragana Ranking	1	2	3	4	5	6	7	8	9	10
Japanese-origin Word Ranking	1	2	4	3	5	7	6	10	8	9

LW: Literary Works/Imaginative Texts, LP: Languages, Linguistics and Philosophy, HE: History and Ethnology, AH: Arts and Other Humanities, PL: Politics and Law, EC: Economics and Commerce, SE: Sociology, Education and Other Social Issues, ST: Science and Technology, BM: Biology and Medicine, IF: Internet Q & A Forum.

As a feature of Japanese language, character types are related to the word origins, by which register variations can be identified as discussed in Chapter 4.

5.6 Conclusion of Chapter 5

In this chapter, I indicated the necessity for new character lists based on a new character database, and then described how I created the Character Database of Japanese (CDJ) and the character lists derived from the database. CDJ is the first Japanese character database made from large corpora composed of books and the internet-forum sites, which contain approximately 33 million running words in total.

After creating the database, its validity was examined. The main findings in this chapter are as follows.

1) The correlation between the frequency (*Fw*) and adjusted frequency (*Uw*) is very high. The distribution of characters is not as uneven as words.
2) The character ranking KWJ (the Ranking for Kanji in Written Japanese), where the rankings are made purely from usage coefficients computed based on the frequencies in the book and internet-forum texts, show higher correlations with F-JLPT (the former Japanese Language Proficiency Test) Kanji lists and Grades (the Japanese primary school Kanji grades) than frequencies in newspaper texts. KIS (the Ranking for Kanji

for International Students) and KGL (the Ranking for Kanji for General Learners) show even higher correlations with F-JLPT Kanji lists and Grades than KWJ.

3) KIS and KGL provide higher text coverage than F-JLPT Kanji lists.
4) The best order for learning Kanji will be different depending on the purpose. KIS will work better for students or academics than KGL, while KGL will work better for conversation than KIS. KWJ will only suit learners who do not need to learn daily conversation but only need to read (and write) Japanese.
5) The proportions of character types in different genres are considerably different. The proportion of Hiragana or Kanji may be an index for informality/formality.

Overall, the rankings in CDJ (i.e. KWJ, KIS and KGL) are shown to be valid and useful for learners and teachers of Japanese. Most of the findings in this chapter, as expected, are similar to the findings with word rankings in VDRJ described in Chapter 3. An additional question is: Is there any discrepancy between rankings for words and rankings for Kanji used in the words? This question is examined in the next chapter.

Chapter 6 Investigating the quantitative relationship between words and characters in Japanese

6.1 Introduction

In this chapter, the mathematical relationship between words and characters is examined using the databases of Japanese vocabulary and characters developed in Chapter 3 and Chapter 5 respectively.

It is widely believed that the burden of learning Japanese vocabulary is relatively heavy compared to other languages because more words are required to gain a certain level of text coverage (Akimoto, 2002; Kindaichi, 1981, 1988; Nagano, 1995; Sato, 1999). Text coverage is the coverage of word tokens in a text. The most frequent 1,000 words cover approximately 60% of Japanese magazine texts (National Language Research Institute, 1962, 2006)[72], while the most frequent 1,000 words cover over 70% in English (e.g. Carroll, Davies, & Richman, 1971). To reach 95% and 98% coverage, 9,500 and 20,000 words (lexemes including proper nouns) are required respectively in Japanese (Matsushita, 2011), while only 5,000 and 9,000 word families including proper nouns are required respectively in English (Nation, 2006).

We should note that the unit of counting for Japanese in Matsushita (2011) is the lexeme while the unit of counting in Nation (2006) is the word family. The unit 'lexeme' is defined by UniDic (Den et al., 2009) which is a digitized dictionary used for morphological analysis and word segmentation in Japanese. The 'short unit' (短単位) of the lexeme, which is the only currently available unit on the computer programme, is an inclusive unit which includes conjugated forms of verbs (e.g. 読む 'yomu' and 読み 'yomi' (read)), phonological variations (e.g. やはり 'yahari' and やっぱり 'yappari' (still, after all, as expected)), orthographical variations (e.g. 足 and 脚 'ashi' (foot, leg) and the combination

[72] As I showed in Chapter 4, text coverage in Japanese is not always as low as generally believed. This issue is to be discussed later in this chapter.

of two minimal units (e.g. 受け入れる 'uke-ireru' (accept)). The *Suru*-verb (e.g. 編集する (edit)) is divided into two units, namely the stem and *-suru* (e.g. 編集／する). Affixes (e.g. 非 'hi-' (non-) and 員 '-in' (member of)) are counted as a unit.

The unit 'word family' adopted by Nation (2006) is set at the Level 6 of Bauer & Nation (1993) which is also an inclusive unit including derived words with frequent affixes and 'regular but infrequent affixes'. For example, members of *abbreviate* are: *abbreviate, abbreviates, abbreviated, abbreviating, abbreviation* and *abbreviations*.

Despite the fact that the lexeme and the word family are different units, both units are considerably inclusive. Yet, why is the required number of words to gain 95% or 98% text coverage so different between Japanese and English? Is the vocabulary learning burden of Japanese really heavier than that of English?

One possible and widely-spread explanation is that many groups of words with different word-origins (語種 'goshu') but similar meanings make Japanese vocabulary larger (Akimoto, 2002; Kindaichi, 1981, 1988; Nagano, 1995)[73]. For example, the words きまり 'kimari' (Japanese-origin), 規則 'kisoku' (Chinese-origin) and ルール 'ru^ru' (Western-origin) are all correspond to the English noun 'rule'.

Nevertheless, there are some questions about the claim of Japanese lexical diversity and the explanation for it. First, the method for the text coverage measure in NLRI (1962), which is cited in many articles and book chapters (e.g. Akimoto (2002), Sato (1999), Tamamura (1984)), is questionable. The text coverage in NLRI (1962) does not include function words. In addition, it is based on magazine texts. As I showed in 4.2.3 in Chapter 4, the text coverage in Japanese books and internet-forum texts is at a similar level to English at least if function words are all included in the coverage. Second, there are also many English synonyms with different word-origins e.g. *liberty/freedom* and *spirit*/*soul*.

[73] All of these books cite Iwabuchi (1970), which is out of service now, to make the claim that Japanese vocabulary is ampler than other languages.

Therefore, the fact that different words with different word origins are used for one meaning is not a persuasive explanation for Japanese lexical diversity.

Third and more essentially, many transparent compounds composed of Kanji, which is one of the major features of Japanese vocabulary, may account for the lower text coverage. For example, the word 春季 'shunki' is a low-frequency word ranked at 28,587 in Matsushita (2011), while the word 春 'haru' and 季節 'kisetsu', both of which share Kanji with the word 春季, are high-frequency words ranked at 1,019 and 1,955 respectively in Matsushita (2011). Even though a learner does not know the word 春季, it is not difficult to infer the meaning of it if s/he knows the meanings of 春 and 季節. In other words, the meaning of the word 春季 is transparent. For those words, learners only need to understand the meanings of the components and the word formation rules, either implicitly or explicitly.

An American comedian Patrick Harlan (known by his nickname Pakkun), who has a good command of Japanese, once made a comment on his Japanese vocabulary learning as follows.

> "After learning a certain numbers of Kanji, I felt much easier to gain vocabulary. Many Kanji are applicable to many words. After learning 100 words, you can acquire another 100 faster. After learning 500, you can gain another 500 or 1,000 twice or three times faster."
>
> "After learning Kanji, you can understand a new word by analysing the meanings of the component Kanji. For example, 冷 'rei' of 冷蔵庫 'reizouko' (refrigerator) means 'to cool', 蔵 'zou' is also read as 'kura' which means 'storehouse', and 庫 'ko' sounds like a 'storeroom' as it is used for 車庫 'shako' (garage). You can somehow make out the meaning of the whole word by combining the meanings of the component Kanji."
>
> (Harlan, 2011. Translated by the author of this thesis.)

His comment seems to contain the key for the current question. That is to say, a limited number of Kanji will make up a large vocabulary, which reduces the learning burden of Japanese vocabulary.

The most basic unit for meaning and syntax is the word. Words still seem to be a more important level of learning than individual characters. Nevertheless, there may be some high-frequency Kanji which are used for many low-frequency words. If so, there may be basic Kanji which should be learned at an early stage but are not used for high-frequency words. This issue is related to the central concern of this research: the most efficient learning order of Japanese vocabulary.

6.2 Research questions

The main research questions (MRQs) are repeated below.

MRQs): In what order should learners of Japanese as a second language learn words and characters in order to be able to read Japanese? How will the order vary according to the purpose of learning?

To estimate the true learning burden of Japanese vocabulary and to think about an efficient order for learning Japanese vocabulary, I set the two sub-research-questions (SRQs) for this chapter as follows. (The SRQ number follows the previous section.)

SRQ 17) How many 'characters' do learners need to learn to attain a certain level of text coverage of 'words'?

Note that it is not to check the simple text coverage by characters as in previous studies (Chikamatsu, Yokoyama, Nozaki, Long, & Fukuda, 2000; NLRI, 1963;

Tamaoka, 2004). To know the meaning of a single character 節 is NOT enough to understand the meaning of 季節. In addition, the coverage to be examined in this chapter also includes the coverage by words composed of Roman alphabet, Hiragana and Katakana as they are generally learned before Kanji.

SRQ 18) Do the characters which provide a certain level of text coverage (in SRQ 17) cover all the high frequency words? If not, what Kanji are further required to cover the words? In other words, is there any discrepancy between the word frequencies and character frequencies?

6.3 Method

For the sub-research-question 1-17), computing the coverage of word tokens by different numbers of characters follows the steps shown below.

1) Calculate character frequencies in the Balanced Corpus of Contemporary Written Japanese (BCCWJ) 2009 monitor version (NINJAL, 2009)[74]
2) Add a 'learning order ranking' to each character
    I. Rank the types of characters as Roman alphabet, Hiragana, Katakana and Kanji[75]
    II. Rank Kanji by frequency[76]

[74] For the details of BCCWJ 2009 monitor version used for this study, see 3.3.2.

[75] The category 'Kanji' includes some signs such as 々 which indicates repeating the previous Kanji.

[76] The adjusted frequency (*U*) is also a possible index to order Kanji; however, the pure frequency (*F*) is used for this chapter. The reasons are as follows. 1) It is easier to interpret the results without the factor of dispersion. For example, when the former Japanese Language proficiency Test Kanji lists are made, only frequency was taken into account as objective data. 2) It is easy to compare the results with other frequency data such as the frequency in newspapers. 3) Even if *F* is used to order Kanji, the overall rankings are not very different from the rankings by *U*. *F* and U have a very high correlation at .99 or higher, and among the most frequent 2,000 Kanji, there are only 162 Kanji which have a ranking gap of 100 or more between the *U* ranking and the *F* ranking (See 5.5.2).

3) List all words in their orthographic forms (書字形), i.e. the word types, of the 'short unit' (短単位) defined by UniDic (Den et al., 2009) in BCCWJ

   Note that the unit of counting for this chapter is not the lexeme but the word type. For this chapter, it is essential to identify which characters are used because the relationship between words and characters is the main concern.

   i.e. 書く 'kaku' / 書か 'kaka' / かく 'kaku' (write), or 足 'ashi' (foot) / 脚 'ashi' (leg) are counted as different 'orthographic forms' or 'types' (but as one 'lexeme').

4) Separate each word into characters

5) Add the learning order ranking to each character

6) Calculate the text coverage by filtering the character of the words by learning order ranking. For example, if a word is composed of two characters which are ranked at 300 and 500 respectively; the word will remain in the list when the filtering level is set at character ranking 600 or higher. The word will be filtered off if the filtering level is set at 400, as one of the characters of the word is ranked lower than the set level.

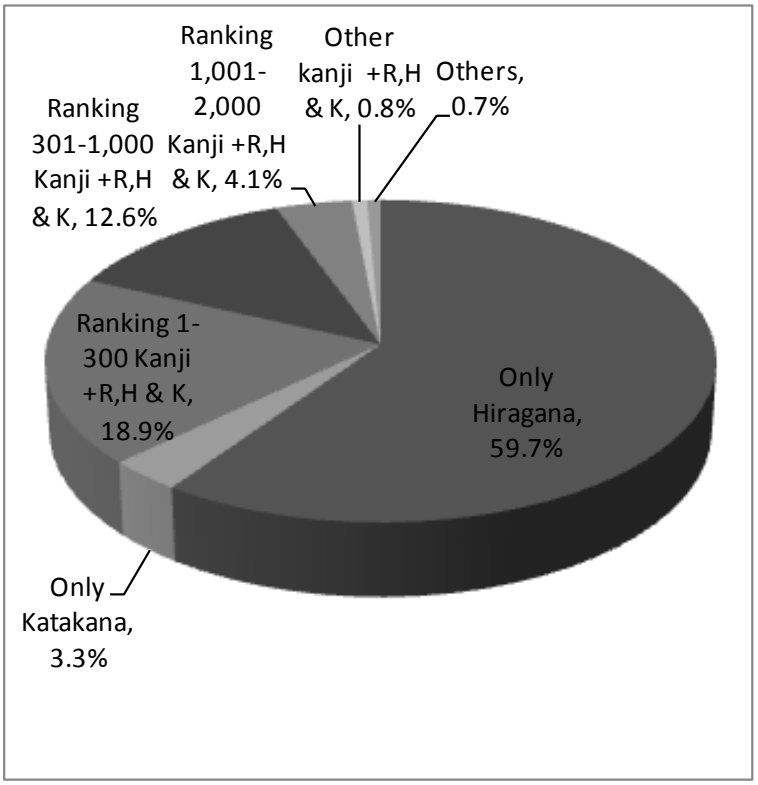
For the sub-research-question 1-18), the number of Kanji by Kanji frequency and levels in CDJ and the former Japanese Language Proficiency Test (F-JLPT) Kanji levels are counted to check if the JLPT Kanji are ranked properly. If there are identified words which are not covered by the high-frequency Kanji, then check what Kanji are used in those words.

6.4 Results

The first question (SRQ 17 shown in 6.2) in this chapter is: How many 'characters' do learners need to learn to attain a certain level of text coverage of 'words'? As shown in Graph 6-1 and Table 6-1, Hiragana alone covers almost 60% of the tokens. Half of the

tokens are function words. 3.3% of the tokens are covered only by Katakana, that is, one out of 30 tokens is a Katakana word. On average, 64% of the words are covered only by the phonographic characters (Hiragana, Katakana and Roman alphabet). 82% of the words are covered by phonographic characters plus the most frequent 300 Kanji. The most frequent 100 Kanji cover 10.1% of the text. The second most frequent 100 Kanji cover 5.2% and the third cover 3.6%. As the Kanji frequency level goes down, coverage by each 100 Kanji also decreases. To gain 95 to 96% text coverage, which is the proposed threshold level for reading comprehension in Japanese (Komori et al., 2004), phonographic characters (i.e. Hiragana, Katakana and Roman alphabet) plus 1,000 to 1,100 Kanji are required. To gain 98% coverage, which is the desired text coverage level proposed by Hu & Nation (2000), phonographic characters plus 1,500 Kanji are required.

Graph 6-1 Text Coverage of BCCWJ by Word Tokens by Character Types



BCCWJ: The Balanced Contemporary Corpus of Written Japanese, 2009 monitor version (NINJAL, 2009)

Table 6-1 Number and Proportion of Word Tokens (Orthographic Forms) and Text Coverage by Character Types (+Level of Kanji) in Japanese

Type of Chracter (+Level of Kanji)(*) R: Roman alphabet, H: Hiragana, K: Katanaka	Number of Word Types (Orthographic Forms) Covered by the Characters	Cumulative Number of Word Types (Orthographic Forms) Covered by the Characters	Text Coverage by the Word Tokens	Cumulative Text Coverage by the Word Tokens	Text Coverage by the Characters	Cumulative Text Coverage by the Characters
Only Roman alphabet	17,712	17,712	0.7%	0.7%	1.1%	1.1%
Only Hiragana (*)	20,272	37,984	**59.7%**	60.4%	51.9%	52.9%
Mixture of R & H	1	37,985	0.0%	60.4%	0.0%	52.9%
Only Katakana (*)	49,349	87,334	**3.3%**	63.6%	**7.3%**	60.2%
Mixture of R/H/K	625	87,959	0.0%	**63.6%**	0.0%	60.2%
Ranking 1- 100 Kanji +R,H & K	**7,187**	95,146	**10.1%**	73.8%	**9.7%**	70.0%
Ranking 101-200 Kanji +R,H & K	**7,360**	102,506	**5.2%**	79.0%	5.8%	75.8%
Ranking 201-300 Kanji +R,H & K	**7,318**	109,894	**3.6%**	**82.6%**	4.1%	79.9%
Ranking 301-400 Kanji +R,H & K	6,636	116,530	2.8%	85.4%	3.3%	83.1%
Ranking 401-500 Kanji +R,H & K	6,830	123,360	2.6%	88.0%	2.9%	86.0%
Ranking 501-600 Kanji +R,H & K	6,820	130,180	2.0%	90.0%	2.4%	88.4%
Ranking 601-700 Kanji +R,H & K	6,585	136,765	1.6%	91.6%	1.8%	90.2%
Ranking 701-800 Kanji +R,H & K	6,393	143,158	1.4%	93.0%	1.6%	91.8%
Ranking 801-900 Kanji +R,H & K	6,186	149,344	1.1%	94.1%	1.4%	93.2%
Ranking 901-1,000 Kanji +R,H & K	5,427	154,771	1.0%	**95.1%**	1.2%	94.4%
Ranking 1,001-1,100 Kanji +R,H & K	4,703	159,474	0.8%	**96.0%**	1.0%	95.3%
Ranking 1,101-1,200 Kanji +R,H & K	4,262	163,736	0.7%	96.6%	0.8%	96.1%
Ranking 1,201-1,300 Kanji +R,H & K	4,222	167,958	0.6%	97.2%	0.7%	96.8%
Ranking 1,301-1,400 Kanji +R,H & K	3,691	171,649	0.5%	97.7%	0.5%	97.4%
Ranking 1,401-1,500 Kanji +R,H & K	3,541	175,190	0.4%	**98.1%**	0.4%	97.8%
Ranking 1,501-1,600 Kanji +R,H & K	2,909	178,099	0.3%	98.4%	0.4%	98.2%
Ranking 1,601-1,700 Kanji +R,H & K	2,793	180,892	0.3%	98.6%	0.3%	98.5%
Ranking 1,701-1,800 Kanji +R,H & K	2,554	183,446	0.2%	98.9%	0.3%	98.7%
Ranking 1,801-1,900 Kanji +R,H & K	2,164	185,610	0.2%	99.0%	0.2%	98.9%
Ranking 1,901-2,000 Kanji +R,H & K	1,993	187,603	0.2%	99.2%	0.2%	99.1%
Ranking 2,001-2,100 Kanji +R,H & K	1,933	189,536	0.1%	99.3%	0.1%	99.3%
Ranking 2,101-2,200 Kanji +R,H & K	1,495	191,031	0.1%	99.4%	0.1%	99.4%
Ranking 2,201-2,300 Kanji +R,H & K	1,427	192,458	0.1%	99.5%	0.1%	99.5%
Ranking 2,301-6,323 Kanji +R,H & K	*15,373*	*207,831*	*0.5%*	*100.0%*	*0.5%*	*100.0%*
Ranking 1-6,323 Kanji +R,H & K	**207,831**	**207,831**	**100.0%**	**100.0%**	**100.0%**	**100.0%**

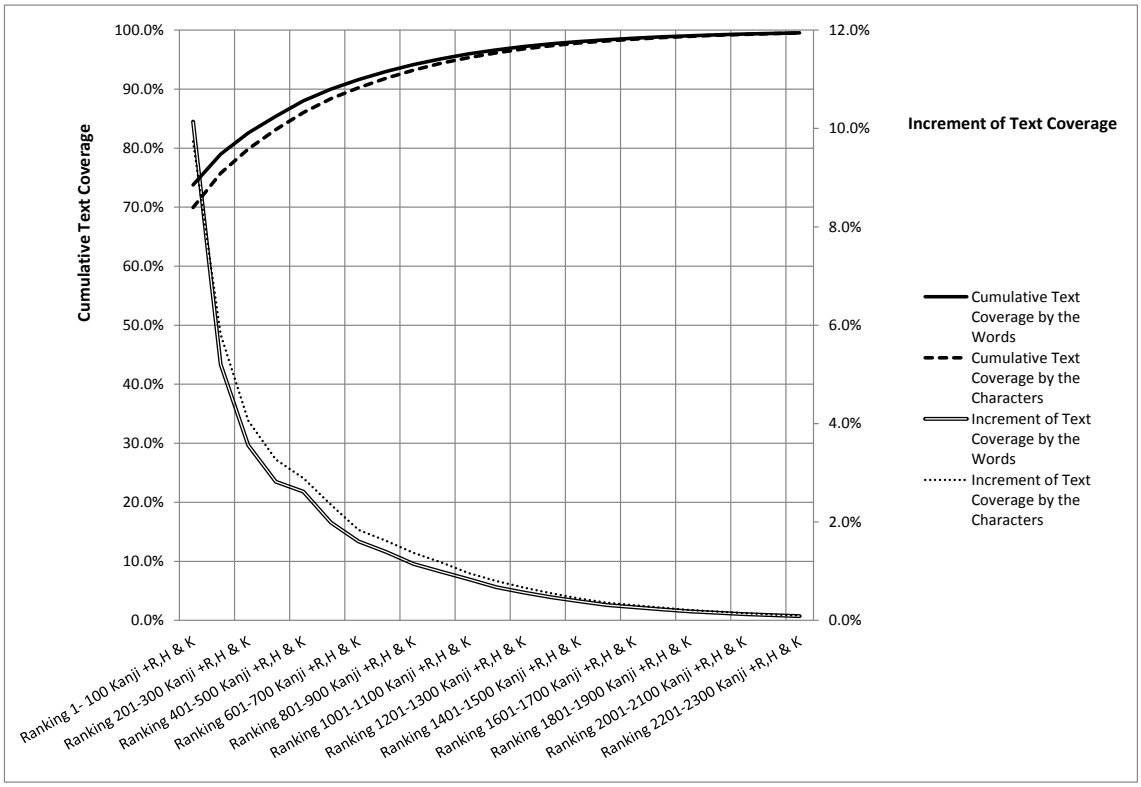
*Hiragana and Katakana include long vowel sign and Kanji includes the other signs.

*Rankings of Kanji are based on CDJ (The Character Database of Japanese) (See Chapter 5).

Learning 100 Kanji in the most frequent 1,000 Kanji means potential understanding of 6,000 to 7,000 types (orthographic forms). For example, as shown in table 6-1, the most frequent 100 Kanji are used for 7,187 word types. The second 100 are used for 7,360 types. 6,000-7,000 word types are equivalent to 3,000 to 4,000 lexemes. The higher the Kanji ranking level is, the more types and tokens are covered by the Kanji. In particular, within the most frequent 300 Kanji, each 100 Kanji contribute to more than 7,000 types. As the Kanji frequency level goes down, types composed of the Kanji decrease in number. This means that the higher a Kanji frequency level is, the more words the Kanji will occur in.

Graph 6-2 shows the increased and cumulative text coverage by words and characters in Japanese at different Kanji frequency levels.

Graph 6-2 Increment of Text Coverage and Cumulative Text Coverage by Words and Characters in Japanese at Different Kanji Frequency Levels



The cumulative text coverage by words is greater than the coverage by characters, while the increased coverage by learning 100 Kanji is greater in a character-based count than in a word-based count except for the most frequent 100 Kanji level (10.1% in a word-based count and 9.7% in a character-based count (See also Table 6-1). This is mainly because the average length of words at the highest-frequency level is shorter than the average length of the whole vocabulary. This is particularly striking in Katakana. The word-based coverage by Katakana is only 3.3% while the character-based coverage is much higher at 7.3% (Table 6-1).

It is clear from this data that examining text coverage merely by characters is different from the coverage by words. That is why the coverage by words depending on the

number of characters is necessary to estimate how many characters (especially Kanji) are required to learn to gain a certain level of reading comprehension.

Another question for this chapter (SRQ 18 shown in 6.2) is: Do the characters which provide a certain level of text coverage (in SRQ 17) cover all the high frequency words? If not, what Kanji are further required to cover the words? In other words: Is there any discrepancy between the word frequencies and character frequencies? Or more critically: Can low-frequency Kanji be a barrier to learning high frequency words? To answer this question, the number of Kanji by the frequency levels for Kanji in CDJ and the former Japanese Language Proficiency Test (F-JLPT) Kanji levels are counted first as shown in Table 6-2[77].

As shown in italics in Table 6-2, there is a narrow gap between the frequency level in CDJ and the former Japanese Proficiency Test Kanji Level. As shown in bold in Table 6-2, among the most frequent 1,000 Kanji, more than 800 Kanji are covered by the Kanji at the former JLPT Level 4, 3 and 2. The remaining 173 Kanji are shown in Table 6-3.

Table 6-2 Number of Kanji by the Frequency Levels for Kanji in CDJ and the Former Japanese Language Proficiency Test (F-JLPT) Kanji Levels

Kanji Frequency Level	F-JLPT Level				Subtotal	Others	Total
	4	3	2	1			
1-100	**46**	**36**	**18**	0	100	0	100
101-300	**34**	**58**	**104**	*3*	199	*1*	200
301-1,000	**23**	**75**	**433**	*166*	697	*3*	700
1,001-2,000	0	*12*	*183*	658	853	147	1,000
Others	0	0	*1*	190	191	4,140	4,331
Total	103	181	739	1,017	2,040	4,291	6,331

* This table reflects the revision of the former Japanese Language Proficiency Test Kani lists in 2001.

Taking account of the data shown in Table 6-1 and 6-2, it is estimated that more than 96% of the word tokens in general texts (i.e. the Balanced Contemporary Corpus of Written

[77] For a more detailed distribution, see Table 5-8 in Chapter 5.

Japanese 2009 monitor version which is the corpus CDJ is created from) will be covered by Hiragana, Katakana, Roman alphabet and 1,200 Kanji which are all Kanji at the former Japanese Language Proficiency (F-JLPT) Test Level 4, 3, and 2 (total 1023 Kanji) plus the most frequent 200 Kanji at the F-JLPT Level 1.

Table 6-3 The Most Frequent 173 Kanji in the Former Japanese Language Proficiency Test (F-JLPT) 'Level 1' or 'beyond Level 1' ('Kyuugai')

Within the Top 300	々 義 態 郎
Within the Top 1000	士 氏 視 素 護 離 証 企 誰 提 姿 井 統 ヴ 振 吉 策 影 頃 紀 為 宮 江 派 藤 僕 従 系 衛 皇 松 隊 施 我 及 織 響 遺 宗 昭 撃 株 源 養 項 興 裁 沢 端 障 激 弁 俺 益 嫌 佐 眼 密 載 己 債 訳 之 症 納 請 挙 貴 徳 推 岡 描 崎 抗 属 盛 監 傷 患 徴 創 街 掛 援 衆 模 敵 津 拠 継 隠 称 尾 聖 鮮 厳 攻 妙 融 丈 筋 帝 秘 敷 伊 驚 射 壊 刑 染 功 訴 跡 討 幕 扱 脱 範 契 弾 診 詳 房 避 酸 倉 充 繰 典 儀 至 削 博 瞬 阪 縁 憲 択 就 聴 握 詩 秀 柄 浜 滅 惑 踏 華 闘 微 雄 維 隣 如 審 誘 賀 郷 霊 釈 黙 魔 携 遣 掲 艦 剣 致

* 々 is a sign to indicate repeating the previous Kanji
* ヴ is a Katakana only to be used for the sound 'v' in loan words. This character is not included in the category of Katakana as it is generally not to be taught at the elementary level.

95% coverage of the word tokens in the books and internet-forum texts requires the most frequent 9,446 lexemes (ranked by the adjusted frequency *U*) or the most frequent 20,749 types (orthographic forms)[78]. As shown in Table 6-1, 95% of word tokens are covered by the most frequent 1,000 Kanji plus Hiragana, Katakana and Roman alphabet; however, this count contains low-frequency words. Therefore, the Kanji used for the most frequent words but not listed in the equivalent Kanji level should be checked. Within the most frequent 9,500 lexemes, 1,700 lexemes are estimated to require Kanji beyond the 1,000 Kanji level. By checking the words which are within the most frequent 9,500 lexemes but composed of Kanji beyond 1,000 Kanji level, two types of Kanji are identified.

[78] This lexeme count is shown in Table 4-2 in 4.2 (Chapter 4).

One type is Kanji only used for high-frequency words but not frequently used for other words. Here are some examples. (The bold characters are this type of Kanji.)

比**較**　記**憶**　**批**判　**距**離　指**摘**　**希**望　分**析**　**韓**国

基**礎**　**誕**生　**監**督　**雰**囲気　**卒**業　洗**濯**　細**胞**

This type of Kanji should be learned when needed even if it is not frequent as an individual Kanji.

The other type is the Kanji used for high-frequency words but which are often also written in Hiragana or Katakana. Here are some examples. (Another frequent orthographic form is in the brackets.)

即ち（すなわち）　駄目（だめ／ダメ）　奴（やつ／ヤツ）

凄い（すごい）　頑張る（がんばる／ガンバル）　挨拶（あいさつ）

嘘（うそ／ウソ）　煙草（タバコ）　匂い（におい）　只（ただ）

是非（ぜひ）　無駄（むだ／ムダ）　喧嘩（けんか）　噂（うわさ）

伺う（うかがう）　頁（ページ）　又（また）

These Kanji are less important than the first type as learners are generally not required to write them; however, it would be better to be able to recognize these characters for reading.

6.5 Discussion

As introduced in previous chapters, for general texts, learners can attain more than 70% comprehension with 95 to 98% coverage (For English, see Hu & Nation (2000), Laufer & Ravenhorst-Kalovski (2010); for Japanese, see Komori et al., (2004)). According to Zipf's law, high-frequency words account for much more text coverage than low-frequency words

(Zipf, 1949). Zip's law can also be applied to Kanji learning. That is, learning Kanji by order of frequency is much more efficient to gain higher text coverage. To read authentic Japanese without dictionary use, learners will need to learn Kanji up to the 1,000 Kanji level at least. The most frequent 1,000 to 1,500 Kanji might be enough for general purposes, with occasional use of a dictionary. However, this may also mean that learning Kanji without reaching the threshold level is of little use. Therefore, keeping motivation for learning Kanji up to the threshold level is important as there are few authentic passages which can be understood without this number of known Kanji.

To attain 95% coverage, 1,000 Kanji are required; however, there are some important words not covered by the top 1,000 Kanji. In other words, some low-frequency Kanji are used for high-frequency words. Many of those Kanji has low productivity, that is, they are rarely used to form other words (e.g. 雰 for 雰囲気 'fun'iki' (atmosphere), 卒 for 卒業 'sotsugyou' (graduation), 濯 for 洗濯 (washing clothes)). To cover the most frequent 9,500 lexemes, a further 200 to 500 Kanji are estimated to be required.

Certainly, the burden of learning Japanese 'characters' is heavier than most other languages. Nevertheless, despite the fact that the text coverage is lower than English at all word frequency levels, the burden of learning Japanese vocabulary may be rather lighter once the learner knows 1) 1,000 to 1,500 characters, 2) word formation rules of Kanji, and 3) metaphors of Kanji compounds.

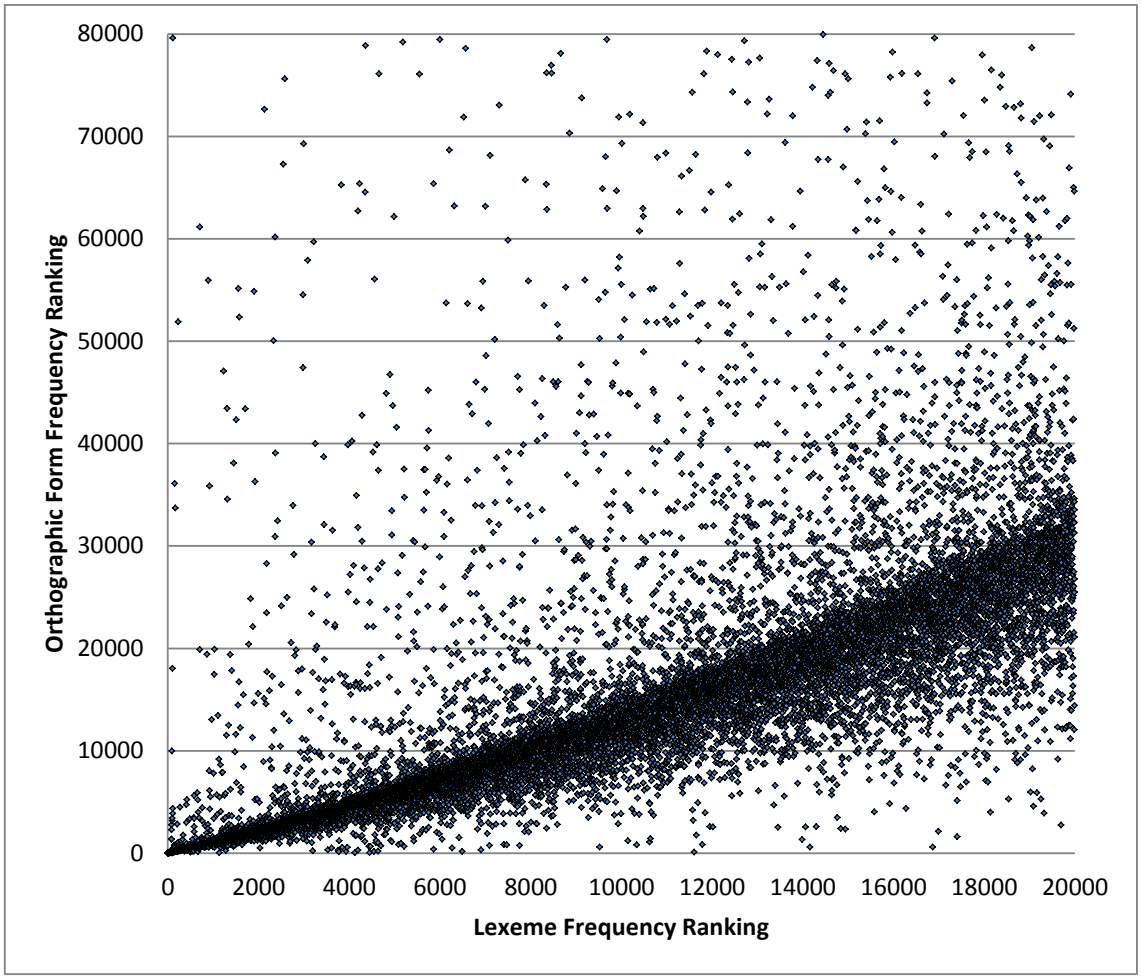
The third point, to understand a metaphorical meaning of Kanji compound, is, for example, to understand the word 入門 'nyuumon' which means 'the first step' or 'start training' from the meanings of the components 入 'nyuu' (enter) and 門 'mon' (gate).

In other words, it is possible that the number of 'units of learning Japanese vocabulary' is not so many as generally perceived. It will also be important for students and teachers to learn or teach the association of different readings (typically the On-reading and Kun-reading) of each Kanji, which will reduce the burden of learning Japanese vocabulary.

For example, the word 入門 ‘nyuumon’ (first step, start training), which is an On-reading (Chinese-origin) word, is composed of 入 and 門. The first Kanji 入 is also used for the word 入る ‘hairu’ (enter) which is a high-frequency Kun-reading (i.e. Japanese-origin) word ranked at 117 in the Word Ranking for Written Japanese (WWJ) in VDRJ. The second Kanji 門 is ranked at 1,476 in WWJ. Both words are much more likely to be learned earlier than the word 入門 ranked at 6,369 in WWJ. Even if a learner does not know the word 入門, if s/he knows 入る and 門, and is able to guess the meaning, learners can increase Japanese vocabulary much easier and faster. Without this kind of association, learners have to learn many related words separately. Thus, the association of words mediated by a Kanji, especially the relationship between the On-reading and Kun-reading with a Kanji, is very important in learning Japanese vocabulary.

Lastly, I would like to mention the quantitative relationship between lexemes and orthographic forms (i.e. word types). As I mentioned in 6.1, the unit lexeme includes conjugated forms of verbs. For example, 書く ‘kaku’, 書か ‘kaka’ and かく ‘kaku’ are sub-members of the lexeme ‘書く’ (write). The total number of lexeme members is 141,949 in VDRJ while the total number of orthographic forms is 207,831. The ratio of lexemes to orthographic forms is approximately 1:1.46. However, the relationship may not be linear. There are two possible reasons for this. One is that the conjugating words (i.e. verbs and adjectives) occur mainly in the high frequency range. The other is that the larger the corpus size, the greater the proportion of one-timers, which are typically proper nouns in the low-frequency range. Graph 6-3 shows the frequency rankings of orthographic forms (word types) and lexemes in VDRJ. The densest part is almost linear but slightly curved. A dot located far from the densest part means a rarely-used orthographic form of a lexeme. There are many forms of this type, though they do not account for a high proportion. We should be aware of those forms when we use the lexeme as a unit of counting.

Graph 6-3 Frequency Rankings of Orthographic Forms (Word Types) and Lexemes in VDRJ



6.6 Conclusion of Chapter 6

In this chapter, to answer the question if the learning burden of Japanese vocabulary is really as heavy as widely-believed, the required number of characters to attain certain levels of text coverage by words was investigated first. And then, the number and types of Kanji which do not have a high-frequency but are used in high-frequency words were identified. The main findings in this chapter are as follows.

1) 63% of the tokens of the Balanced Contemporary Corpus of Written Japanese (2009 monitor version) texts are covered without Kanji (but more than half of these tokens are function words).
2) To attain 95% coverage, 1,000 Kanji are required; however, some important words are

not covered by the most frequent 1,000 Kanji.

3) To cover those words, several hundred more Kanji will be required.

4) Most high-frequency and mid-frequency Japanese words are composed of a limited number of Kanji, therefore, the burden of learning Japanese vocabulary may not be heavy as expected from the text coverage studies, once the learner knows:

    1. the most frequent 1000 to 1500 characters.
    2. forms, meanings and compounding rules of Kanji.
    3. metaphors of Kanji compounds.
    4. create the links between different readings (e.g. On-reading and Kun-reading) of each Kanji.

I explored the lexical features of Japanese and the mathematical relationship between words and characters of Japanese by developing and analysing vocabulary and character databases from Chapter 3 to 6. Based on these, more detailed word tiers will be explored in Chapter 7 to answer the main research question.

Chapter 7 Exploring the word tiers of Japanese by extracting domain-specific words: In what order should learners learn groups of words?

7.1 Introduction

In Chapter 3, I created a database which includes different word rankings to meet different types of needs. To validate the rankings, I confirmed that different rankings provide different text coverage in different target domains. In Chapter 4, I examined how different the lexical features are between genres. The database has only three types of rankings; however, if a learner has a certain major domain or field, learning domain-specific words will be a more efficient way to gain text coverage.

A character database was developed in Chapter 5. Based on this, the importance of understanding word formation rules for reducing the burden of learning Japanese vocabulary was claimed in Chapter 6. However, the findings in Chapter 6 will not directly mention an efficient learning order of words, but only implies that some difficult or low-frequency Kanji should also be learned earlier and that some semantically transparent compounds can be 'skipped' as they can be counted as known words.

In this chapter, lexical domain-specificity is explored by extracting domain-specific words and checking text coverage in different types of texts by different types of words. Then, I will answer the main research questions for this whole thesis: "In what order should learners of Japanese as a second language learn words and characters in order to be able to read Japanese? How will the order vary according to the purpose of learning?"

Specifically, the following steps are to be taken. Firstly, the previous word list studies about basic vocabulary and domain-specific words in English and Japanese are reviewed. The concept of 'word tiers' is also briefly introduced. Secondly, the research questions for this chapter are proposed. Thirdly, some types of domain-specific words, namely academic words, limited-academic-domain words, literary words, are extracted. The features of those

groups of words are also discussed. Fourthly, using test corpora, text coverage by the extracted group of words is examined. Using a proposed index entitled Text Covering Efficiency (TCE) to evaluate a group of words as a source for covering a text, how the word tiers work in different type of texts is also examined. Then, the specific method for deciding the most efficient learning order of words according to the learner needs will be proposed. How learner's language background possibly affects the understanding of texts will also be discussed in terms of the proportion of word-origins of the domain-specific words. Remaining issues and a conclusion will follow these discussions.

7.1.1 Significant research

7.1.1.1 English word lists

In English language teaching, the vocabulary selection (or control) movement arose in the 1920s or 1930s (Richards, 2001, p 8; Schmitt, 2000, p 15). That mainly focuses on selecting basic vocabulary. The most important outcome was Michael West's General Service List (West, 1953). This list is the classic and influential English basic vocabulary list (Nation, 2001, p 11; Schmitt, 2000, p 16–17).

There are also many vocabulary lists serving for specialised uses (compact reviews are in Coxhead & Hirsh (2007, p 66–68) and Nation (2001, p 187–188, 198–203)). One of the most influential lists is the Academic Word List (AWL) (Coxhead, 1998, 2000) which provide notably higher text coverage in different genres of academic texts than in general texts. The 'academic words' in this list are different from technical words in that it is expected to provide high text coverage in any academic genre. In other words, 'academic words' are words which are commonly used frequently across a range of academic genres. There were similar attempts before the Academic Word List such as the University Word List (Xue & Nation, 1984). However, the Academic Word List consisting of 570 word families, which is fewer than the University Word List by over 200 hundred words,

provides high text coverage at 8.5 to 10.0% in academic texts (Coxhead, 2000). This coverage is at the same as or even higher level than the University Word List.

Technical vocabulary is selected from a more specific domain such as economics (Sutarsyah, Nation, & Kennedy, 1994), applied linguistics (Chung, 2003a; Chung & Nation, 2003) or medicine (Wang, Liang, & Ge, 2008). There are many discussions about technical vocabulary in needs analysis (Ward, 1999), theory and history (Castellví, 2003), selecting methods (Chung, 2003a, 2003b), useful indices (Chujo & Utiyama, 2006) and so on. Technical vocabulary has been studied not only for second language learning and teaching, but also for various purposes such as controlling the creation of new terms, standardization of terms, technical translation and so on. In this study, technical terms in a single academic field are not extracted; however, the level of specificity and methodological issues in selecting vocabulary[79] are related to this study.

From the viewpoint of the level of specificity, one attractive idea is extracting vocabulary which is located between academic words and technical vocabulary. Tajino, Terauchi, Sasao, & Maswana (2007) and Tajino, Dalsky, & Sasao (2009) propose incorporating a vocabulary learning programme at different levels of specialization of university curricula, such as learning 'academic words' for general academic purposes for the first year students, and then narrow down to 'arts' (文系 'bunkei') vocabulary or 'science' (理系 'rikei') vocabulary at the next step and so on. 'Arts' include humanities and social sciences. 'Science' includes medical sciences and physical sciences. The next step beyond these large disciplines is each major field such as law or pharmacy. A similar idea is realized in the Science-specific Word List (Coxhead & Hirsh, 2007) which focuses on a similar level to the science vocabulary in Tajino et al. (2009).

All of abovementioned vocabularies are for academic purposes except for the basic vocabulary. Besides the academic texts, one possibly specific domain is literary works;

[79] Methodological issues are mentioned in later sections of this chapter.

however, there seems no attempt to extract literary vocabulary. This may be because there have been no need for that, or there seems no specific vocabulary in the field because many literary works deal with a wide range of topics including daily-life ones. Nevertheless, there are literary words which are likely to occur more frequently in literary texts (at least in Japanese) such as 瞳 'hitomi' (eye). Outside literary works, the word 目 'me' is generally used for referring to eyes.

7.1.1.2 Japanese word lists

It is obvious that, in Japanese applied linguistic studies, there are many studies of basic words particularly for international students as well as many technical words for different fields. However, there are few studies of the vocabulary in-between, namely academic words and arts/science vocabulary.

The most influential vocabulary lists are the former Japanese Language Proficiency Test word lists (Japan Foundation & Association of International Education, Japan, 2002). As introduced in 3.3.1, words in these lists were selected subjectively by the expert committee. When they select the words, eleven types of textbooks and other references including one frequency list were compared and checked. Therefore, these lists were created from relevant studies at that time. The lists include words from Level 4 (elementary) to Level 1 (advanced). The Level 4 and Level 3 lists are the basic vocabulary lists which have a similar social impact in teaching Japanese to the General Service List (West, 1953) in teaching English[80]. Though it is primarily based on subjective judgement, as shown in 3.5, these lists provide higher text coverage in conversation texts than the frequency list made from the Balanced Contemporary Corpus of Written Japanese. This means that the frequency list made from a written corpus is less useful for daily-life needs. As there is no Japanese spoken corpus suitable for counting frequency currently, only for

[80] The number of words in the Level 4 & 3 lists is only around 1300.

daily-life conversation needs, subjective selection of basic vocabulary is still useful.

There are many attempts at selecting technical vocabulary for second language learners in various fields such as economics (Komiya, 1995; Oka, 1992), business administration (Terajima, 2010), chemistry (Komiya, 2005), agricultural science (Muraoka, Kagehiro, & Yanagi, 1997; Muraoka & Yanagi, 1995) and environmental engineering (Mizumoto et al., 2005).

As for the academic words, some researchers have pointed out the existence of a group of words which are common in different academic fields but not in basic daily-life domains. Fudano & Fukasawa (1995) use the term 'in-between expressions' (はざま表現 'hazama-hyougen') for the group of words and phrases. Fukao (2001) describes the words more specifically and precisely as "cross-disciplinary academic vocabulary which is located between daily-life vocabulary and technical vocabulary" (日常語に使用される語彙と専門用語との間に位置する専門分野を超えた学術的な語彙). Mizumoto & Ikeda (2003) use a simpler term 'basic technical terms' (基礎専門語 'kiso-senmon-go') to refer to a similar concept. Despite these indications of the existence of the academic words, there had been no attempts to extract the academic words before Sumi (2010) and Butler (2010).

Sumi (2010) mentions the usefulness of the English Academic Word List (Coxhead, 1998, 2000) and selected 434 words as the 'Basic Academic Terms' (学術基本用語 'gakujutsu-kihon-yougo'); however, the words in this list are at a much lower frequency level than the Academic Word List. 341 words (78.6%) out of 434 words are not included in the former Japanese Language Proficiency Test word lists. That is, most of the selected words are seemingly at an advanced or super-advanced level. It is questionable to call the terms 'basic' even for academic purposes. Also, the words in the Basic Academic Terms tend to be more related to social sciences and humanities, especially on modern thought, because the selection of the terms mainly relies on five word lists for preparing for the modern Japanese exams held for university admission. This seems problematic as the

selected words are not really common in different academic fields. In sum, the Basic Academic Terms have different features from the Academic Word List mainly in levels and domains, which are caused by the different selection methods. It is hard to tell how useful the list is as Sumi (2010) does not provide text coverage data in any texts.

Butler (2010) also mentions the usefulness and the selection method of Coxhead's Academic Word List and selected 1,230 words as Japanese Academic Vocabulary for Elementary and Junior High School Students. It is obvious from its name that the list is not for adult learners. The selection method is similar to the method used for making the Academic Word List; however, the corpus used for making this list was a textbook corpus which contains textbooks used for nine subjects in primary and junior high schools (Year 1 to Year 9) in Japan. Also, the selected words are adjusted by an expert committee consisting of teachers from primary to university levels, aimed at both first and second language learners. The main differences between this list and the Academic Word List are the number of words and the target learners. Butler (2010) also did not provide text coverage data in any texts.

All in all, there seems no Japanese word list equivalent to the Academic Word List whose text coverage is higher in a wide range of academic texts than in general texts. Both Sumi (2010) and Butler (2010) lack in quantitative evidence of the usefulness of the list. In Japanese studies, there seems to be no attempt at selecting Japanese literary words, either.

7.1.1.3 Needs and importance of the lists for domain-specific words

Whatever the target language is, one important point to be confirmed is the value and importance of selecting domain-specific words. As is discussed in 2.5.1, for the list of words common in different domains, for example the Academic Word List, there is a general debate about its needs. Ward (1999) and Hyland & Tse (2007) claim that it is more efficient to follow the order of word frequency in the learner's specialised field to read the

texts in the field rather than to learn academic words first.

Coxhead & Hirsh (2007) raised four reasons to show the value of the Academic Word List to answer the abovementioned question. The four reasons are: 1) EAP classrooms tend to group students by proficiency and/or undergraduate or postgraduate levels, 2) The lexical background knowledge of EAP students cannot be considered to be the same, 3) The first year tends to comprise a range of papers that may be core to several subject areas to be studied by the final years of study, and 4) Not all students enter university with a clear view of their path of study. All of the four reasons are, in short, related to the issue how a curriculum can match individual learner needs, readiness and background. If we can offer a programme which totally matches each individual learner, that might be better. Nevertheless, as Coxhead & Hirsh (2007) point out, a second or foreign language program is generally required to match with needs from a group of learners with different needs, readiness and background, and they are also generally expected to learn a wide range of disciplines as they move on to their major studies. This viewpoint is in line with Tajino et al. (2009, 2007) who claim vocabulary learning should go from a wider to narrower range of domains according to the learners' level of study, namely first year, undergraduate major and postgraduate studies.

As discussed in Chapters 3, 4 and 5, different genres admittedly have different lexical features. The closer the corpus is to the learner needs, the better the frequency list reflects the efficient order of learning. However, as discussed above, the ideas in Coxhead & Hirsh (2007) and Tajino et al. (2009, 2007) are practical and useful.

In addition, it is also important that the selected academic words also show the common lexical features of academic texts. Nation (2001) reviewed relevant studies and discussed the nature and role of academic vocabulary (p.194-196)[81]. Learning academic words inevitably involves how to manage academic information. In a wider sense, any type

[81] The nature of academic words is to be discussed specifically after extracting the Japanese academic words in 7.2.

of domain-specific words would somehow reflect the features of the domain, whether the domain is a wide one such as ‘academic texts’ or a more specific one such as medical texts.

Thus, this study first tries to extract different levels of academic vocabulary: from the more common academic words to less common ones. Also, as the corpus for this study contains a large proportion of literary texts, literary words will also be extracted as a trial. And then, how these words work in different genres will be tested by checking the text coverage of the test corpora.

7.1.1.4 Word tiers

For describing different groups of words, the term ‘tier’ is used in Beck & McKeown (1985) and Beck, McKeown, & Kucan (2002). They classify English vocabulary into three tiers in the American school education context where the majority of learners are first language learners. The three tiers roughly correspond to basic vocabulary, academic vocabulary and the others (low-frequency words). For this study, I use the term ‘word tiers’ to describe the whole Japanese vocabulary composed of different groups of words which are defined by ‘domain’ and ‘frequency level’, for example ‘intermediate literary words’ or ‘advanced academic words’. If a word is neither academic nor literary, I tentatively call it a ‘general’ word. Three domains are assumed for this study, namely general, academic and literary domains. Literary words are only selected from literary works (LW in VDRJ), i.e. imaginative texts, but not from technical texts in literature.

As one major topic for this chapter is academic vocabulary, I define the frequency levels by the Word Ranking for International Students (WIS) introduced in Chapter 3.

The domains and levels for this chapter are as follows.

<u>Domain</u>

- General / Academic / Literary

Level *****Assumed Known Words (proper nouns etc.) are not included

- Basic: the top 1,288 words = words in the Level 4 and 3 lists of the former Japanese Language Proficiency Test (F-JLPT)

  (The F-JLPT Level 4 words, which are all ranked at 681 or higher in WIS (the Word Ranking for International Students in VDRJ), will not be classified into different domains as they are very basic for most domains. F-JLPT Level 3 words are also basic, but some of them have domain-specificity.)
- Intermediate: ranked at 1,289-5,000
- Advanced 1: ranked at 5,001-10,000 (6K-10K)
- Advanced 2: ranked at 10,001-15,000 (11K-15K)
- Super-Advanced: ranked at 15,001-20,000 (15K-20K)
- 21K+: 20,001+ (21K+)

The method of extracting different tiers of words will be explained in 7.2.2 and 7.3.1.

7.1.2 Research questions

At the end (7.4.5) of this chapter, I will answer the main research questions (MRQs) as shown below.

MRQs): In what order should learners of Japanese as a second language learn words and characters in order to be able to read Japanese? How will the order vary according to the purpose of learning?

Specific sub-research-questions (SRQs) to be answered before answering the main research questions in this chapter are as follows. (The SRQ number follows the previous section.)

SRQ 19) What words are commonly used more frequently in different academic domains than in general texts?

SRQ 20) What words are commonly used more frequently only in a limited number of academic domains than in general texts? Are there limited-academic-domain words frequently used in one or two domain(s) such as 1) humanities and arts, 2) social sciences, 3) technological sciences and 4) biomedical sciences[82]? If yes, what words are those?

SRQ 21) What words are commonly used more frequently in literary works than in other types of texts?

SRQ 22) How high is the text coverage by different groups of words such as basic vocabulary, academic words and limited-academic-domain words and literary words in different types of texts? Does each group of words provide significantly higher text coverage in the target domain than in the other domains?

Based on the results of the questions above, various types of texts are analysed mainly by checking the proposed index entitled Text Covering Efficiency (TCE). This is to clarify register variations as well as to explore the most efficient learning order of words depending on the type of texts. The research questions for this purpose are shown below.

SRQ 23) What features does each text genre have in terms of its Text Covering Efficiency (TCE) of grouped words at each level?

SRQ 24) How can the efficiency in covering texts by a group of words be measured? How should the most efficient learning order of words be decided?

SRQ 25) How efficient is learning each group of words in covering texts in different

[82] 'Technological' and 'biological' natural sciences are respectively equivalent to 'physical' and 'medical' sciences in Tajino, Dalsky, & Sasao (2009).

genres?

7.2 Academic vocabulary

7.2.1 Classification of 'academic vocabulary'

As is to be explained in 7.2.2, the method for extracting academic vocabulary is to extract domain-specific words first from the four academic domains of humanities and arts, social sciences, technological natural sciences and biological natural sciences, and then check how many domains each extracted word is extracted from. For example, if a word is extracted from 3 domains, I call it a '3-domain word' here. Among the extracted 4-domain words, 3-domain words, 2-domain words and 1 domain words, the first two will be categorised as '(common) academic words' (AWs) as they will be used frequently for a wide range of academic fields, which will have similar features to the words in the Academic Word List (Coxhead, 1998, 2000)[83]. The latter two, i.e. 2-domain words and 1-domain words will be categorised as 'limited-academic-domain words' (LADs). I want to include all the words from the four categories as 'academic vocabulary' (Table 7-1). The 4-domain words and 3-domain words are expected to have similar lexical features to the words widely known as 'academic words'[84]; however, I call them 'common academic words' to avoid confusion. I use the terms as shown in Table 7-1.

[83] The Academic Word List is made from four large sub-corpora of arts, commerce, law and science. The construction of the sub-corpora is different from this study in that two of the four sub-corpora for AWL are from social sciences (i.e. commerce and law) and only one from (natural) science. The classification into the four large science domains adopted for this study basically follows Tajino, Dalsky, & Sasao (2009) and Tajino, Terauchi, Sasao, & Maswana (2007). Following this approach, any 3 domains out of the four domains must include at least one art (文系 'bunkei') domain and one (natural) science (理系 'rikei') domain. Thus, the common features among different types of academic domains are expected to be guaranteed for the extracted words.

[84] In Japanese, I named 'academic words' as 学術共通語(彙) 'gakujutsu-kyoutsuu-go(i)' (Matsushita, 2011) which literally means 'common academic words'. 'Limited-academic-domain words' can be literally translated into Japanese as 限定学術領域語 'gentei-gakujutsu-ryouiku-go'.

Table 7-1 Classification of academic vocabulary for this study

academic vocabulary	common academic words (AWs)	4-domain words
		3-domain words
	limited-academic-domain words (LADs)	2-domain words
		1-domain words

7.2.2 Method for extracting academic vocabulary

There are several ways to extract domain-specific words. For this study, I adopt a statistical index called log-likelihood ratio (LLR) (Dunning, 1993). There are three reasons for this decision which basically agree with Leech, Rayson, & Wilson (2001, p 16). One is that the log-likelihood ratio does not require a particular distribution pattern such as the normal distribution. Another reason is that, compared to other indices, the log-likelihood ratio will return a moderate result (For further discussion, see Chujo & Utiyama (2006)). In other words, the extracted words will be neither too specific nor too general. The last reason is that the log-likelihood ratio can be applied to comparing differently-sized target corpora. That is, log-likelihood ratio figures can be compared even if they are calculated from differently-sized corpora. This is very important as the sizes of the target corpora are considerably different in this study.

The construction of the whole corpus (NINJAL, 2009) is shown in Table 7-2 (=Table 3-4). (For more details, see 3.2.2.) The 'technical texts' shown in the table are identified by the C-code which is attached to each text file. If the C-code has '3' at the thousands digit, that means the book which contains the text, is written for experts. Therefore, the text from the book is identified as a technical text.

Table 7-2 Number of Types and Tokens by Field in VDRJ (=Table3-4) *The corpus is made from books and internet-forum sites contained in NINJAL (2009).

Field	Code for the ten domains	G (General)		T (Technical)		Total	
		G Type	G Token	T Type	T Token	Type	Token
Literary Works/Imaginative Texts	LW	**68,446**	**8,251,999**	--	--	**68,446**	**8,251,999**
Humanities and Arts							
Languages and Linguistics	LP	21,252	403,305	7,831	102,504	23,708	505,809
Philosophy and Religion		36,253	1,503,013	9,269	125,917	38,229	1,628,930
History	HE	49,700	2,096,004	11,835	138,139	51,514	2,234,143
Ethnology		39,759	1,083,009	3,040	19,666	40,150	1,102,675
Fine Arts	AH	35,501	967,809	5,042	39,744	36,177	1,007,553
Literature (G=Literary works=Imaginative texts)		--	--	5,592	36,852	5,592	36,852
Other Humanities and Arts		46,304	1,973,098	683	3,414	46,337	1,976,512
The Whole of Humanities and Arts		**88,953**	**8,026,238**	**23,787**	**466,236**	**92,810**	**8,492,474**
Social Sciences							
Politics	PL	26,299	920,841	8,814	115,166	27,900	1,036,007
Law		16,502	511,059	10,074	333,946	19,542	845,005
Economics	EC	20,015	684,404	12,534	367,555	23,525	1,051,959
Commerce and Business		22,087	846,432	10,788	310,716	24,489	1,157,148
Sociology and Social Issues	SE	30,362	1,318,930	12,960	333,772	33,008	1,652,702
Education		20,157	621,050	10,417	262,063	22,675	883,113
Other Social Matters		18,993	424,164	4,114	36,168	19,652	460,332
The Whole of Social Sciences		**54,613**	**5,326,880**	**29,386**	**1,759,386**	**60,762**	**7,086,266**
Technological Natural Sciences							
Mathematics	ST	3,497	40,397	1,959	19,472	4,352	59,869
Physics		2,368	25,239	1,280	9,430	2,920	34,669
Astronomy, Earth and Planetary Science		8,181	101,565	2,583	21,765	9,035	123,330
Chemistry, Metal and Mine		4,682	37,469	2,553	23,275	6,017	60,744
Technology (Architecture, Civil Engineering)		16,242	307,617	7,662	114,099	18,443	421,716
Technology (Mechanics, Electricity, Marine Engineering)		12,993	195,762	5,495	72,049	14,820	267,811
Other Technological Natural Sciences		18,530	399,470	8,426	145,175	21,018	544,645
The Whole of Technological Natural Sciences		**32,125**	**1,107,519**	**15,864**	**405,265**	**36,309**	**1,512,784**
Biological Natural Science							
Biology	BM	14,680	262,283	4,064	41,071	15,672	303,354
Agriculture		14,932	238,989	3,376	28,584	15,860	267,573
Pharmacy		3,610	24,703	1,103	10,197	4,017	34,900
Medicine		16,657	485,896	5,955	82,800	17,961	568,696
Dentistry		1,740	11,551	874	3,814	2,174	15,365
Nursing		2,348	19,255	2,491	23,505	3,744	42,760
Other Biological Natural Sciences		28,254	943,822	6,749	74,567	29,490	1,018,389
The Whole of Biological Natural Science		**40,160**	**1,986,499**	**13,117**	**264,538**	**42,674**	**2,251,037**
Internet Q & A Forum (Yahoo Chiebukuro)	IF	**54,215**	**5,224,852**	--	--	**54,215**	**5,224,852**
The Whole of VDRJ		**135,794**	**29,923,987**	**46,996**	**2,895,425**	**144,231**	**32,819,412**

Note 1: Published books and library books are added together.
Note 2: The figures contain number of signs. Unidic and MeCab were used for word segmentation. No additional processing was made for extracting noises.
Note 3: If the C-code of a text is 3,000-3,999, it is counted as a technical text.

The examples of the book titles for the technical texts in linguistics and language studies are shown below. (The titles in the brackets are the translation by the author of this thesis.)

- 続昭和(→平成)日本語方言の総合的研究 (A Comprehensive Study of Japanese Dialects, Second Series: from Showa era to Heisei era)
- 国際コミュニケーションと国際関係 (International Communication and International Relationships)
- 日英対照動詞の意味と構文(A Contrastive Study of Verbs between Japanese and English: Meanings and Structure)
- 英語から日本が見える (Viewing Japan through the English Language)
- 国語文字史の研究 (A Study of the History of Japanese Characters)
- 「た」の言語学 (A Linguistic Investigation into *–ta*)
- ことばの歴史 (History of Language)
- 京阪系アクセント辞典 (A Dictionary of the Kyoto and Osaka Accents)
- 日本語モダリティの史的研究 (A Historical Study of the Modality of the Japanese Language)

To extract domain-specific words by a statistical index, two types of corpora are required: a target corpus (i.e. the corpus from which the domain-specific words are extracted) and a reference corpus (i.e. general corpus). For this study, four target corpora are used. Each of the four target corpora is a group of technical texts from one of the four large academic fields: 1) Humanities and Arts, 2) Social Sciences, 3) Technological Natural Sciences and 4) Biological Natural Sciences (Table 7-2). (See also footnote 35 in 3.3.2 for the difference on the sub-divisions of the corpus between Coxhead's study and this study.) The reference corpus, which is all general (non-expert) book texts and all the internet-forum

texts in VDRJ, contain around 30 million tokens. For extracting the academic vocabulary from humanities and arts, all technical texts from humanities and arts are compared with the reference corpus for calculating the log-likelihood ratio. The same procedures are repeated for the four large academic fields using the 'Keyness' function of the software AntConc version 3.2.1 (Anthony, 2007).

After adding the log-likelihood figures to the database (VDRJ), academic vocabulary is extracted using the filtering function. The cut-off points are set at a higher level for more narrowly distributed words. This decision may look arbitrary; however, the fewer the criteria, the higher the cut-off point should be. Otherwise, the extracted vocabulary will include more inappropriate words. Specifically, the cut-off points are set as shown below. (LLR: log-likelihood ratio)

-Common academic words (AWs): high specificity in 3+ academic domains

- 4-domain words (cut-off point: LLR > 0[85])
- 3-domain words (cut-off point: LLR > 0)

-Limited-academic-domain words (LADs) : high specificity only in 1 or 2 academic domains

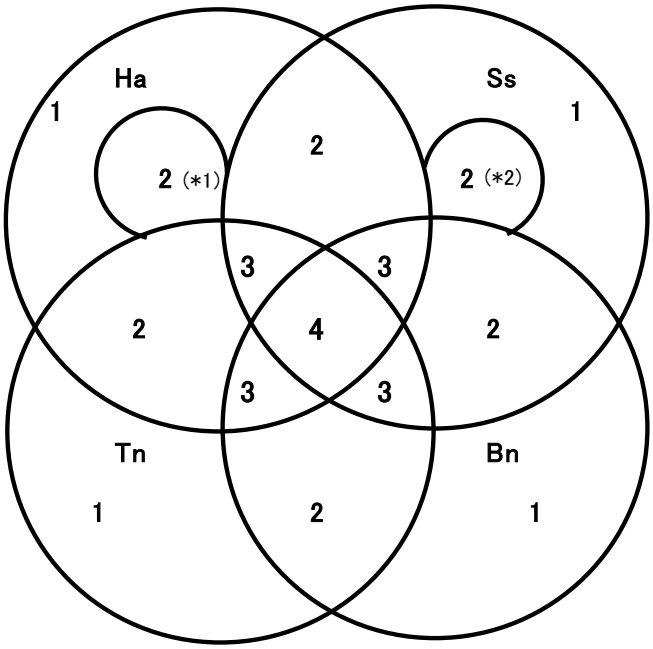
- 2-domain words (cut-off point: LLR > 1)
- 1-domain words (cut-off point: LLR > average value at a domain)

4 domain-words are extracted first. 3-domain words are extracted from the remainder. The same approach is applied to extracting 2-domain and 1-domain words. For example, if the log-likelihood ratio figures of a word for the four large academic domains are 4.8, 0.9, -2.5 and 0.4, the word is a 3-domain word. If the figures are 89.6 (LLR > average), 0.8, -1.8 and -14.8, the word is a 1-domain word.

[85] In the column 'Specificity Level' in VDRJ, '1' means LLR>0, '2' means LLR>1, and '3' means LLR > average value at a domain.

When checking overlapping words extracted from the four academic domains, different combinations of the domains where the words occur are found (Figure 7-1). There is only one combination for the 4-domain words; however, there are four combinations for the 3-domain words, six combinations for the 2-domain words and four types of the 1-domain words. In total, fifteen groups are identified for academic vocabulary at the four different combination levels.

Figure 7-1 Number of Shared Academic Domains among the Four Academic Domains



Ha: Humanities & Arts, **Ss**: Social Sciences,
Tn: Technological Natural Sciences, **Bn**: Biological Natural Sciences
* 1: The overlapping domains are Ha and Bn.
* 2: The overlapping domains are Ss and Tn.

After extracting the words from the four domains, the former JLPT Level 4 vocabulary (the words ranked at 681 or higher in WIS, the Word Ranking for International Students in VDRJ) were eliminated because the words such as 左 'hidari' (left (side)) or 百 'hyaku' (hundred) are too basic even if they are statistically specific to some domain(s). The words ranked at 20,001 or lower were also eliminated as their frequencies are too low.

All the remaining words are classified into different frequency levels from basic to

super-advanced by the Word Rankings for International Students (WIS) as shown in 7.1.1.4.

7.2.3 Common academic words (AWs) listed in the Japanese Common Academic Word List (JAWL)

I will first show and discuss the distribution and examples of common academic words, followed by the semantic features, parts of speech, word-origins and Kanji of the common academic words. After describing and discussing different groups of domain-specific words i.e. the limited-academic-domain words (LADs) and literary words (LWs), I will examine the text coverage by the common academic words as a proof of the usefulness of the Japanese Common Academic Word List (JAWL) in 7.4, along with an analysis of texts in different genres by different groups of words.

7.2.3.1 Distribution and examples of Japanese common academic words

The 4-domain words and 3-domain words are included in the Japanese Common Academic Word List (JAWL) version 1, which is available from the accompanying CD or http://www.geocities.jp/tatsum2003/. Not only the word lists, but the database version of JAWL, which contains types of information including word rankings, level of domain-specificity, reading of the word, sub frequencies, is also available there. In VDRJ (the Vocabulary Database for Reading Japanese), words which are labelled 'Aca4D' or 'Aca3D' in the 'Word Tier Label' column are the common academic words.

JAWL includes 2,591 words which are labelled from Level 0 to Level VIII for the user's convenience (Table 7-3); however, the Level 0 (70 basic 4-domain and 3-domain words) list and Level Ⅰ list (559 intermediate 4-domain words) are the most important lists. At the basic level, there are only 70 words. It is not surprising as most basic words are not specific to academic texts but commonly used in various types of texts. However, once a learner enters into the intermediate level, a large proportion of common academic words

must be learned if s/he learns Japanese for academic purposes. 1,101 (559+542) words are listed as common academic words at the intermediate level in the Word Rankings for International Students (WIS). These words account for 29.8% of the 3,688 intermediate words (ranked at 1,289-5,000). The numbers for both four-domain words and three-domain words are highest at the intermediate level and the number decreases as the level goes lower.

The number of Japanese common academic words may seem to be too high as low-frequency common academic words are included in the list. However, it is sure that such many common academic words exist as they are common words extracted from the four large academic domains at different frequency levels. As is discussed later in 7.4.2.1, these words are still worth being included in the list.

Table 7-4 shows the number and proportion of Japanese common academic words by JAWL level and the former Japanese Language Proficiency Test (F-JLPT) word level. 55.1% of JAWL Ⅰ words are listed at F-JLPT Level 2, and 30.6% are at Level 1. More than 80% of JAWL Ⅰ and II (intermediate) words are listed in F-JLPT Level 2 (Intermediate) or Level 1 (Advanced). This also suggests that the intermediate academic vocabulary is very important for learning academic Japanese.

80 words (14.3%) of JAWL Level 1 (intermediate 4-domain words) are not listed in F-JLPT word lists, and 101 words (18.6%) of JAWL Level 2 (intermediate 3-domain words) are not listed in F-JLPT word lists, either. These words include 挙げる 'ageru' (mention, cite), 捉える 'toraeru' (capture, grasp, see), 時点 'jiten' (a point of time, moment), 多数 'tasuu' (a large number), 層 'sou' (layer, stratum), 初期 'shoki' (the early days, the beginning), 両者 'ryousha' (the two), 次元 'jigen' (dimension), 反論 'hanron' (counterargument), 組み合わせ 'kumi-awase' (combination), 示唆 'shisa' (suggestion), and 仮説 'kasetsu' (hypothesis). These words seem to be essential for academic language; thus, F-JLPT lists seem to be inappropriate at least for academic purposes, though it is still used for university admission purposes at some universities.

Table 7-3 Distribution and Examples of Japanese Common Academic Words listed in JAWL Ver.1

JAWL Label	The Former JLPT Level	Word Rankings for International Students (WIS rankings)	Level	Number of High-Specificity Domains among the 4 Large Academic Domains	Number of Lexemes	Least Frequent 6 Words in Each Domain	Translation of the Least Frequent 6 Words in Each Domain
JAWL 0	L3	682-1,291	Basic	4	31	科学　規則　割合 生産　産業　講義	science, rule, proportion, production, industry, lecture
				3	39	人口　スクリーン　数学 競争　工業　地理	population, screen, mathmatics, competition, manufacture, geography
JAWL I		1,292-5,000	Inter.	4	559	発足　半数　配分 縮小　適正　見直し	inauguration, half the number, allocation, downsize, proper, reconsider
JAWL II				3	542	演説　大小　実情 ステージ　ライフ　担保	speech, size, real situation, stage, life, guarantee
JAWL III	L2 L1	5,001-10,000	Adv. 1	4	212	難問　能動　付随 定型　除　本稿	difficult problem, active, accompany, standard, except, this article
JAWL IV				3	452	交錯　カウント　精度 一因　箇年　エンド	mixture, count, accuracy, one cause, -year, end
JAWL V	Other	10,001-15,000	Adv. 2	4	103	併存　親和　盛況 散在　補填　関わり合う	coexistence, affinity, prosperity, straggle, compensation, implicated
JAWL VI				3	328	帰着　編著　沿海 拮抗　常套　内情	come down to, written and edited, coastal, close competition, conventional, internal condition
JAWL VII		15,000-20,000	Super-adv.	4	56	閉　増刊　含意 複　活路　所与	closed, extra edition, implication, double-, way out, given
JAWL VIII				3	269	極小　付則　深度 概算　頒布　円錐	minimal, additional clause, depth, rough estimate, distribution (of goods/paper), cone

* JAWL: Japanese Common Academic Word List

Table 7-4 Number and Proportion of Japanese Common Academic Words by JAWL Level and the F-JLPT Word Level

JAWL Label	Word Rankings for International Students (WIS rankings)	Level	Number of High-Specificity Domains Out of the 4 Large Academic Domains	F-JLPT Word Level (Counted by Lexeme)					F-JLPT Word Level (Counted by Lexeme, %)				
				Level 3	Level 2	Level 1	Others	Total	Level 3	Level 2	Level 1	Others	Total
JAWL 0	682-1,291	Basic	4	31	--	--	--	31	100.0	--	--	--	100.0
			3	39	--	--	--	39	100.0	--	--	--	100.0
JAWL I	1,292-5,000	Inter.	4	--	**308**	**171**	**80**	**559**	--	**55.1**	**30.6**	**14.3**	100.0
JAWL II			3	--	**268**	**173**	**101**	**542**	--	**49.4**	**31.9**	**18.6**	100.0
JAWL III	5,001-10,000	Adv. 1	4	--	28	46	138	212	--	13.2	21.7	65.1	100.0
JAWL IV			3	--	39	118	295	452	--	8.6	26.1	65.3	100.0
JAWL V	10,001-15,000	Adv. 2	4	--	2	5	96	103	--	1.9	4.9	93.2	100.0
JAWL VI			3	--	5	28	295	328	--	1.5	8.5	89.9	100.0
JAWL VII	15,000-20,000	Super-adv.	4	--	2	3	51	56	--	3.6	5.4	91.1	100.0
JAWL VIII			3	--	8	10	251	269	--	3.0	3.7	93.3	100.0
JAWL 0-VIII	682-20,000	All	4 or 3	70	660	554	1307	2591	2.7	25.5	21.4	50.4	100.0

* JAWL: Japanese Common Academic Word List

* F-JLPT: the former Japanese Language Proficiency Test

7.2.3.2 Semantic features of Japanese common academic words

Common academic words are highly abstract, and essential for managing academic information. Below are some examples.

- Range: 占める ‘shimeru’ (occupy, account for),

  特殊 ‘tokushu’ (special, particular)
- Relation: 属する ‘zokusuru’ (belong to), 依存 ‘izon’ (reliance/rely)
- Comparison/Evaluation: 後者 ‘kousha’ (the latter),

  優れる ‘sugureru’ (superior)
- Quantitative change: 減少 ‘genshou’ (decrease),

  強化 ‘kyouka’ (reinforcement)
- Stage: 当初 ‘tousho’ (beginning), 現状 ‘genjou’ (present condition)
- Development of enunciation: 取り上げる ‘toriageru’ (take up [an issue]),

  まとめる ‘matomeru’ (summarize)

Besides these notions and functions, social or scientific aspects of academic information such as cause-effect, degree, agent, action, object, direction, goal, instrument, time are managed by common academic words[86].

Some of the 3-domain words have concrete meanings e.g. 署名 'shomei' (signature) and 保健 'hoken' (health, hygiene). Nevertheless, few 4-domain words have concrete meanings. This nature of the words seems to be the same at all levels.

7.2.3.3 Part of speech of Japanese common academic words

Among the 2,591 common academic words, 1,072 words (41.4 %) are common nouns such as 背景 'haikei' (background). There are 882 (34.0 %) verbal nouns such as 連続 'renzoku(-suru)' (establish/-ment). Adding other types of nouns together, 2,104 words (81.2 %) can be nouns. Excluding verbal nouns, there are 225 verbs (8.7 %) such as 認める 'mitomeru' (recognize/approve) and 述べる 'noberu' (describe/mention). Including verbal nouns, 1,107 words (42.7%) can be verbs.

There are only 95 (3.7 %) nominal adjectives (e.g. 詳細 'shousai' (detail/-ed), 平等 'byoudou' (equal/-ity)) and 9 (0.3 %) adjectives (e.g. 著しい 'ichijirushii' (remarkable).

There are 106 (4.1 %) affixes (e.g. -期 '-ki' (period), -種 'shu' (type)). As discussed in 4.3 and 4.4, Chinese-origin affixes are frequent in Japanese academic expressions.

There are only 34 (1.3 %) adverbs (e.g. しばしば 'shibashiba' (frequently) and 22 (0.8 %) other parts of speech (e.g. particle, auxiliary verb). In this category, there are remarkably many archaic words. Examples are のみ 'nomi' (only), つつ '-tsutsu' (while doing), べし '-beshi' (ought to), あらゆる 'arayuru' (every), いかなる 'ikanaru' (any), 我が 'waga' (my), 漠然 'bakuzen' (vague). The auxiliary verb れる/られる '-reru/rareru'

[86] Hirsh (2004) classified the functions of English academic words (Coxhead, 2000) in academic texts into three large categories of Textual, Ideational and Interpersonal. And then, the Textual is classified into subcategories of metatextual, extratextual and intratextual, Ideational is classified into scholarly process, states of affairs and relations between entities, and Interpersonal is labelled as authoritative but not classified. This is a classification of the functions of academic words but not the classification of words itself.

(used for passives/potentials/spontaneous/honorifics) is also extracted as a common academic word. This is probably because passive sentences are more frequently used in academic texts than in general texts.

Comparing these proportions to general proportions as shown in Table 4-17 in Chapter 4, common academic words have more verbal nouns (common academic words: the VDRJ vocabulary from 1K to 20 K = 34.0:18.2) and affixes (4.1% for common academic words vs. 0.5% and 2.0% for prefixes and suffixes respectively) but fewer verbs (8.7:13.7), adjectives (0.3:1.6) and adverbs (1.3:2.9) (The ratios are based on lexeme counts). This result is in line with the result of 4.4 where I claim that the proportion(s) for the total tokens of suffixes, verbal nouns can be the index for formality and the proportion(s) for the total tokens of adverbs, verbs, adjectives can be the index for informality.

7.2.3.4 Word origins of Japanese common academic words

As shown in Table 7-5, Chinese-origin words, which are mostly written in Kanji, account for around three quarters of the words at any levels. 'Other-origins', which are mostly English-origin words, account for 7-11% (counted by lexemes) at the advanced level or above; however, they only account for 2.1% at JAWL Ⅰ which is the most important level of all. Japanese-origin words account for more than 20% at JAWL 0 and Ⅰ but 9-16% at the other levels, which are considerably lower than the proportion in the total Japanese lexemes at around one third or more (Matsushita, 2009, 2010).

These facts tell us that the first language effect, especially understanding Kanji vocabulary, will possibly make a gap in burden of learning Japanese academic texts depending on the learner's language background. This issue will be further discussed in 7.5.

Table 7-5 Number and Proportion of Word Origins of Japanese Common Academic Words by Frequency Level

JAWL Label	The Former JLPT Level	Word Rankings for International Students (WIS rankings)	Level	Number of High-Specificity Domains Out of the 4 Large	Total Number of Lexemes	Word-origin (Counted by Lexeme)						Word-origin (% at Each Level)					
						Japanese-origin	Chinese-origin	Other-origins	Mixed-origin	Proper Nouns	Signs, Unknown and Others	Japanese-origin	Chinese-origin	Other-origins	Mixed-origin	Proper Nouns	Signs, Unknown and Others
JAWL 0	L3	682-1,291	Basic	4	31	8	21	0	1	0	1	25.8	67.7	0.0	3.2	0.0	3.2
				3	39	8	28	3	0	0	0	20.5	71.8	7.7	0.0	0.0	0.0
JAWL I	L2	1,292-5,000	Inter.	4	**559**	**115**	**417**	12	14	0	1	**20.6**	**74.6**	**2.1**	2.5	0.0	0.2
JAWL II				3	542	77	416	35	7	6	1	14.2	76.8	6.5	1.3	1.1	0.2
JAWL III		5,001-10,000	Adv. 1	4	212	27	163	16	6	0	0	12.7	76.9	7.5	2.8	0.0	0.0
JAWL IV				3	452	56	343	41	7	4	1	12.4	75.9	9.1	1.5	0.9	0.2
JAWL V	Other	10,001-15,000	Adv. 2	4	103	9	85	8	1	0	0	8.7	82.5	7.8	1.0	0.0	0.0
JAWL VI				3	328	43	246	31	5	1	2	13.1	75.0	9.5	1.5	0.3	0.6
JAWL VII		15,000-20,000	Super-adv.	4	56	9	37	6	2	0	2	16.1	66.1	10.7	3.6	0.0	3.6
JAWL VIII				3	269	38	192	30	5	0	4	14.1	71.4	11.2	1.9	0.0	1.5
JAWL 0-VIII	All	682-20,000	All	4 or 3	2591	390	1948	182	48	11	12	**15.1**	**75.2**	**7.0**	1.9	0.4	0.5

* JAWL: Japanese Common Academic Word List

7.2.3.5 Kanji used for Japanese common academic words

Even if we limit the Kanji use to the common Japanese Kanji (常用漢字 'jouyou Kanji')[87], 70% of the characters used for the most representative orthographic forms of common academic words are Kanji. As shown in Table 7-6, at the basic and intermediate levels, three quarters are Kanji; however, after the intermediate level, the proportion decreases to 59.3% at JAWL VIII (super-advanced level, ranked from 15,001 to 20,000). At JAWL 0 (basic) and JAWL II (intermediate), most Kanji appear for the first time if learning common academic words from the basic level; however, at JAWL II or above, more than half the Kanji are repeatedly used ones. In other words, Kanji which are new to learners are fewer than half. Many Kanji are repeatedly used. This can also be understood by comparing the proportion of first appearing Kanji and common academic words at each level. At JAWL 0 and Ⅰ, the proportions of Kanji are higher at 4%, 5% and 36% than the proportions of common academic words at 1%, 2% and 22% (Table 7-6). This shows that the Kanji at the basic and intermediate levels are repeatedly used. Learning JAWL 0 and Ⅰ Kanji should be very important.

(From here down blank.)

[87] 2,136 Kanji are currently listed in the revised list of common Japanese Kanji (改定常用漢字表 'kaitei-jouyou-Kanji-hyou') (Agency for Cultural Affairs, 2010).

Table 7-6 Number and Proportion of Kanji which are New to Learners in Common Academic Words *Learning common academic words in order of the level is assumed.

JAWL Label	The Former JLPT Level	Level	Number of High-Specificity Domains Out of the 4 Large Academic Domains	Number of Kanji Types	Number of New Kanji (*)	Proportion of New Kanji at the Level	Proportion of New Kanji among All the Kanji Used in Common Academic Words	Cumulative Number of New Kanji	Cumulative Proportion of New Kanji (*)	Number of Academic Words at the Level	Proportion of Common Academic Words among All the Common Academic Words
JAWL 0	L3	Basic	4	42	42	100%	**4%**	42	4%	31	**1%**
			3	56	51	91%	**5%**	93	9%	39	**2%**
JAWL I	L2 L1 Other	Inter.	4	439	378	86%	**36%**	471	45%	559	**22%**
JAWL II			3	472	202	43%	19%	673	64%	542	21%
JAWL III		Adv. 1	4	263	51	19%	5%	724	69%	212	8%
JAWL IV			3	478	150	31%	14%	874	83%	452	17%
JAWL V		Adv. 2	4	146	21	14%	2%	895	85%	103	4%
JAWL VI			3	386	85	22%	8%	980	93%	328	13%
JAWL VII		Super-adv.	4	86	14	16%	1%	994	94%	56	2%
JAWL VIII			3	312	62	20%	6%	1056	100%	269	10%
JAWL 0-VIII	All	All	4/3	**1056**	**1056**		**100%**			**2591**	**100%**

The eleven most frequently used Kanji for common academic words are 合 (combine, together), 定 (fix, certain), 分 (divide, minute), 一 (one), 同 (same), 数 (number), 上 (up), 体 (body), 出 (out), 大 (large), 実 (real, actual). These Kanji show, as discussed in 7.2.3.2, the abstract features of common academic words which are used for managing academic information. Each of these Kanji appears in 38 to 23 common academic words. Other frequent Kanji are 用 (use), 要 (need), 明 (bright, clear), 度 (degree), 発 (start, emerge), 論 (theory, logic), 入 (enter), 有 (exist), 行 (act, behave), 成 (become), 学 (study), 生 (live, raw), 理 (reason, theory), 前 (front, before), 動 (move), 法 (law), 点 (point), 面 (face, surface), 付 (attach), 当 (hit, equivalent), 特 (special), 中 (middle, inside), 変 (change), 質 (quality), 自 (self), 部 (part, section), 進 (proceed). Each of these appears in 22 to 15 common academic words.

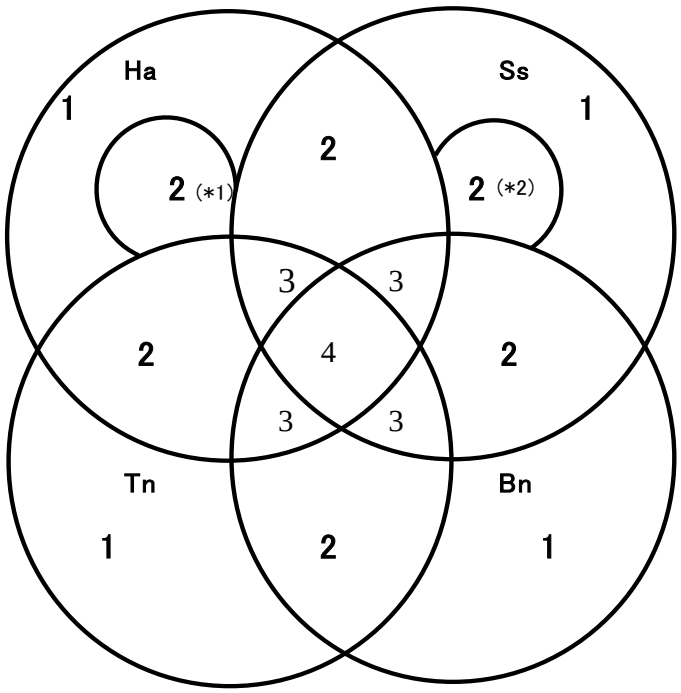
Similar to the discussion in Chapter 6, some Kanji at JAWL 0 and Ⅰ appear in only one common academic word and are not used in other common academic words. There are five such Kanji at JAWL 0 which are 十 (for 十分 'juubun' (sufficient)), 研 (for 研究 'kenkyu' (research)), 紹 (for 紹介 'shoukai' (introduction)), 糸 (for 糸 'ito' (thread)) and

険 (for 危険 ‘kiken’ (danger)). 十 and 糸 are fairly common in non-academic Japanese; however, 紹 seems rarely used for other words than 紹介. There are 46 such Kanji at JAWL Ⅰ as well. Examples are 互 (for 相互 ‘sougo’ (mutual)), 刺 (for 刺激 ‘shigeki’ (stimulus)), 唆 (for 示唆 ‘shisa’ (suggestion)), 徴 (for 特徴 ‘tokuchou’ (feature)), 摘 (for 指摘 ‘shiteki’ (point out)). These Kanji are not often used for other words but still need to be learned at the level as the words composed of the Kanji are essential for academic texts.

7.2.4 Limited-academic-domain Words (LADs)

Limited-academic-domain words (LADs) are the words which are specific to 1 or 2 academic domains out of the four domains of 1) Humanities and Arts, 2) Social Sciences, 3) Technological Natural Sciences and 4) Biological Natural Sciences (Figure 7-2) (For more detail, see 7.2.1 and 7.2.2). As mentioned in 7.2.2, for the 2-domain words, the cut-off point is set at more than 1.0 of the log-likelihood ratio, and for the 1-domain words, the cut-off point is set at more than the average value in the target domain. Actually, LADs were not intended to be extracted first but were a kind of ‘by-product’ produced through the process of extracting the common academic words, namely the 4-domain and 3-doman words. Looking at those 2-domain and 1-domain words, they seem not to be the unimportant left-overs of common academic words but seem to be useful groups of words. LADs are something between ‘academic’ and ‘technical’. They are expected to provide higher text coverage in some academic fields than non-academic vocabulary. As discussed in 7.1, there are similar ideas in English vocabulary studies (Coxhead & Hirsh, 2007; Tajino et al., 2009, 2007); however, there seems no similar attempt in Japanese. In the university curriculum, these words should be learned before the learners select their major. The lists of LADs are available from the accompanying CD. Also, these words are easily identified in the columns for ‘Specificity Level’ (see 7.2.2) and ‘Word Tier Label’ in VDRJ.

Figure 7-2 Number of Shared Academic Domains among the Four Academic Domains with the Domains for Limited-academic-domain words Highlighted in Bold Type *1 and 2 in bold type show the domains for limited-academic-domain words.



Ha: Humanities & Arts, **Ss**: Social Sciences,
Tn: Technological Natural Sciences, **Bn**: Biological Natural Sciences
* 1: The overlapping domains are Ha and Bn.
* 2: The overlapping domains are Ss and Tn.

7.2.4.1 Distribution, examples and semantic features of Japanese limited-academic-domain words

I will show the distribution, examples and semantic features of 2-domain words first, followed by the 1-domain words.

2-domain words

The distribution of 2-domain words by frequency level and shared domains is shown in Table 7-7.

Table 7-7 Number of 2-domain Words of Japanese Limited-academic-domain Words (LAD) by Frequency Level and Shared Domains

LAD Label (*)	F-JLPT Level (**)	Word Rankings for International Students	Level	Number of Lexemes in LAD of Ha & Ss	Number of Lexemes in LAD of Ha & Tn	Number of Lexemes in LAD of Ha & Bn	Number of Lexemes in LAD of Ss & Tn	Number of Lexemes in LAD of Ss & Bn	Number of Lexemes in LAD of Tn & Bn	Total
LAD 0	L3	682-1,291	Basic	15	5	4	5	6	10	45
LAD I		**1,292-5,000**	**Inter.**	139	27	30	77	57	61	**391**
LAD III	L2	**5,001-10,000**	**Adv. 1**	138	38	25	86	50	92	**429**
LAD V	L1 Other	10,001-15,000	Adv. 2	91	28	22	58	37	60	296
LAD VII		15,000-20,000	Super-adv.	93	23	17	43	16	40	232
Total				**476**	121	98	**269**	166	**263**	1,393

Ha: Humanities & Arts, **Ss**: Social Sciences, **Tn**: Technological Natural Sciences, **Bn**: Biological Natural Sciences

*LAD II, IV, VI and VIII are the labels for 1-domain words.

**F-JLPT: The former Japanese Language Proficiency Test

'Humanities and arts' (Ha) and social sciences (Ss) share 476 domain-specific words which is the largest group among the six combinations. Technological natural sciences (Tn) and biological natural sciences (Bn) share 263 specific words which are also a large group. Interestingly, social sciences and technological natural sciences also share 269 specific words. In contrast to that, 'humanities and arts' and biological natural sciences share only 98 words.

From the viewpoint of the level, intermediate to advanced are the most important levels at which to learn the 2-domain words. In any combination of the two domains, intermediate (Inter.) and advanced 1 (Adv. 1) levels offer the largest number of 2-domain words. For common academic words (3-domain and 4-domain words), their importance is more related to the intermediate level. Generally speaking, the more specific a word is, the lower the frequency of the word will be. 1-domain words and technical vocabulary are expected to be distributed at lower-frequency levels.

Examples of 2-domain words and their English translations are in Table 7-8 and 7-9.

Table 7-8 Examples of 2-domain Words of Japanese Limited-academic-domain Words (LAD) by Frequency Level and Shared Domains

LAD Label (*)	F-JLPT Level (**)	Word Rankings for International Students	Level	Least Frequent 2 Words in LAD of Ha & Ss	Least Frequent 2 Words in LAD of Ha & Tn	Least Frequent 2 Words in LAD of Ha & Bn	Least Frequent 2 Words in LAD of Ss & Tn	Least Frequent 2 Words in LAD of Ss & Bn	Least Frequent 2 Words in LAD of Tn & Bn
LAD 0	L3	682-1,291	Basic	貿易 輸出	砂 テキスト	発音 ステレオ	製 レポート	以内 パート	アルコール テニス
LAD I	L2 L1 Other	1,292-5,000	Inter.	孤立 融資	オール ペーパー	静岡 書簡	ニーズ 顧客	総務 性的	スイッチ 液
LAD III		5,001-10,000	Adv. 1	容れる 教義	音響 流布	発現 海域	本件 セクション	閉塞 弱める	多用 部位
LAD V		10,001-15,000	Adv. 2	払い戻し ユニバーシティ	落差 コロン	目付け 生長	VTR リハーサル	所見 救命	光学 ペーハー
LAD VII		15,000-20,000	Super-adv.	峻別 公債	目配り テクノ	太極 増量	パレット 軽微	マンガン 居宅	棒状 雨水

Ha: Humanities & Arts, **Ss**: Social Sciences, **Tn**: Technological Natural Sciences, **Bn**: Biological Natural Sciences

*LAD II, IV, VI and VIII are the labels for 1-domain words.

**F-JLPT: The former Japanese Language Proficiency Test

Table 7-9 Examples of 2-domain Words of Japanese Limited-academic-domain Words (Translation) by Frequency Level and Shared Domains

LAD Label (*)	F-JLPT Level (**)	Word Rankings for International Students	Level	Translation of the Least Frequent 2 Words in LAD of Ha & Ss	Translation of the Least Frequent 2 Words in LAD of Ha & Tn	Translation of the Least Frequent 2 Words in LAD of Ha & Bn	Translation of the Least Frequent 2 Words in LAD of Ss & Tn	Translation of the Least Frequent 2 Words in LAD of Ss & Bn	Translation of the Least Frequent 2 Words in LAD of Tn & Bn
LAD 0	L3	682-1,291	Basic	trade export	sand text	pronunciation stereo	made (in) report	within part(-timer)	alcohol tennis
LAD I	L2 L1 Other	1,292-5,000	Inter.	isolation loan	all paper	Shizuoka pref./city	need (n.) customer	general affairs sexual	switch liquid
LAD III		5,001-10,000	Adv. 1	compatible doctrine	accoustic circulation	manifestation waters	this matter section	impasse weaken	frequent use region (of body)
LAD V		10,001-15,000	Adv. 2	refund university	a drop cologne	overseer growth	VTR rehearsal	remark (n.) lifesaving	optics pH
LAD VII		15,000-20,000	Super-adv.	sharp distinction	meticulous care	tai ji increase in	pallet slight	manganese dwelling	stick-shaped rainwater

Ha: Humanities & Arts, **Ss**: Social Sciences, **Tn**: Technological Natural Sciences, **Bn**: Biological Natural Sciences

*LAD II, IV, VI and VIII are the labels for 1-domain words.

**F-JLPT: The former Japanese Language Proficiency Test

Overall, 2-domain words have much more concrete and specific meanings than common academic words. Then, what are the semantic features of each group of 2-domain words? By looking at the members of each group, some features are found for some combinations.

'Humanities and art' and social sciences tend to share many words on social studies,

especially on political history. The examples are 支配 ‘shihai’ (domination), 戦後 ‘sengo’ (after world war II), 封建 ‘houken’ (feudalism) and 中世 ‘chuusei’ (the middle ages). This is probably because history texts are classified as a part of ‘humanities and arts’.

Social sciences and technological natural science share many words in industry. The examples are コスト ‘kosuto’ (cost), 賠償 ‘baishou’ (compensation), マネージャー ‘mane^ja^’ (manager) and 家電 ‘kaden’ (home electrical appliances). Social sciences and biological natural sciences share many words on social security, medical and nursing service. The examples are 予防 ‘yobou’ (prevention), 麻痺 ‘mahi’ (anesthesia), 届け出 ‘todokede’ (notification, entry) and 母性 ‘bosei’ (maternity). These two combinations show social aspects of natural sciences.

Not surprisingly, technological and biological natural sciences share natural science vocabulary. The examples are エネルギー ‘enerugi^’ (energy), 細胞 ‘saibou’ (cell), 酸化 ‘sanka’ (oxidization) and イオン ‘ion’ (ion). Many of these words at basic and intermediate levels are essential for science students.

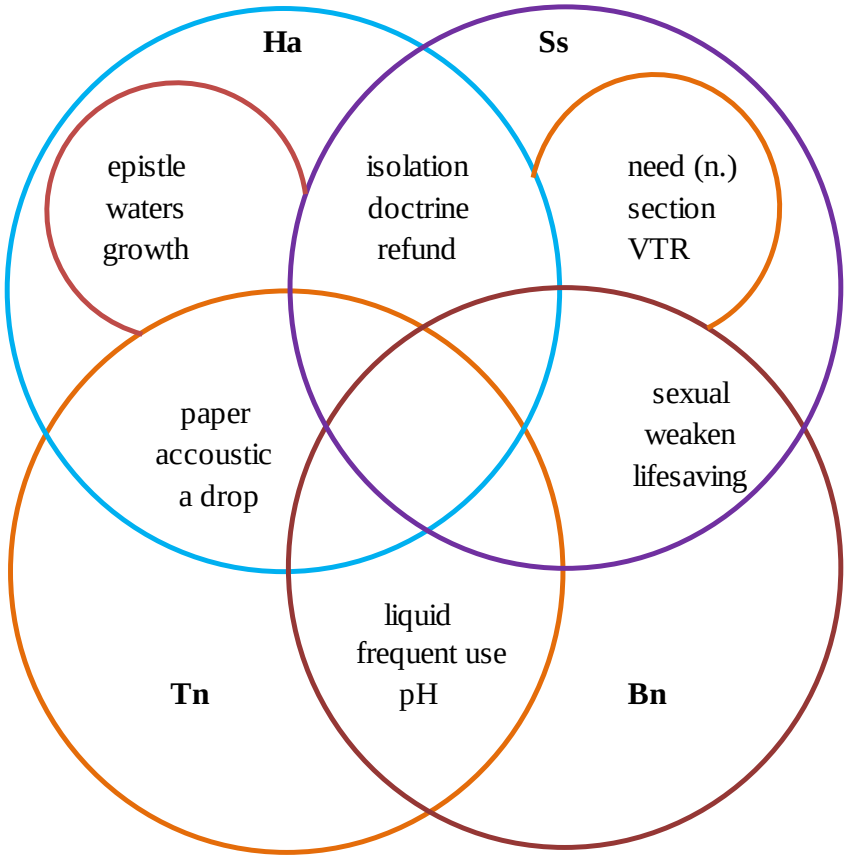
The features of the other two combinations, i.e. ‘humanities and arts’ and each of the two natural science domains, are not clear. Some shared words in ‘humanities and arts’ and technological natural sciences are on information science. The examples are タイプ ‘taipu’ (type), 文字 ‘moji’ (letter, character), 印刷 ‘insatsu’ (printing) and 蔵書 ‘zousho’ (the book stock, a collection of books). This may be because library science and information science are classified as a part of technological natural sciences. However, there are also many words which do not show distinctive features. It is not easy to find a common feature from the three words of 哲学 ‘tetsugaku’ (philosophy), 照明 ‘shoumei’ (lighting, illumination) and 木材 ‘mokuzai’ (timber).

Similarly, it is hard to detect common features shared by ‘humanities and arts’ and biological natural sciences. Examples of the scientific words in this category are 感覚 ‘kankaku’ (feeling, sensation), 脳 ‘nou’ (brain), 栽培 ‘saibai’ (cultivation) and 光線

‘kousen’ (ray, beam). Some of these words such as 光線 (ray, beam) may be used for metaphorical expressions in humanities texts. Some non-scientific words in this category are 儀式 ‘gishiki’ (ceremony), 細工 ‘saiku’ (workmanship, elaboration), 平坦 ‘heitan’ (flat, even) and 美的 ‘bi-teki’ (aesthetic). The non-distinctiveness of these categories will probably come from the nature of humanities and arts. As discussed in Chapter 4, humanities and arts are generally more lexically diverse than other academic fields and closer to daily-life words.

Example words in a Venn diagram are shown in Figure 7-3.

Figure 7-3 Examples of 2-domain Words (Translation) in a Venn Diagram



Ha: Humanities & Arts, **Ss**: Social Sciences,
Tn: Technological Natural Sciences, **Bn**: Biological Natural Sciences

1-domain words

Extracting 1-domain words is merely a trial. One is because the corpus size is not

large enough[88], especially for natural sciences, as the corpus is not dedicatedly designed for academic purposes, but a balanced corpus. Another reason is that extracting words from only one target corpus will require a more complete target corpus. Extracting something common across domains is much easier. Therefore, the precision of extraction seems lower than the common academic words (4-domain and 3-domain words).

The distribution of 1-domain words is shown in Table 7-10.

Table 7-10 Number of 1-domain Words of Japanese Limited-academic-domain Words (LAD) by Frequency Level and Domain

LAD Label (*)	F-JLPT Level (**)	Word Rankings for International Students	Level	Number of Lexemes in Ha	Number of Lexemes in Ss	Number of Lexemes in Tn	Number of Lexemes in Bn	Total
LAD 0	L3	682-1,291	Basic	13	6	5	9	33
LAD II	L2 L1 Other	**1,292-5,000**	**Inter.**	104	111	46	52	**313**
LAD IV		**5,001-10,000**	**Adv. 1**	104	127	60	68	**359**
LAD VI		10,001-15,000	Adv. 2	71	74	48	54	247
LAD VIII		15,000-20,000	Super-adv.	60	55	29	53	197
Total				352	373	188	236	1,149

Ha: Humanities and Arts, **Ss**: Social Sciences, **Tn**: Technological Natural Sciences, **Bn**: Biological
*LAD I, III, V and VII are the labels for 2-domain words.
**F-JLPT: The former Japanese Language Proficiency Test

The number of words is higher in arts and social sciences than in natural sciences. It is not sure if this is because of the corpus sizes or the nature of the domains. From the viewpoint of the levels, the highest number is in Adv.1 at 6K to 10K level, which is, as expected,

[88] As shown in Table 7-2, technological natural science texts have 1.51 million tokens and biological natural science texts have 2.25 tokens; however, the distribution of words is uneven in some sub-sections. For example, physics texts only have 0.03 million tokens, and pharmacy and dentistry only have 0.03 million and 0.02 million tokens respectively.

higher than common academic words and 2-domain words.

Examples of 1-domain words and their English translations are in Tables 7-11 and 7-12.

Table 7-11 Examples of 1-domain Words of Japanese Limited-academic-domain Words (LAD) by Frequency Level and Domain

LAD Label (*)	F-JLPT Level (**)	Word Rankings for International Students	Level	Least Frequent 2 Words in LAD of Ha	Least Frequent 2 Words in LAD of Ss	Least Frequent 2 Words in LAD of Tn	Least Frequent 2 Words in LAD of Bn
LAD 0	L3	682-1,291	Basic	辞典 文法	工場 遊び	海岸 汽車	退院 柔道
LAD II	L2 L1 Other	1,292-5,000	Inter.	色彩 滋賀	紛争 犯	原子 コンクリート	拳 杉
LAD IV		5,001-10,000	Adv. 1	王家 呪術	超過 欠席	硬化 ドラッグ	臓器 左足
LAD VI		10,001-15,000	Adv. 2	報国 遍歴	持ち分 受諾	PM 蒸留	卵子 緑茶
LAD VIII		15,000-20,000	Super-adv.	厳寒 鼎	卸売り 引き当て	プログラミング バラック	居合 微小

Ha: Humanities & Arts, **Ss**: Social Sciences, **Tn**: Technological Natural Sciences, **Bn**: Biological Natural Sciences

*LAD II, IV, VI and VIII are the labels for 1-domain words.

**F-JLPT: The former Japanese Language Proficiency Test

Table 7-12 Examples of 1-domain Words of Japanese Limited-academic-domain Words (Translation) by Frequency Level and Domain

LAD Label (*)	F-JLPT Level (**)	Word Rankings for International Students	Level	Translation of the Least Frequent 2 Words in LAD of Ah	Translation of the Least Frequent 2 Words in LAD of Ss	Translation of the Least Frequent 2 Words in LAD of Tn	Translation of the Least Frequent 2 Words in LAD of Bn
LAD 0	L3	682-1,291	Basic	dictionary grammar	factory play(ing)	seashore train	leave hospital judo
LAD II	L2 L1 Other	1,292-5,000	Inter.	coloring Shiga (pref.)	conflict offense	atom concrete (n.)	fist/martial art cedar
LAD IV		5,001-10,000	Adv. 1	royal family incantation	excess absence	harden(ing) drag/drug	organ left leg/foot
LAD VI		10,001-15,000	Adv. 2	patriotic itinerancy	quota acceptance	PM distillation	ovum green tea
LAD VIII		15,000-20,000	Super-adv.	intense cold three-legged vessel	wholesale researve fund	programming shanty	*iai* (martial arts) micro

Ha: Humanities & Arts, **Ss**: Social Sciences, **Tn**: Technological Natural Sciences, **Bn**: Biological Natural Sciences

*LAD II, IV, VI and VIII are the labels for 1-domain words.

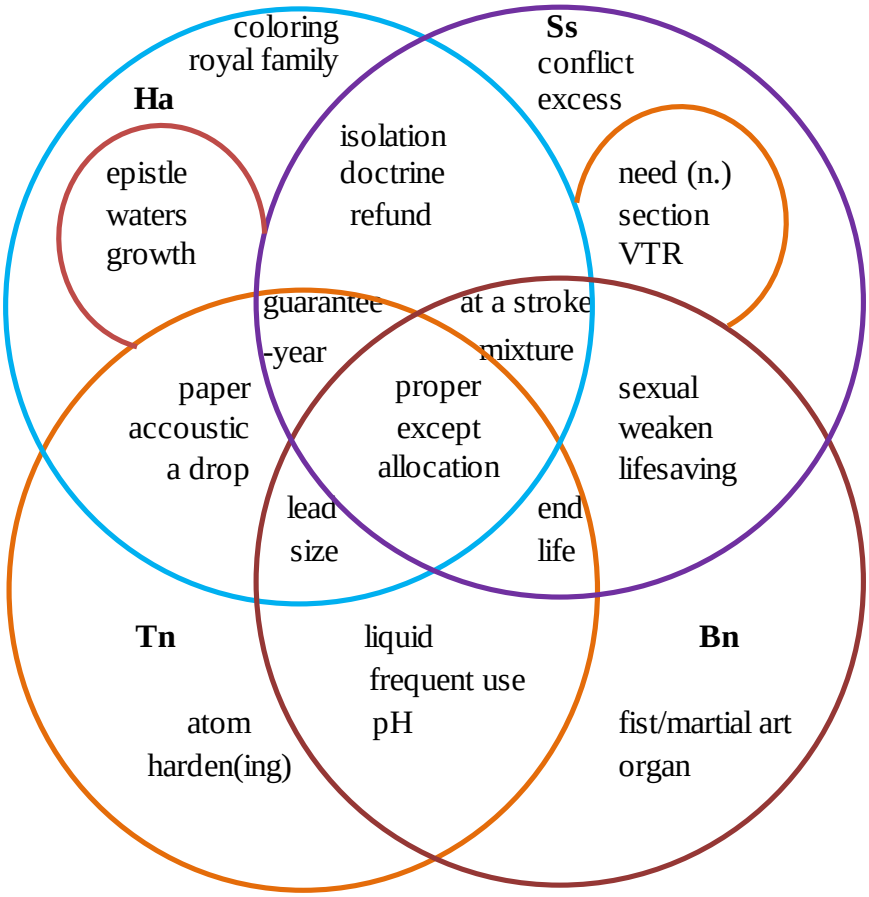
**F-JLPT: The former Japanese Language Proficiency Test

The semantic features of 1-domain words are much clearer than the 2-domain words, let alone 3-domain and 4-domain words. Typical examples are 神社 'jinja' (shrine), 人間 'ningen' (human being), 国語 'kokugo' (national language) for humanities and arts, 労働 'roudou' (labour), 予算 'yosan' (budget), 国籍 'kokuseki' (nationality) for social sciences, 材料 'zairyou' (material), インストール 'insuto^ru'(install), 原子 'genshi' (atom) for technological natural sciences, and 医学 'igaku' (medical science), 栄養 'eiyou' (nutrition), 熱帯 'nettai' (the tropics) for biological natural sciences.

There are some words which should not be extracted as 1-domain words from the domain. Examples are 同じく 'onajiku' (likewise), 年寄り (aged person) for humanities and arts, 園 '-en/sono' (garden), ホワイト 'howaito' (white) for social sciences, 呼び出す 'yobidasu' (summon, call), 禅 'zen' (Zen) for technological natural sciences, and 鏡 'kagami' (mirror), 県立 'kenritsu' (prefectural) for biological natural sciences. There are not many of these exceptions. They should be eliminated in some way when creating word lists for 1-domain words.

Examples of academic vocabulary including all 4-domain to 1-domain words in one Venn diagram are shown in Figure 7-4.

Figure 7-4 Examples (Translations) of Academic Vocabulary (4-domain to 1-domain Words)



Ha: Humanities & Arts, Ss: Social Sciences,
Tn: Technological Natural Sciences, Bn: Biological Natural Sciences

7.2.4.2 Part of speech of Japanese limited-academic-domain words

The overall proportion of parts of speech in limited-academic-domain words (LADs) (total 2,542 lexemes) is similar to common academic words (AWs); however, there is a difference to be pointed out.

LADs have more common nouns (1,605 words; LADs:LWs = 63.1:41.4) and fewer verbal nouns (633 words; LADs:LWs =24.9:34.0). The proportion of common nouns to all LADs (63.1%) is even higher than the proportion of nouns (including numerals) in the most frequent 20,000 VDRJ vocabulary (54.1%). Verbal nouns are fewer than AWs but are still more than the proportion of verbal nouns in the most frequent 20,000 VDRJ vocabulary (18.2%). This result shows the inclination to nouns in technical words. Adding all types of

nouns including verbal nouns, 2,104 words (87.9 %) can be nouns which is more than AWs (81.2%). Verbs (81 words (3.2 %) excluding verbal nouns) are even fewer than AWs (8.7%). Adding other types of verbs including verbal nouns together, 714 words (28.1%) can be a verb, which is less than AWs (42.7%).

Nominal adjectives (e.g. フル 'furu' (full), 偉大 'idai' (great)) make up 88 words (3.5 %) whose proportion is at the same level as AWs (3.7%). There are only 3 adjectives (e.g. 'katai' 硬い (stiff)) (0.1 %) listed in LADs, whose proportion is even lower than AWs (0.3%).

There are 109 affixes (e.g. –犯 '-han' (offense)) whose proportion (4.3%) is at the same level as AWs (4.1%). Affixes are very important in academic Japanese. There are 15 adverbs (e.g. 現に 'genni' (surely)) whose proportion (0.6 %) is less than AWs (1.3%). There are 9 words (0.8 %) which belong to other parts of speech such as particles or auxiliary verbs. In this category, similar to AWs, there are remarkably many archaic words, namely なり [affirmative aux.], とも (even though), たり [affirmative aux.], ごとし (as/like), 単なる (mere), しめる（＝しむ） [causative aux.] and かかる (such).

7.2.4.3 Word origins of Japanese limited-academic-domain words

Among 2,542 limited-academic-domain words (LADs = 2-domain and 1-domain words), there are 314 Japanese-origin words (12.4%), 1,757 Chinese-origin words (69.1%), 429 Western-origin (overwhelmingly English-origin) and other words (including some proper nouns) (16.9%) and 42 mixed-origin words (1.7%). Chinese-origin words account for a high proportion at 69.1% which is a little lower than common academic words (AWs) (75.2%) but still higher than the whole VDRJ (48.2%)[89]. The gap between LADs and AWs (75.2 - 69.1 = 6.1%) all goes to Western-origin words and other words (mostly English-origin words) at 16.9% which is more than double the proportion of Western origin for

[89] See Table 4-6 or 4-9 in Chapter 4.

common academic words at 7.0%. Japanese-origin words account for only 12.4% which is even less than 15.1% for common academic words. These proportions are much lower than the proportion of Japanese-origin words in the whole VDRJ (31.8%). These results show that Chinese-origin words are very dominant in academic vocabulary and Western-origin words are not generally used for a wide range of domains but for a more particular domain.

7.2.5 Conclusion of 7.2

In this section, after describing the classification of academic vocabulary and the method for extracting academic vocabulary, the distribution, semantic features and parts of speech of academic vocabulary are described.

The main findings in 7.2 are as follows.

1) Common academic words (4-domain and 3-domain words) are distributed mainly at the intermediate level. As the number of shared domains decreases to 2 and 1, the distribution of words moves to the lower-frequency range.
2) Many of the common academic words (4-domain and 3-domain words) are used for managing academic information. The meanings of common academic words are highly abstract. The Kanji used for common academic words also represent this feature. Limited-academic-domain words (2-domain words and 1-domain words) have more concrete meanings than common academic words.
3) Among the 2-domain words, the words specific to ‘humanities and arts’ and social sciences are mainly about history, especially political history. The words specific to social sciences and technological natural sciences are mainly about industry. The words specific to social sciences and biological natural sciences are mainly about social security, medical and nursing service. The words specific to technological and biological natural sciences are mostly common natural science words. However, there

seems no clear tendency for the words specific to ‘humanities and arts’ and technological or biological natural sciences.

4) Compared to the whole VDRJ vocabulary, academic vocabulary (common academic words and limited-academic-domain words) contains a much higher proportion of Chinese-origin words.
5) The proportions of verbal nouns and affixes in academic vocabulary are higher than the proportions in VDRJ, namely general Japanese. In contrast to that, the proportions of verbs, adjectives and adverbs in academic vocabulary are lower than the proportions in VDRJ.
6) The proportions of verbal nouns and verbs are higher for common academic words (4-domain and 3-domain words) than for limited-academic-domain words (2-domain and 1-domain words), while the proportion of common nouns is higher for limited-academic-domain words than for common academic words.

7.3 Literary words (LWs)

Literary vocabulary must be a group of words which are useful for reading literary works; however, there are various types of literary works (e.g. novels, poems, children’s stories) with a variety of topics (love, murder, family, religion and almost everything). It is still not clear if there is a ‘literary vocabulary’; however, it is possible to try to extract it using a statistical index such as the log-likelihood ratio if we have a large literary text corpus. In this section, after introducing the method for extracting literary words, I will show and discuss their distribution and examples, followed by their semantic features, parts of speech and word origins. The usefulness of the extracted literary words will be examined by checking text coverage in 7.4, along with an analysis of texts in different genres by the distribution of different groups of words.

7.3.1 Method for extracting Japanese literary words

The target corpus is the literary work texts (LW)[90], which are all imaginative texts, in the Balanced Contemporary Corpus of Written Japanese (BCCWJ) 2009 monitor version (NINJAL, 2009), the corpus used for this study. The literary work texts contain 8.25 million tokens.

The index used for extracting literary words is the log-likelihood ratio (Dunning, 1993) by using the 'keyness' function in a software tool AntConc (Anthony, 2007), which is the same index and the tool as for extracting academic vocabulary.

Nevertheless, the method is different from extracting academic vocabulary because there are no sub-sections in the literary text corpus. For extracting academic vocabulary, the academic texts were divided into four academic domains, and overlapping words extracted from the four academic domains were checked. However, the literary texts are not divided into sections but packed in one corpus as the target corpus. Therefore, for extracting literary words, I use four different 'reference' corpora shown below.

- Technical texts
- General texts in humanities and arts (Ha)
- General texts in the other 3 academic domains of social sciences (Ss), technological natural sciences (Tn) and biological natural sciences (Bn).
- Internet-forum (Yahoo Chiebukuro) texts

After extracting domain-specific words from literary texts using the four different reference corpora, the overlapping words from the four results are identified as 'literary words'. The cut-off point is set as the average value for each of the four extraction trials. The former JLPT Level 4 words (681 lexemes) are eliminated. The words ranked at 20,001 or lower are also eliminated. The remaining words are classified into basic to super-advanced levels by

[90] The literary work section is not a ready-made one. The texts are identified as part of the process of making sub-sections of the corpus by the author of this thesis. For details of the classification, see 3.3.2 in Chapter 3.

the Word Rankings for International Students (WIS).

7.3.2 Extracted Japanese 'literary words'

The list for literary words is available from the accompanying CD. Also, these words are easily identified in the columns for 'Possible Literary Keywords' in VDRJ.

7.3.2.1 Distribution and examples of Japanese literary words

The number and examples of literary words are shown in Table 7-13.

Table 7-13 Number and Examples of Japanese Literary Words (LWs) by Level

LW Label	F-JLPT Level	Word Rankings for International Students	Number of Lexemes of Literary Words	Least Frequent 2 Literary Words at Each Level	Translation of the Least Frequent 2 Literrary Words at Each Level
Basic Lit.	L3	682-1,291	142	ちっとも 引き出し	(not) at all drawer
Inter. Lit	L2 L1 Other	1,292-5,000	446	戸惑う 吐き出す	puzzled vent
Adv. 1 Lit.		5,001-10,000	483	不吉 銀色	ominous silver
Adv. 2 Lit.		10,001-15,000	345	敵機 口笛	hostile aircraft whistle
Super-adv.		15,000-20,000	200	香菜 樹海	coriander sea of trees
	Total		1,616		

*F-JLPT: The former Japanese Language Proficiency Test

The literary words are mainly distributed from intermediate to advanced level. The 05K to 10K (ranked between 5,000 and 10,000) level has the largest number of words at 483, and intermediate comes the second at 446. This distribution is slightly more biased to lower frequency than common academic words (Table 7-3) but at a similar level to limited-academic-domain words (Table 7-7). On the other hand, there are also 142 literary words at the basic level, which are more than the 70 common academic words and 75 limited-

academic-domain words at the basic level. Literary words are seemingly closer to the daily-life words.

How many literary words overlap with academic vocabulary (4-domain to 1-domain words)? The answer is only 27 words, which is 1.7% of literary words and 0.5% of academic vocabulary. This result means that academic texts and literary texts have considerably different lexical features.

Most of the overlapping words (24 words out of 27 words) overlap with 1-domain words while no literary words overlap with 4-domain words. 17 words overlap with the words specific to biological natural science. Many physical words such as words for body parts, e.g. 左手 'hidari-te' (left hand), こぶし 'kobushi' (fist), 血 'chi' (blood), 頭上 'zujou' (overhead), ひざ 'hiza' (knee), 全身 'zenshin' (whole body). Other examples of overlapping words are 音 'oto' (sound), 光 'hikari' (light), 棚 'tana' (shelf), 組 'kumi' (class), 岩 'iwa' (rock), 興奮 'koufun' (excitement), 帝 'mikado' (emperor), ネズミ 'nezumi' (mouse) and 帆 'ho' (sail). The overlapping words are mainly at the intermediate level but not at 11K or above. These words seem to be used frequently in the daily-life domain but are sometimes used for a scientific topic.

7.3.2.2 Semantic features of Japanese literary words

Looking over the extracted 'possible literary keywords', they are of course useful for learners who want to read Japanese literary works. There are some obvious features of literary words.

First, they contain numerous words related to the body. Not only basic words for body parts such as 首 'kubi' (neck) or 腕 'ude' (arm) but also many words for detailed body parts such as 指先 'yubisski' (fingertip) or まぶた 'mabuta' (eyelid).

Second, not surprisingly, there are also hundreds of words for body action. Examples are 立ち上がる 'tachiagaru' (stand up, rise to one's feet, (metaphor) rise up), 飛び出す

‘tobidasu’ (rush out, bounce out), 振り向く ‘furimuku’ (turn around, turn face about).

Third, there are more adverbs in literary words than in academic vocabulary. These literary adverbs often connote modal elements of sentences. Examples are ‘kitto’ ちっとも ‘chittomo’ ((not … as expected) at all) and たちまち ‘tachimachi’ (surprisingly instantly). There are also many mimetic words (擬態語 ‘gitai-go’) used as adverbs. Examples are きらきら ‘kirakira’ (sparkling, twinkling), ぐずぐず ‘guzuguzu’ (shilly-shally, dilly-dally), にやにや ‘niyaniya’ (grinning, simpering).

Fourth, there are many interjections. Examples are おや ‘oya’ (*mmm, oh*, expressing suspicion or surprise), へー ‘he^’ (*really?* Expressing a small surprise) and ほら ‘hora’ (*Look!*).

Fifth, there are some forms for colloquial contraction such as こりゃ ‘korya’ (= これは ‘kore wa’ (this is)) and ちまう ‘-chimau’ (= ‘-teshimau’, expressing completion of an action). There are also a few colloquial forms for the Kansai dialect such as はる ‘-haru’ (equivalent to ‘-irassharu’, used for honorific durative forms of verbs), どす ‘-dosu’ (equivalent to ‘-desu’ (be)) and さかい ‘-sakai’ (equivalent to ‘-dakara’ (because)).

Sixth, not surprisingly, there are numerous words which can be used for metaphorical expressions. For example, 振り向く ‘furimuku’ (turn around, turn face about) also means ‘to pay attention’. Other examples are 横たわる‘yokotawaru’ (lie down) for 前途に多くの困難が横たわる ‘zento ni ookuno kon’nan ga yokotawaru’ (many difficulties lie before us), かみしめる ‘kamishimeru’ (bite hard, chew thoroughly) for 幸せ／よこころび をかみしめる ‘shiawase/yorokobi o kamishimeru’ (deeply appreciate one’s happiness / savour the joy).

There are some problems particularly with some nouns such as トロッコ ‘torokko’ (trolley train) or 舞子 ‘maiko’ (dancing girl who is studying to be a geisha) which seem to be extracted from a particular text. Also, there are some nouns meaning daily-life things which are often described in literary texts but do not sound ‘literary’. Examples are ビール

‘bi^ru’ (beer), 馬 ‘uma’ (horse), 岩 ‘iwa’ (rock, crag) and ソファー ‘sofa^’ (sofa). These words should be excluded when elaborating a set of literary words.

7.3.2.3 Part of speech of Japanese literary words

The proportions of some parts of speech in literary words show a sharp contrast to the proportions in academic vocabulary. Proportions (counted by lexemes) of verbs (LWs:AWs:LADs = 34.0:8.7:3.2 (%)), adverbs (10.5:1.3:0.6) and interjections (2.6:0.0:0.0) are higher for literary words (LWs) than for academic vocabulary (common academic words (AWs) and limited-academic-domain words (LADs)). These proportions for literary words are also higher than the proportions among the most frequent 20,000 lexemes in VDRJ (Table 4-17 in Chapter 4) at 13.7%, 2.9% and 0.4% for verbs, adverbs and interjections respectively. On the other hand, proportions for verbal nouns (LWs:AWs:LADs = 5.3:34.0:24.9) and affixes (1.6:4.1:4.3) are lower for literary words than for academic vocabulary. These proportions for literary words are also lower than the proportions among the most frequent 20,000 lexemes in VDRJ (Table 4-17 in Chapter 4) at 18.2% and 2.5% for verbal nouns and affixes respectively.

These results are in accordance with the results of 4.4 in Chapter 4. This inevitably means literary words have fewer loanwords but more indigenous (Japanese-origin) words because verbs and interjections are basically of Japanese-origin while verbal nouns are mostly of Chinese or Western origins.

7.3.2.4 Word origins of Japanese literary words

As expected, the proportion (counted by lexemes) of word origins for Japanese literary words also shows a sharp contrast to academic vocabulary. Among all 1,616 Japanese literary words, 1,159 words (71.7%) are Japanese-origin, 352 words (21.8%) are Chinese-origin, 40 words (2.5%) are of Western and other origins, 50 words (3.1%) are of

mixed-origins, and 15 words are others (signs such as 々 (indicating repeating the previous Kanji), proper nouns and unknown word origin). The ratio between Japanese-origin words and Chinese-origin words in Japanese literary words is almost the opposite to the ratio in academic vocabulary. This result tells us how a learner's language background will possibly affect the understanding of texts in different genres. This issue is to be discussed in 7.4.5.2.3.

7.3.3 Conclusion of 7.3

In this section, I described the method for extracting literary words and the features of literary words from various aspects. The main findings in 7.3 are as follow.

1) Literary words are mainly distributed from intermediate to advanced level.
2) Only 27 literary words overlap with academic vocabulary. The 27 words account for 1.7% of literary words and 0.5% of academic vocabulary.
3) Literary words contain numerous words for body parts and body actions.
4) Literary words contain many modal adverbs and interjections.
5) Literary words contain many words for metaphorical expressions.
6) Extracted literary words contain some words for daily-life things which are often described in literary texts but do not sound 'literary'.
7) The proportions (counted by lexemes) of verbs, adverbs and interjections are high in literary words. The proportions of verbal nouns and affixes are low in literary words. These show a sharp contrast to common academic words.
8) In contrast to the Japanese academic vocabulary, the proportion (counted by lexemes) of Japanese-origin words is very high in literary words. On the other hand, the proportion of Chinese-origin words is low in literary words.

All in all, literary words are the words for describing human actions and feelings vividly and effectively. Though there are some exceptions, the words seem to be worth being included in a list for learners who want to read Japanese literary works. If the literary texts can be divided into some sub-genres such as romance, detective stories and so on, we may be able to create a better word list for reading literary works.

7.4 Testing word tiers by lexical profiling

In this section, I will examine what position the extracted domain-specific words in the previous sections (i.e. common academic words, limited-academic-domain words, literary words) occupy in different genres to prove their usefulness by checking text coverage and Text Covering Efficiency (TCE, to be proposed in 7.4.1.2).

Based on these analyses and the 'word tier analysis' to be proposed in 7.4.5, I will give an answer to the main research questions for this thesis: In what order should learners of Japanese as a second language learn words and characters in order to be able to read Japanese? How will the order vary according to the purpose of learning?

How the word tiers work in different genres (register variations) and how a learner's language background possibly affects the understanding of texts in different genres will also be discussed.

7.4.1 Methods

7.4.1.1 Testing text coverage

There are both qualitative and quantitative ways for evaluating a vocabulary list developed for learning and teaching. In order to look at the efficiency of vocabulary learning which is the main purpose of study, I will first look at text coverage by the extracted common academic words, limited-academic-domain words and literary words[91].

[91] See 2.2.2 for the importance of text coverage. Average text coverage per lexeme (entitled Text Covering Efficiency: TCE) will also be proposed as a measure of efficiency in 7.2.4.2.

For testing text coverage, baseword files of these groups of words were created for AntWordProfiler (Anthony, 2009).

The hypotheses to be tested are:

1) Text coverage by common academic words is higher in different types of academic text than in other types of texts e.g. conversation or literary texts.
2) Text coverage by limited-academic-domain words is higher in the academic texts in the target domain than in other types of texts e.g. texts in a non-target academic domain or literary texts.
3) Text coverage by literary words is higher in the texts of literary works than in other types of texts e.g. academic texts or newspaper texts.

These will be tested in both 1) the texts used for developing the list i.e. the technical texts in the Balanced Contemporary Corpus of Written Japanese, 2009 monitor version (NINJAL, 2009) and 2) test corpora which are not used for developing the list. It is important to test lists on corpora which are not those from which they were made. Eleven test corpora shown below are used for this study. (The number of tokens in each text is also shown in related tables from 7-15 to 7-33.)

JS-Bn: J-Stage texts in biological natural sciences. Journal articles on environmental studies, physical education, health and sports science, which were downloaded from J-STAGE (Japan Science & Technology Information Aggregator) at http://www.jstage.jst.go.jp/browse/-char/ja. This test corpus contains 0.72 million running words from four types of academic journals.

MTT-Bn: Meidai Technical Texts in Biological Natural Sciences. 0.01 million running words from the a volume of model lecture texts out of the nine volumes of "Technical

Lectures in Japanese for International Students" edited by the members of Nagoya University (Meidai)[92].

JS-Tn: J-Stage texts in technological natural sciences. Journal articles on electricity and civil engineering, which were downloaded from J-STAGE (Japan Science & Technology Information Aggregator) at http://www.jstage.jst.go.jp/browse/-char/ja. This test corpus contains 2.71 million running words from four types of academic journals.

MTT-Tn: Meidai Technical Texts in Technological Natural Sciences. 0.07 million running words from the five volumes of model lecture texts out of the nine volumes of "Technical Lectures in Japanese for International Students" edited by the members of Nagoya University (Meidai)[93].

MTT-Ss: Meidai Technical Texts in Social Sciences. 0.05 million running words from the three volumes of model lecture texts out of the nine volumes of "Technical Lectures in Japanese for International Students" edited by the members of Nagoya University (Meidai)[94].

TB: Text Bank in Social Sciences for Intermediate and Advanced Learners of Japanese. 0.19 million running words from the body of the text bank.

TIS: Texts for International Students (Shinya & Matsushita, 1994). An edited textbook in international studies, which mostly contains social science texts but a few texts on humanities. 0.04 million thousand running words.

UYN: Utiyama Yomiuri Newspaper Corpus. 5.68 million running words from the Yomiuri newspaper articles published from 1989 to 2001. The Japanese data from the Japanese-English News Article Alignment Data (JENAAD) (Utiyama & Isahara, 2003).

---

[92] MTT texts are lecture texts; yet, they basically have the features of written texts. See also footnote 61 in 3.5.1.

[93] See above.

[94] See above.

BSB: Best Seller Books. This is contained in BCCWJ 2009 monitor version but was not used to create VDRJ and JAWL. This corpus is mostly composed of literary works (novels etc.) but includes some different types of texts such as critiques and essays. Total of 2.10 million running words.

UPC: Utiyama Parallel Corpus. 2.30 million running words from literary works including essays, novels and stories. The Japanese data from the English-Japanese translation alignment data (Utiyama & Takahashi, 2003). Downloaded from http://www2.nict.go.jp/x/x161/members/mutiyama/align/index.html on 16 November 2010.

MC: Meidai Conversation Corpus: 1.13 million running words from various types of pair or group conversation at cafés, schools, homes or other places. Compiled by the members of Nagoya University (Meidai). Downloaded from http://dbms.ninjal.ac.jp/nknet/ndata/nuc/ on 10 December 2010.

7.4.1.2 The idea of Text Covering Efficiency (TCE)

To evaluate a group of words as a target for learning, I propose an index entitled Text Covering Efficiency (TCE). TCE is calculated by dividing text coverage (tokens) of a group of words by the number of lexemes of the group extracted from the BCCWJ (the corpus used for this study), and then dividing the quotient by the total number of tokens in the target text (domain) to adjust the difference in size of the texts and make the figures from differently-sized texts comparable. For the user's convenience, the figure is multiplied by 1,000,000. The solution means the expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain. Therefore, it is comparable with the standardized frequency per million. In other words, TCE is an expected standardized frequency of a grouped lexeme in a text.

The formula for TCE is as follows.

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

E: Text covering efficiency = Expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain

F_t: Number of tokens of the tested group of words in the target text

L_{tw}: Number of lexemes of the tested group of words extracted from BCCWJ

N_t: Number of tokens in the target text

The idea behind TCE is simply that it is better to gain more text coverage by a smaller number of learned lexemes. In other words, even if a group of words provide high text coverage, it will not always be efficient to learn the group of words if the group has many lexemes to learn. Therefore, the average number of tokens to be covered by a word in the group needs to be calculated. High efficiency in vocabulary learning is that more words in a text are covered by fewer learned words. TCE is assumed to predict the average efficiency in gaining text coverage by learning a word of the group.

This is a converse idea to the type/token ratio (TTR) which is an index to measure the lexical diversity of a text mainly adopted in first language acquisition research. TTR is calculated by dividing the number of types by the number of tokens. For language development, the more types in a text, the better. However, the task here is to evaluate a group of words as a source for covering a text. Therefore, the more the average number of tokens in a text covered by a lexeme, the better. If a group of words returns a high TCE, learning that particular group of words will be an efficient way to gain the coverage of the target text.

As argued about TTR (Richards & Malvern, 1997), the relationship between the numbers of tokens and lexemes will be different depending on the text size. Nevertheless, it

is not a problem for TCE because the formula does not use the number of lexemes occurring in the text but uses the number of lexemes of the target group of words. This is a reasonable idea because learners generally do not know which words will occur in a particular text. For example, to evaluate the value of the intermediate literary words as a source for gaining the text coverage, it is reasonable to divide the tokens by the number of lexemes of the intermediate literary words which a learner will learn before s/he reads the text.

Table 7-14 Mean Frequency per Million for Each 1,000 Word Level in Word Ranking for International Students (WIS)

Level	WIS	Mean Frequency	1,000 Word Level	Mean Frequency
Basic	1-1,291	548.5	01K	694.6
			02K	102.0
Intermediate	1,292-5,000	46.5	03K	40.6
			04K	23.1
			05K	15.3
Adv. 1	5,001-10,000	7.9	06K	11.5
			07K	9.1
			08K	7.4
			09K	6.2
			10K	5.2
Adv. 2	10,001-15,000	3.5	11K	4.4
			12K	4.0
			13K	3.4
			14K	2.9
			15K	2.7
S-Adv.	15,001-20,000	1.9	16K	2.3
			17K	2.1
			18K	1.9
			19K	1.7
			20K	1.5

TCE figures can be compared with standardized frequency per million. Table 7-14 shows the mean frequency for each 1,000 word level. Comparing the TCE figures with the figures in 7-14, we can see what ranking of a general word a domain-specific word is equivalent to. For example, if a TCE figure of a grouped word is over 15, the words are at least as valuable as general intermediate words because the standardized frequency per million for 05K is 15.3. When checking TCE figures, it will be useful to remember the figures shown in Table 7-14 to assess the value of TCE figures.

7.4.1.3 Domain-specified analysis and domain-unspecified analysis

When testing text coverage of 3-domain, 2-domain or 1-domain words of academic

vocabulary, there are two ways for testing. One is domain-specified analysis and the other is domain-unspecified analysis. Let us suppose a 3-domain word is specific in three domains of 'humanities and arts' (Ha), social sciences (Ss) and technological natural sciences (Tn) but not specific in biological natural sciences (Bn). When you test the coverage of an Ha text, Ss text or Tn text, the 3-domain words can be included in the coverage; however, for the biological natural science text, the word may only be able to behave as a general word. In this case, if you do not include the word in the coverage by the 3-domain words, that is domain-specified analysis. If you still include the word in the coverage by the 3-domain words, that is domain-unspecified analysis.

Specifying a domain for an analysis will be more important for 2-domain and 1-domain words. 1-domain words for humanities and arts are not likely to show high text coverage for a medical text. If all 1-domain words for the four academic domains are included in the coverage of a biology or politics text, it is hard to tell which group of 1-domain words provide high text coverage.

To conduct the domain-specified analysis, many different sets of baseword lists need to be created. However, the results will be more elaborated and useful. If you cannot specify a domain for the target text (e.g. non-academic texts or academic texts with mixed genres), you can only conduct a domain-unspecified analysis. For each analysis in this chapter, I will show which type of analysis method I adopt.

7.4.2 The usefulness of JAWL (common academic words)

7.4.2.1 Text coverage and Text covering efficiency by Japanese common academic words

Table 7-15 shows text coverage of the BCCWJ (the whole), BCCWJ-T (the academic texts used for extracting the academic vocabulary) and the test corpora in different genres by different levels of the common academic words as well as non-JAWL

(non-academic) basic words on the top. The genres are sorted in JAWL Ⅰ text coverage order (high JAWL Ⅰ coverage on the right).

The table clearly shows that academic texts have higher text coverage than non-academic texts. It also shows that JAWL Ⅰ and II are the most important levels. (Common academic words at the basic level are also important; however, I do not put much focus on them because they are much fewer in number and all basic words are important anyway.)

First of all, I will look at JAWL Ⅰ since the number of words is 559 which is very close to the Academic Word List (570 words). The text coverage of the technical texts used for extracting the common academic words is 11.1% (see 'BCCWJ-T' in Table 7-15) which is close but higher than the figure of the Academic Word List at 10.0%. Of course, we cannot attempt an easy comparison since the Academic Word List does not contain the words listed in the General Service List which contains around 2,000 words while JAWL Ⅰ only excludes basic 1,288 lexemes listed in the former Japanese Language Proficiency Test Level 4 and 3. The units of counting are not exactly the same and the structures of the languages are also different. However, JAWL Ⅰ at least can provide coverage which can be compared favourably with AWL.

Text coverage of the academic texts in test corpora by the Academic Word List is 8.5% (Coxhead, 2000) or 9.3-11.1% (Hyland & Tse, 2007). JAWL Ⅰ also provides consistently high text coverage of the academic texts of the test corpora in different science fields at 9.7-15.1% (Table 7-15). Coverage by JAWLⅠ is highest in journal articles at 13.5% (Bn) and 15.1% (Tn). JAWLⅠ also has high coverage of the other academic texts including introductory ones at 9.7-11.1%. Newspapers seems to have similar lexical features to academic texts as they contain 8.7% JAWLⅠ words. Newspapers also contain many JAWL II words at 6.6% which is the highest among all genres.

Table 7-15 Text Coverage in Different Genres by the Different Levels of Japanese Common Academic Words *Domain-unspecified

Corpus Label																				
Text Genre					MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
					Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
Text Size (Total Tokens)					1,129,538	2,298,828	2,102,178	**32,819,424**	5,675,357	186,768	50,601	42,152	13,904	74,645	**2,895,425**	719,802	2,705,026
JAWL Label	# of Words	F-JLPT	WIS	Level	# of HSD	Text Coverage (%)											
Basic (non-JAWL)	1,242	L4 L3	1-1,291	Basic	0	80.6	72.4	73.0	**68.6**	57.0	62.5	66.2	62.0	60.4	59.8	**58.5**	52.2	50.8
JAWL 0	31	L3	682-1,291	Basic	4	0.6	1.3	1.2	**1.6**	2.1	2.8	3.0	2.4	2.7	3.7	**2.7**	3.4	3.3
	39				3	0.6	1.3	1.4	**1.4**	1.9	2.4	2.2	2.4	1.6	1.8	**1.9**	1.7	1.5
JAWL I	**559**		1,292-5,000	**Inter.**	**4**	**0.8**	**2.7**	**3.1**	**4.6**	**8.7**	**9.7**	**9.8**	**10.2**	**11.1**	**11.1**	**11.1**	**13.5**	**15.1**
JAWL II	541				3	0.5	1.6	1.5	**2.6**	6.6	5.0	4.8	4.7	4.2	2.9	**5.6**	5.2	4.5
JAWL III	212	L2	5,001-10,000	Adv. 1	4	0.0	0.1	0.1	**0.2**	0.3	0.4	0.4	0.2	0.4	0.7	**0.5**	0.9	1.2
JAWL IV	451	L1			3	0.1	0.2	0.2	**0.4**	0.8	0.7	0.4	0.6	2.1	1.9	**0.9**	1.2	1.2
JAWL V	103	Othe	10,001-15,000	Adv. 2	4	0.0	0.0	0.0	**0.0**	0.1	0.1	0.0	0.1	0.1	0.1	**0.1**	0.2	0.2
JAWL VI	327	r			3	0.0	0.1	0.1	**0.1**	0.2	0.2	0.1	0.2	0.3	0.6	**0.3**	0.4	0.4
JAWL VII	56		15,000-20,000	Super-adv.	4	0.0	0.0	0.0	**0.0**	0.0	0.0	0.0	0.0	0.0	0.1	**0.0**	0.1	0.1
JAWL VIII	268				3	0.0	0.0	0.0	**0.0**	0.1	0.1	0.1	0.1	0.1	0.4	**0.1**	0.2	0.3

* JAWL: Japanese Common Academic Word List
* F-JLPT: the former Japanese Language Proficiency Test
*WIS: Word Rankings for International Students
*# of HSD: Number of High-Specificity Domains Out of the 4 Large Academic Domains
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

To confirm the high text coverage of academic texts by JAWL Ⅰ and II, a Chi-square test (test of independence) was conducted on the tokens of JAWL Ⅰ and II and the other words between the three non-academic test corpora (MC, BSB and UPC) and seven academic test corpora (TB, MTT-Ss, TIS, MTT-Bn, MTT-Tn, JS-Bn and JS-Tn). The result is significant (χ^2 = 8653486.191, df = 2, p < .001) showing the distribution of JAWL Ⅰ and II is not the same as the other words across the non-academic and academic texts.

Text coverage of non-academic texts, on the other hand, by the Academic Word List is very low at 1.4% (fiction texts) while the coverage of non-academic texts by JAWL Ⅰ is 0.8% (conversation), 2.7% (UPC, general books including novels and essays) and 3.1% (BSB, dominantly literary texts). JAWL Ⅰ's coverage is a little higher than the Academic Word List but it is lower than the coverage of academic texts by 7-14%. This also proves that JAWL Ⅰ is a valid and useful list.

It is also obvious that text coverage by the common academic words (especially JAWL Ⅰ and II) are in inverse proportion to the coverage by non-JAWL basic words. As the proportion of the non-JAWL basic words decreases, the proportion of JAWL Ⅰ and II increases. Table 7-16 shows that the cumulative text coverage by all the basic words (including JAWL 0) and JAWL Ⅰ and II (2,412 lexemes in total). The coverage keeps almost the same levels at around 80% throughout the genres except for academic journals where many technical words are expected to be contained.

Table 7-16 Cumulative Text Coverage in Different Genres by the Basic and JAWL Ⅰ and II words *Domain-unspecified

Corpus	MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
# of Words	Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal	Tn (Journal
2,412	83.1	79.2	80.2	**78.8**	76.2	82.3	86.0	81.6	80.0	79.2	**79.9**	75.9	75.3

* JAWL: Japanese Common Academic Word List
* F-JLPT: the former Japanese Language Proficiency Test
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

Let us look at what common academic words are frequently used in these academic

texts. JAWL Ⅰ provides 15.1% coverage of journal articles in technological natural sciences (Tn). This is a notably high coverage. The most frequent common academic words in this corpus are 拠る (according to), 的 '-teki' (-like (a suffix which changes a noun into an adjectival noun)) , 示す (show, indicate), 性 (-ity, (a suffix)), 於く (in, at (formal)). These words account for 2.0% in total. It is high, yet it is not only one or two words that provide the high text coverage. Some high-frequency words are highly abstract which behave like function words. However, there are also some high-frequency content words such as 用いる 'mochiiru' (use (formal)), 図 'zu' (chart, diagram, figure), 値 (value, count, number), 結果 'kekka' (result) and 変化 'henka' (change).

Below is a sample text from an academic item from Wikipedia. The bold types without underlining show basic words (including JAWL 0) and the underlined types show the words listed in JAWLⅠ.

Sample Text

人類学は一般に、人類の進化や生物学的側面を研究する自然人類学と、人類の社会的・文化的側面を研究する文化人類学(Cultural Anthropology)あるいは社会人類学(Social Anthropology)に大別される。文化人類学の名称はアメリカにおいて用いられ、イギリスおよび多くのヨーロッパ諸国では「社会人類学」の名称が用いられてきた。他のヨーロッパ諸国や日本においては民族学（英語圏での Ethnology、ドイツ語圏での Ethnologie）の名称も用いられている（民族学を一分野とする場合も多い）。民俗学（Folklore）もまた隣接分野として共通の研究テーマを共有することが多い。

自然人類学は、人類を進化の過程によって形作られてきた生物学的側面から捉える。それに対して、文化人類学は自然の対義としての文化から人類を研究しようとする学問分野である。文化とは、進化の過程を経て形成された遺伝的な形質のことではなく、人類が後天的に学習した行動パターンや言語、人工物の総体を指している。したがって文化人類学の隣接科学には言語学と考古学があり、アメリカの学部ではこれらの学問に加えて自然人類学をあわせて総合的に教育されている。

(Cited from the item 文化人類学 'Bunka-jinrui-gaku' (Cultural Anthropology) in Wikipedia)

In this text, basic words account for 57.7% (including 6.8% JAWL 0 words) and JAWLⅠ account for 20.4% (78.1% in total) of the total tokens in the text[95]. Adding 6.4% JAWL II (9 lexemes, 17 tokens, e.g. 進化 'shinka' (evolution), 生物 'seibutsu' (creature, living thing), 自然 'shizen' (nature)) and 11.7% non-JAWL intermediate words (11 lexemes, 31 tokens, e.g. 人類 'jinrui' (the human species), 名称 'meishou' (name, title), ヨーロッパ 'yo^roppa' (Europe)), cumulative text coverage reaches 96.2%.

Let us look at JAWL III or above. Text coverage is not high by JAWL III or above; however, the number of lexemes of JAWL III or above is also smaller than JAWLⅠ or II. Therefore, Text Covering Efficiency (TCE) should be checked. (For the formula for TCE, see 7.4.1.2.)

As shown in Table 7-17, JAWL III to VIII also provide much higher TCE (the expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain) for academic texts than for non-academic texts. TCE of JAWL III and IV (05K-10K) ranges from 10 to 54 for academic texts but from 1 to 5 for non-academic texts. As shown in Table 7-18, learning JAWLⅠ and II is 4.7 times more efficient in covering academic texts than non-academic texts and JAWL III-VIII is around 8 times (7.4-9.6 times) more efficient. The efficiency level increases as the frequency level goes to lower levels. Compared to the JAWLⅠ and II, learning JAWL III-VIII is less efficient; however, it is around 8 times more efficient in covering academic texts than non-academic texts. Considering the fact that thousands of words are required to gain 1% coverage at this level, JAWL III-VIII are also good lists for academic purposes.

[95] Academic items of Wikipedia seem to contain more academic words than other academic texts. I tested text coverage of JAWLⅠ words on a few academic items of Wikipedia. The results are all 15-20%. If this is generally true, academic items of Wikipedia should be a very good resource for learning academic words. Also, it may be true that some academic words are encyclopaedic words used for explaining various ideas and concepts. Wikipedia seem to contain more proper nouns and low-frequency words (21K+) as well.

Table 7-17 Text Covering Efficiency (TCE) of the Different Levels of Japanese Common Academic Words by Genre *Domain-unspecified

Corpus Label																	
Text Genre					MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
					Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
Text Size (Total Tokens)					1,129,538	2,298,828	2,102,178	**32,819,424**	5,675,357	186,768	50,601	42,152	13,904	74,645	**2,895,425**	719,802	2,705,026
JAWL Label	# of Words	F-JLPT Level	WIS	Level	# of HSD	TCE: Text Covering Efficiency = Expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.											
Basic (non-JAWL)	1,242	L4 L3	1-1,291	Basic	0	649	583	588	**552**	459	503	533	499	486	481	**471**	420	409
JAWL 0	31	L3	682-1,291	Basic	4	187	405	382	**525**	667	888	966	773	882	1,178	**856**	1,099	1,069
	39				3	166	337	356	**354**	475	604	569	611	422	460	**493**	425	384
JAWL I	**559**		1,292-5,000	**Inter.**	**4**	**14**	**47**	**56**	**82**	**156**	**174**	**175**	**182**	**198**	**198**	**199**	**241**	**271**
JAWL II	541				3	10	29	27	**48**	122	93	89	86	78	53	**104**	96	84
JAWL III	212	L2	5,001-10,000	Adv. 1	4	1	4	5	**8**	15	17	18	11	21	32	**24**	41	54
JAWL IV	451	L1			3	1	4	5	**8**	17	16	10	13	47	42	**20**	27	27
JAWL V	103	Other	10,001-15,000	Adv. 2	4	1	2	2	**3**	5	5	5	6	6	8	**10**	23	18
JAWL VI	327				3	1	2	2	**3**	8	6	4	5	9	19	**9**	13	13
JAWL VII	56		15,000-20,000	Super-adv.	4	0	1	1	**2**	3	4	4	7	8	18	**6**	10	21
JAWL VIII	268				3	0	1	1	**2**	4	3	3	2	2	14	**5**	8	11

*TCE means the expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.
* JAWL: Japanese Common Academic Word List
* F-JLPT: the former Japanese Language Proficiency Test
*WIS: Word Rankings for International Students
*# of HSD: Number of High-Specificity Domains Out of the 4 Large Academic Domains
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

Table 7-18 Means and Standard Deviations for TCE of Common Academic Words in Academic and Non-academic Texts by Level

Level	Non-academic Texts		Academic Texts		Ratio for M (Aca/Non-aca.)
	M	SD	M	SD	
Basic	305.5	93.9	737.8	266.1	2.4
Inter.	30.5	16.7	144.2	66.6	4.7
Adv. 1	3.4	1.6	26.9	13.9	7.9
Adv. 2	1.3	0.6	10.0	6.1	7.4
S-Adv.	0.8	0.3	8.1	5.9	9.6

*Non-academic texts include MC, BSB and UPC. Academic texts include TB, TIS, all MTT and JS texts.

These figures prove that JAWL is a set of appropriate word lists for efficient vocabulary learning for academic purposes. Also, the method for extraction is also proven to be appropriate.

7.4.2.2 Different behaviour of Japanese common academic words in different domains

Newspapers show a similar text coverage and TCE to social science (Ss) texts (Table 7-15 and 7-17). Newspapers contain slightly fewer basic words but slightly more JAWL II and IV (3-domain words) than social science texts; however, newspaper articles will be a good resource for learning common academic words, especially for social sciences.

It is also clear that (both technological and biological) natural science texts (Tn and Bn) contain more JAWL words at the advanced levels. TCE of JAWL II (intermediate 3-domain words) ranges from 86 to 93 for social science (Ss) texts but from 53 to 78 for introductory natural science texts, while TCE of JAWL III and IV (advanced 4-domain and 3-domain words) ranges only from 10 to 18 for social science texts but from 21 to 47 for natural science texts. This result is in line with English studies (Coxhead & Hirsh, 2007; Coxhead, Stevens, & Tinkle, 2010). This should be re-examined when examining the limited-academic-domain words later.

Journal articles show notably higher coverage and TCE than other types of academic texts. In particular, TCE figures for journal articles at the super-advanced level (16K-20K) are surprisingly high at 8-11, compared to the average standardized frequency per million for this level at 1.92. This is also strong support evidence for the validity of JAWL.

Remaining issues and future research for common academic words will be mentioned in 7.5, taking account of the results of the tests for the other domain-specific words.

7.4.3 The usefulness of Japanese limited-academic-domain words

Table 7-19 and 7-20 show the text coverage in different genres by Japanese limited-academic-domain words (7-19 for domain-unspecified analysis and 7-20 for domain-specified analysis). (The genre order follows Table 7-15 and 7-17 for common academic words.) For domain-specified analysis, the specified domain is fixed as the domain whose intermediate 1-domain words show the highest Text Covering Efficiency (TCE).

Text coverage for limited-academic-domain words (LADs) is much lower than common academic words; however, not surprisingly, the overall distribution pattern is similar to common academic words. According to the domain-unspecified analysis shown in Table 7-18, text coverage by LAD Ⅰ and II (intermediate, 704 words in total) ranges from 0.8% to 3.5% for academic texts while it ranges from 0.5% to 1.3% for non-academic texts. According to the domain-specified analysis shown in Table 7-19, text coverage by LAD Ⅰ and II (intermediate, 300, 384, 211 or 200 words in total in each domain) ranges from 0.4% to 3.2% for academic texts while it ranges from 0.1% to 0.9% for non-academic texts.

To confirm the high text coverage of academic texts by LAD Ⅰ and II, a Chi-square test (test of independence) was conducted on the tokens of LAD Ⅰ and II and the other words between the three non-academic test corpora (MC, BSB and UPC) and seven academic test corpora (TB, MTT-Ss, TIS, MTT-Bn, MTT-Tn, JS-Bn and JS-Tn). The result is significant ($\chi^2 = 9085386.25$, $df = 2$, $p < .001$) showing the distribution of LAD Ⅰ and II is not the same as the other words across the non-academic and academic texts.

Table 7-19 Text Coverage in Different Genres by Different Levels of Japanese Limited-academic-domain Words *Domain-unspecified

Corpus Label																	
Text Genre					MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
					Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
Text Size (Total Tokens)					1,129,538	2,298,828	2,102,178	**32,819,424**	5,675,357	186,768	50,601	42,152	13,904	74,645	**2,895,425**	719,802	2,705,026
LAD Label	# of Words	F-JLPT Level	WIS	Level	# of HSD	Text Coverage (%)											
LAD 0	45	L3	682-1,291	Basic	2	0.2	0.3	0.3	**0.4**	1.2	0.9	0.8	1.4	0.2	0.5	**0.6**	0.5	0.5
	33				1	0.2	0.3	0.2	**0.3**	0.5	0.4	0.3	0.6	0.6	0.1	**0.3**	0.2	0.2
LAD I	**391**		1,292-5,000	**Inter.**	**2**	**0.5**	**1.3**	**1.0**	**1.9**	**4.4**	**3.5**	**3.1**	**3.5**	**3.1**	**1.7**	**3.2**	**2.3**	**1.9**
LAD II	313				1	0.5	1.2	1.1	**1.6**	2.8	2.5	2.2	2.1	1.0	1.0	**2.4**	1.3	0.8
LAD III	429	L2	5,001-10,000	Adv. 1	2	0.1	0.2	0.2	**0.4**	0.9	0.6	0.6	0.7	0.9	1.2	**0.8**	0.9	1.1
LAD IV	359	L1			1	0.1	0.2	0.2	**0.4**	0.7	0.6	0.5	0.3	0.2	0.2	**0.8**	0.4	0.5
LAD V	296	Other	10,001-15,000	Adv. 2	2	0.0	0.1	0.1	**0.1**	0.3	0.1	0.1	0.2	0.4	0.7	**0.3**	0.3	0.4
LAD VI	247				1	0.0	0.1	0.1	**0.1**	0.2	0.2	0.0	0.1	0.1	0.3	**0.3**	0.2	0.4
LAD VII	232		15,000-20,000	Super-adv.	2	0.0	0.0	0.0	**0.1**	0.1	0.1	0.0	0.0	0.2	0.2	**0.2**	0.2	0.2
LAD VIII	197				1	0.0	0.0	0.0	**0.1**	0.1	0.1	0.0	0.0	0.1	0.1	**0.2**	0.1	0.1

* LAD: Limited-academic-domain words
*WIS: Word Rankings for International Students
*# of HSD: Number of High-Specificity Domains Out of the 4 Large Academic Domains
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

Table 7-20 Text Coverage in Different Genres by Different Levels of Japanese Limited-academic-domain Words *Domain-specified

Corpus Label																			
Text Genre					MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
					Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
Specified-domain for Domain-specified Analysis					Tn	Ha	Ha	**Ss**	Ss	Ss	Ss	Ss	Bn	Tn	**Ss**	Bn	Tn
Text Size (Total Tokens)					1,129,538	2,298,828	2,102,178	**32,819,424**	5,675,357	186,768	50,601	42,152	13,904	74,645	**2,895,425**	719,802	2,705,026
LAD Label	# of Words for Ha	# of Words for Ss	# of Words for Tn	# of Words for Bn	F-JLPT Level	WIS	Level	# of HSD	Text Coverage (%)								
LAD 0	24	26	20	20	L3	682-1,291	Basic	2	0.1	0.2	0.1	**0.3**	1.1	0.8	0.7	1.2	0.1	0.4	**0.5**	0.3	0.4
	13	6	5	9				1	0.0	0.2	0.1	**0.0**	0.1	0.1	0.0	0.1	0.5	0.0	**0.1**	0.1	0.0
LAD I	**196**	**273**	**165**	**148**		1,292-5,000	**Inter.**	**2**	**0.2**	**0.9**	**0.6**	**1.4**	**3.9**	**3.2**	**2.9**	**2.9**	**2.6**	**1.5**	**2.7**	**1.4**	**1.5**
LAD II	104	111	46	52				1	0.1	0.6	0.6	**0.6**	1.9	1.9	1.3	0.9	0.5	0.8	**1.4**	0.5	0.4
LAD III	201	274	216	167	L2 L1	5,001-10,000	Adv. 1	2	0.1	0.1	0.1	**0.2**	0.7	0.5	0.6	0.7	0.8	1.1	**0.6**	0.7	1.0
LAD IV	104	127	60	68				1	0.0	0.1	0.1	**0.1**	0.5	0.5	0.4	0.2	0.1	0.1	**0.5**	0.1	0.3
LAD V	141	186	146	119	Other	10,001-15,000	Adv. 2	2	0.0	0.0	0.0	**0.1**	0.2	0.1	0.1	0.1	0.4	0.7	**0.2**	0.2	0.4
LAD VI	71	74	48	54				1	0.0	0.0	0.0	**0.0**	0.2	0.2	0.0	0.1	0.1	0.3	**0.2**	0.1	0.3
LAD VII	133	152	106	73		15,000-20,000	Super-adv.	2	0.0	0.0	0.0	**0.0**	0.1	0.1	0.0	0.0	0.2	0.2	**0.1**	0.2	0.2
LAD VIII	60	55	29	53				1	0.0	0.0	0.0	**0.0**	0.1	0.1	0.0	0.0	0.0	0.0	**0.1**	0.1	0.1

* LAD: Limited-academic-domain words
* F-JLPT: the former Japanese Language Proficiency Test
*WIS: Word Rankings for International Students
*# of HSD: Number of High-Specificity Domains Out of the 4 Large Academic Domains
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

There are some interesting differences between common academic words and limited-academic-domain words. As I mentioned in 7.4.2.1, text coverage by common academic words is inversely proportional to non-JAWL basic words. Interestingly, common academic words are also in inverse proportion to the coverage by LAD Ⅰ and II. In addition, LAD III to VIII seem to be in inverse proportion to LAD Ⅰ and II but in proportion to JAWL Ⅰ and II.

In sum, natural science texts are covered more by JAWL Ⅰ and II (4-domain and 3-domain intermediate words) and LAD III to VIII (2-domain and 1-domain advanced words) than social science texts while social science texts are covered more by non-JAWL basic words and LAD I and II (2-domain and 1-domain intermediate words).

Another interesting thing is that newspaper texts show the highest LAD coverage among all the genres. Overall text coverage of newspapers is similar to social science texts; however, newspapers contain more advanced LADs. Newspapers will not use too many technical words as they are published for the general public; however, they tend to use fewer basic words and wide range words but more intermediate words and limited-academic-domain words than other genres. Newspaper articles seem to be expected to provide technical information to some extent in a way that general adult readers can understand.

How different is the efficiency level depending on the genre? How efficient is learning LADs compared to common academic words (4-domain and 3-domain words)? To answer these questions, Text Covering Efficiency is calculated for LADs (Table 7-21 for domain-unspecified analysis and Table 7-22 and 7-23 for domain-specified analysis). Text covering efficiency figures for LADs are combined with the ones for the common academic words (the words listed in JAWL) in Table 7-24 (domain-specified analysis is done only for LADs but not for JAWL).

Table 7-21 Text Covering Efficiency (TCE) of Different Levels of Japanese Limited-academic-domain Words by Genre *Domain-unspecified

Corpus Label						MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
Text Genre						Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
Specified-domain for Domain-specified Analysis						Tn	Ha	Ha	**Ss**	Ss	Ss	Ss	Ss	Bn	Tn	**Ss**	Bn	Tn
Text Size (Total Tokens)						1,129,538	2,298,828	2,102,178	**32,819,424**	5,675,357	186,768	50,601	42,152	13,904	74,645	**2,895,425**	719,802	2,705,026
LAD Label	# of Words	F-JLPT Level	WIS	Level	# of HSD	TCE: Text Covering Efficiency = Expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.												
LAD 0	45	L3	682-1,291	Basic	2	41	74	59	**96**	274	195	178	302	48	113	**144**	113	119
	33				1	57	99	74	**103**	153	117	86	182	183	35	**95**	61	59
LAD I	**391**		1,292-5,000	**Inter.**	**2**	**14**	**33**	**25**	**47**	**113**	**89**	**80**	**89**	**80**	**45**	**83**	**59**	**48**
LAD II	313				1	15	39	36	**52**	90	80	70	66	32	32	**77**	41	24
LAD III	429	L2	5,001-10,000	Adv. 1	2	2	5	5	**9**	20	15	14	17	20	27	**20**	22	26
LAD IV	359	L1			1	2	6	5	**10**	19	15	14	8	5	4	**23**	10	13
LAD V	296	Other	10,001-15,000	Adv. 2	2	1	2	2	**4**	9	5	4	6	14	23	**10**	9	14
LAD VI	247				1	1	3	2	**5**	9	7	1	3	5	12	**13**	10	17
LAD VII	232		15,000-20,000	Super-adv.	2	0	1	1	**2**	4	4	2	1	7	8	**7**	10	11
LAD VIII	197				1	1	2	1	**3**	5	4	3	2	4	3	**10**	7	5

*TCE means the expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.

* LAD: Limited-academic-domain words

* F-JLPT: the former Japanese Language Proficiency Test

*WIS: Word Rankings for International Students

*# of HSD: Number of High-specificity Domains out of the 4 large academic domains

*Ha: Humanities & Arts

*Ss: Social Sciences

*Tn: Technological Natural Sciences

*Bn: Biological Natural Sciences

Table 7-22 Text Covering Efficiency (TCE) of Different Levels of Japanese Limited-academic-domain Words by Genre *Domain-specified

Corpus Code							MC	BSB	UPC	BCCWJ	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	BCCWJ-T	JS-Bn	JS-Tn
Genre							Conver-sation	Novels, Essays etc.	Essays, Novels etc.	Whole	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	Academic (Various)	Bn (Journal Articles)	Tn (Journal Articles)
Specified Domain							Tn	Ha	Ha	Ss	Ss	Ss	Ss	Ss	Bn	Tn	Ss	Bn	Tn
Total Tokens							1,129,538	2,298,828	2,102,178	32,819,424	5,675,357	186,768	50,601	42,152	13,904	74,645	2,895,425	719,802	2,705,026
LAD Label	F-JLPT Level	WIS	Level	# of HSD	# of Words for Ha	# of Words for Ss	# of Words for Tn	# of Words for Bn	TCE: Text Covering Efficiency = Expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.										
LAD 0	L3	682-1,291	Basic	2+	24	26	20	20	37	80	55	124	441	309	271	472	65	219	269	131	217
--				1+	13	6	5	9	53	125	91	72	230	142	63	127	583	38	91	78	97
--				2-	21	19	25	25	44	67	64	59	44	39	52	70	35	29	44	99	41
--				1-	20	27	28	24	57	82	62	110	136	111	91	194	33	34	96	55	53
LAD I		1,292-5,000	Inter.	2+	196	273	165	148	11	45	31	51	144	116	105	107	173	92	161	95	89
LAD II				1+	104	111	46	52	24	55	57	57	168	168	115	81	90	171	58	89	77
--				2-	195	118	226	243	16	22	19	39	43	26	21	48	23	10	26	38	18
--				1-	209	202	267	261	14	31	25	49	47	32	46	58	21	8	80	32	15
LAD III		5,001-10,000	Adv. 1	2+	201	274	216	167	2	5	6	8	26	20	20	24	50	53	28	41	46
LAD IV				1+	104	127	60	68	3	7	9	11	39	38	29	14	21	15	17	20	49
--				2-	228	155	213	262	1	4	4	9	9	5	2	5	1	2	11	10	5
--	L2			1-	255	232	299	291	2	5	4	10	8	3	5	5	2	2	24	8	6
LAD V	L1 Other	10,001-15,000	Adv. 2	2+	141	186	146	119	1	2	2	4	12	7	6	7	35	46	14	19	27
LAD VI				1+	71	74	48	54	1	3	3	6	21	22	3	7	21	59	10	22	68
--				2-	155	110	150	177	1	2	2	4	3	2	1	4	0	1	6	3	2
--				1-	176	173	199	193	1	3	2	5	3	1	1	2	0	1	14	6	5
LAD VII		15,000-20,000	Super-Adv.	2+	133	152	106	73	1	1	1	2	4	5	3	2	21	16	10	22	20
LAD VIII				1+	60	55	29	53	1	3	1	4	12	14	7	2	4	15	9	18	19
--				2-	99	80	126	159	0	1	1	2	2	1	0	0	0	1	3	4	3
--				1-	137	142	168	144	1	2	1	3	2	0	1	2	4	1	10	3	2
LAD I-VIII	L3+	682-20,000	All	2+/1+	1047	1284	841	763	5.9	19.7	16.5	22.9	68.4	57.6	46.9	48.2	70.4	61.2	54.5	47.4	54.5

*TCE means the expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.
* LAD: Limited-academic-domain words
* F-JLPT: the former Japanese Language Proficiency Test
*WIS: Word Rankings for International Students
*# of HSD: Number of High-specificity Domains out of the 4 large academic domains
*'+' and '-' in '# of HSD' mean that the words are specific in the domain or not. E.g., '2-' means the words are 2-domain words but not specific in the specified domain.
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

Table 7-21 clearly shows the superiority of LADs in gaining text coverage of academic texts; however, the superiority gets greater with domain-specific LADs shown as '2+' and '1+' in Table 7-22. Learning domain-specific LADs is 3-4 times more efficient than domain-unspecific (2- and 1-) words for basic and intermediate levels in gaining text coverage of academic texts and 7-12 times more efficient for advanced to super-advanced levels.

Table 7-23 Means and Standard Deviations for TCE of domain-specific (2+ and 1+) LADs in Academic and Non-academic Texts by Level

Level	Non-academic Texts		Academic Texts		Ratio for M (Aca/Non-aca.)
	M	SD	M	SD	
Basic	73.4	29.0	200.7	156.1	2.7
Inter.	37.0	27.0	112.0	136.6	3.0
Adv. 1	5.5	11.5	31.4	43.6	5.8
Adv. 2	2.1	22.8	24.9	44.7	11.7
S-Adv.	1.2	16.5	11.9	32.6	10.0

*Non-academic texts include MC, BSB and UPC. Academic texts include TB, TIS, all MTT and JS texts.

As indicated in Table 7-23, learning intermediate domain-specific (2+ and 1+) LADs is 3.0 times more efficient, Advanced 1 (6K-10K) gains 5.8 times; beyond 10K (Adv. 2 and S-Adv.) gains more than 10 times. This result suggests the importance of focused and specific purpose vocabulary learning and teaching at the advanced level.

Table 7-24 shows the overall comparison of TCE between common academic words (4-domain and 3-domain words) and LADs (2-domain and 1-domain words). For basic and intermediate levels, learning common academic words is more efficient in gaining text coverage of academic texts; however, at the Adv.1 (6K-10K) level, the highest TCE figure moves from 4-domain to 2 or 1-domain words, and at the levels beyond 10K, the peak moves to 1-domain words in most test corpora. This also suggests that focused vocabulary learning and teaching at the advanced levels is more efficient. (Note that TCE for 3-domain words is calculated by domain-unspecified analysis. If domain-specific analysis is applied to 3-domain words, TCE figures for 3-domain words will exceed 2-domain words in the intermediate level; but I did not do so as it is not realistic.)

Table 7-24 Text Covering Efficiency (TCE) of Different Levels of Japanese Limited-academic-domain Words by Genre

*Domain-specified only for LADs but not for JAWL

Corpus Label								MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
Text Genre								Conver-sation	Novels, Essays etc.	Essays, Novels etc.	Whole	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	Academic (Various)	Bn (Journal Articles)	Tn (Journal Articles)
Specified-domain for Domain-specified Analysis								Tn	Ha	Ha	**Ss**	Ss	Ss	Ss	Ss	Bn	Tn	**Ss**	Bn	Tn
Text Size (Total Tokens)								1,129,538	2,298,828	2,102,178	**32,819,424**	5,675,357	186,768	50,601	42,152	13,904	74,645	**2,895,425**	719,802	2,705,026
Label	# of Words for Ha	# of Words for Ss	# of Words for Tn	# of Words for Bn	F-JLPT Level	WIS	Level	# of HSD	TCE: Text Covering Efficiency = Expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.											
Basic (non-JAWL)		1,242			L4 L3	1-1,312		0	649	583	588	**552**	459	503	533	499	486	481	**471**	420	409
JAWL 0		31						4	187	405	382	**525**	667	888	966	773	882	1,178	**856**	1,099	1,069
		39					Basic	3	166	337	356	**354**	475	604	569	611	422	460	**493**	425	384
LAD 0	24	26	20	20	L3	682-1,291		2	37	80	55	**124**	441	309	271	472	65	219	**207**	131	217
	13	6	5	9				1	53	125	91	**72**	230	142	63	127	583	38	**119**	78	97
JAWL I		559						**4**	**14**	**47**	**56**	**82**	**156**	**174**	**175**	**182**	**198**	**198**	**199**	**241**	**271**
JAWL II		541						3	10	29	27	**48**	122	93	89	86	78	53	**104**	96	84
LAD I	**196**	**273**	**165**	**148**		1,292-5,000	**Inter.**	**2**	**11**	**45**	**31**	**51**	**144**	**116**	**105**	**107**	**173**	**92**	**97**	**95**	**89**
LAD II	104	111	46	52				1	24	55	57	**57**	168	168	115	81	90	171	**126**	89	77
JAWL III		212						4	1	4	5	**8**	15	17	18	11	21	32	**24**	41	54
JAWL IV		451				5,001-10,000	Adv. 1	3	1	4	5	**8**	17	16	10	13	47	42	**20**	27	27
LAD III	201	274	216	167	L2			2	2	5	6	**8**	26	20	20	24	50	53	**22**	41	46
LAD IV	104	127	60	68	L1 Other			1	3	7	9	**11**	39	38	29	14	21	15	**38**	20	49
JAWL V		103						4	1	2	2	**3**	5	5	5	6	6	8	**10**	23	18
JAWL VI		327				10,001-15,000	Adv. 2	3	1	2	2	**3**	8	6	4	5	9	19	**9**	13	13
LAD V	141	186	146	119				2	1	2	2	**4**	12	7	6	7	35	46	**11**	19	27
LAD VI	71	74	48	54				1	1	3	3	**6**	21	22	3	7	21	59	**21**	22	68
JAWL VII		56						4	0	1	1	**2**	3	4	4	7	8	18	**6**	10	21
JAWL VIII		268				15,000-20,000	Super-adv.	3	0	1	1	**2**	4	3	3	2	2	14	**5**	8	11
LAD VII	133	152	106	73				2	1	1	1	**2**	4	5	3	2	21	16	**7**	22	20
LAD VIII	60	55	29	53				1	1	3	1	**4**	12	14	7	2	4	15	**19**	18	19

*TCE means the expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.
* JAWL: Japanese Acaademic Word List
* LAD: Limited-academic-domain words
* F-JLPT: the former Japanese Language Proficiency Test
*WIS: Word Rankings for International Students
*# of HSD: Number of High-specificity Domains out of the 4 large academic domains
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

It is also clear that LADs are also good for reading newspapers. The same thing is also true of common academic words (JAWL vocabulary); however, LADs seem more useful at the intermediate level and above.

In sum, LADs are useful words for reading academic texts and newspapers as well as common academic words. Different levels of LADs are 3 to 12 times more useful in reading academic texts than non-academic texts. The relative importance of LADs is higher at the advanced level or above than basic and intermediate levels.

7.4.4 The usefulness of Japanese literary words

Table 7-25 and 7-26 show text coverage and Text Covering Efficiency (TCE) respectively for literary words in different corpora. Text coverage by the intermediate literary words (446 words) in literary texts (BSB and UPC) is around 2.8%, which is much lower than the coverage by the 4-domain intermediate common academic words (JAWL Ⅰ, 559 words) in academic texts at 9.7-15.1%. Nevertheless, the distribution pattern is clearly the opposite to the academic vocabulary, that is, high in literary texts (non-academic texts) but low in academic texts and newspapers.

To confirm the high text coverage of literary texts by literary words, a Chi-square test (test of independence) was conducted on the tokens of literary words and the other words between the two literary test corpora (BSB and UPC) and eight non-literary test corpora (TB, MTT-Ss, TIS, MTT-Bn, MTT-Tn, JS-Bn and JS-Tn) (The conversation corpus was not used). The results are all significant for each of the five levels and overall (χ^2 = 13421304.09, df = 2, $p < .001$) to prove the distributions of literary words are not the same as the other words across the literary and non-literary texts.

Table 7-25 Text Coverage in Different Genres by Different Levels of Japanese Literary Words (LWs)

	Corpus Code	MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
	Genre	Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
	Specified Domain	Tn	Ha	Ha	**Ss**	Ss	Ss	Ss	Ss	Bn	Tn	**Ss**	Bn	Tn
	Total Tokens	1,129,538	2,298,828	2,102,178	**32,819,424**	5,675,357	186,768	50,601	42,152	13,904	74,645	**2,895,425**	719,802	2,705,026
LW Label	F-JLPT Level / WIS # of Words	Text Coverage (%)												
Basic LW	L3 682-1,291 142	6.73	2.86	3.52	**2.12**	0.78	1.05	1.13	1.25	0.97	1.33	**1.02**	0.63	0.66
Inter. LW	L2 1,292-5,000 446	3.23	2.76	2.83	**1.87**	0.74	0.52	0.39	0.66	0.29	0.81	**0.62**	0.42	0.40
Adv. 1 LW	L2 5,001-10,000 483	0.75	0.64	0.73	**0.41**	0.15	0.06	0.03	0.08	0.16	0.12	**0.08**	0.05	0.07
Adv. 2 LW	L1 10,001-15,000 345	0.07	0.21	0.23	**0.14**	0.02	0.01	0.00	0.01	0.01	0.02	**0.02**	0.02	0.01
Super-adv. LW	Other 15,000-20,000 200	0.01	0.08	0.09	**0.05**	0.01	0.00	0.00	0.01	0.00	0.00	**0.00**	0.02	0.00
Total	1,616	10.79	6.55	7.40	**4.59**	1.70	1.64	1.55	2.01	1.43	2.27	**1.75**	1.14	1.14

*LW: Literary Words
*F-JLPT: The former Japanese Language Proficiency Test
*WIS: Word Rankings for International Students
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

Table 7-26 Text Covering Efficiency (TCE) of Different Levels of Japanese Literary Words (LWs) by Genre

	Corpus Code	MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
	Genre	Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
	Specified Domain	Tn	Ha	Ha	**Ss**	Ss	Ss	Ss	Ss	Bn	Tn	**Ss**	Bn	Tn
	Total Tokens	1,129,538	2,298,828	2,102,178	**32,819,424**	5,675,357	186,768	50,601	42,152	13,904	74,645	**2,895,425**	719,802	2,705,026
LW Label	F-JLPT Level / WIS # of Words	TCE: Text Covering Efficiency = Expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.												
Basic LW	L3 682-1,291 142	474	201	248	**149**	55	74	80	88	68	93	**72**	44	46
Inter. LW	L2 1,292-5,000 446	72	62	63	**42**	17	12	9	15	7	18	**14**	9	9
Adv. 1 LW	L2 5,001-10,000 483	15	13	15	**9**	3	1	1	2	3	3	**2**	1	2
Adv. 2 LW	L1 10,001-15,000 345	2	6	7	**4**	1	0	0	0	0	0	**0**	1	0
Super-adv. LW	Other 15,000-20,000 200	1	4	4	**3**	0	0	0	1	0	0	**0**	1	0
Total	1,616	66.8	41	46	**28**	11	10	10	12	9	14	**11**	7	7

*LW: Literary Words
*F-JLPT: The former Japanese Language Proficiency Test
*WIS: Word Rankings for International Students
*TCE means the expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

Literary words provide a similar level of text coverage and TCE for conversation from basic to Adv. 1 (01K to 10K) level; however, beyond 10K (Adv.2 and S-Adv.), coverage and TCE in conversation texts is not as high as in literary texts. In this sense, literary words beyond 10K (e.g. 瞬き 'matataki' (blink), にやり 'niyari' (snigger)) have truly distinctive lexical features with literary works. When the literary words were extracted, no conversation corpus could be used as a reference corpus. If we add a common conversation corpus as a reference corpus, the number of literary words and its text coverage will be smaller, while the TCE figure is expected to be higher. The current literary words are very different from other types of written texts; however, it is also close to daily conversation words up to 10K. It may be better to exclude conversation words from literary words; however, considering the fact that learners cannot always write spoken words in Kanji, it may also be good to keep the spoken words in the literary words as they are.

Table 7-27 Means and Standard Deviations for TCE of Literary Words in Literary and Non-literary Texts by Level

Level	Literary Texts		Non-literary Texts		Ratio for M (Lit./Non-lit.)
	M	SD	M	SD	
Basic	224.72	23.25	68.62	17.4	3.3
Inter.	62.71	0.75	11.87	3.9	5.3
Adv. 1	14.18	1.02	1.88	0.9	7.6
Adv. 2	6.39	0.26	0.34	0.2	19.0
S-Adv.	4.17	0.31	0.25	0.3	16.5

*Literary texts include BSB and UPC. Non-literary texts include UYN, TB, TIS, all MTT and JS texts.

The average TCE figures of literary words from intermediate to advanced levels, which range from 4.2 to 62.7 (Table 7-27), are around half of the figures for common academic words (AWs) and limited-academic-domain words (LADs), which range from 8.1 to 144.2 (Table 7-18 and 7-23). This result suggests that literary texts are more diverse in vocabulary. Or the figures may be improved if we extract domain-specific words after dividing literary texts into different genres such as detective stories, romances etc. However, the average TCE figures of literary words are still much higher in literary texts than in non-literary texts. The figures range from 5.3 to 19.0, which are higher than the figures of

academic vocabulary (AWs and LADs), particularly at the advanced level or above. This suggests that focused vocabulary learning is also very useful for reading literary works.

In sum, literary words do not provide as high coverage and efficiency for reading literary works as academic vocabulary for academic texts; however, they are still useful words for reading literary texts.

7.4.5 Word tier analysis of text genres in Japanese: Answering the main research questions for this thesis

I have tested text coverage and Text Covering Efficiency with the extracted domain-specific words in different genres. However, there is one more question to examine; Is Text Covering Efficiency with these words higher than 'non-specific (general) words' at the same frequency level? (I compared domain-specific LADs with domain-unspecific LADs but did not compare with non-academic vocabulary.)

Also, I have checked what different genres and levels the groups of domain-specific words are positioned group by group; however, if these different aspects are combined together, what kind of features are found with those text genres?

In sum, what is the most efficient learning order of words according to the main working genre of a learner? This is the main research question for this whole thesis. To answer this question, I propose an analysis entitled 'word tier analysis' by which text coverage and Text Covering Efficiency with different groups of words in different text genres at different frequency levels are analysed together. Using word tier analysis, I will show different lexical features with different text genres. Proportions of word origins with different groups of words are also calculated and discussed together.

7.4.5.1 Method

I developed a 'word tier analyser' which is an Excel sheet (see accompanying CD)

where word profiling (text coverage) and Text Covering Efficiency (TCE) with groups of words in a text can be checked automatically by cutting and pasting the result of the number of word tokens counted by AntWordProfiler with the 'word tier baseword lists'. Using this analyser, the ranking of groups of words by TCE in a genre will also be automatically provided. This analysis can be either domain-specified or domain-unspecified for 3-domain, 2-domain and 1-domain words; however, I just conducted the domain-unspecified analysis here because the domain-specified analysis will be too complicated and confusing for some texts with highly mixed text genres.

Word origins are calculated by group for all frequency levels together.

7.4.5.2 Result and discussion

7.4.5.2.1 Features of word tiers

Table 7-28 Text Covering Efficiency (TCE) of the Grouped Words by Genre (Not Graded by Level) *Domain-unspecified

		Corpus Code	MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
		Genre	Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
		Total Tokens (Million)	1.13	2.30	2.10	**32.82**	5.68	0.19	0.05	0.04	0.01	0.07	**2.90**	0.72	2.71
WIS	F-JLPT Level	Label	Number of Lexemes in VDRJ	TCE: Text Covering Efficiency = Expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.											
1-20,000	L4-L1, Others	General	13,302	61	59	58	**56**	48	50	51	50	46	46	**46**	41	40
682-20,000	L3-L1, Others	AW	2,591	10	28	29	**42**	80	82	81	80	88	89	**90**	103	108
		LAD	2,542	6	15	12	**21**	44	35	30	35	27	23	**36**	26	24
		LW	1,616	67	41	46	**28**	11	10	10	12	9	14	**11**	7	7
20,001+	L2, L1, Others	21K+	91,104	0.1	0.2	0.2	**0.2**	0.2	0.1	0.1	0.1	0.4	0.4	**0.3**	0.3	0.5
--		AKW	30,821	0.6	0.8	0.4	**0.6**	0.4	0.1	0.1	0.3	0.1	0.2	**0.4**	0.2	0.1
1-5,000	L4-L1, Others	1K-05K	5,024	184	178	177	**177**	177	183	187	183	171	168	**177**	163	159
1-10,000	L4-L1, Others	1K-10K	10,024	95	93	93	**92**	94	96	96	96	90	89	**93**	86	85

*WIS: Word Rankings for International Students
*F-JLPT: The former Japanese Language Proficiency Test
*VDRJ: Vocabulary Database for Reading Japanese
*AW: Common Academic Words
*LAD: Limited-academic-domain words
*LW: Literary Words
*AKW: Assumed Known Words (mostly proper nouns)
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

Table 7-28 shows Text Covering Efficiency (TCE) of the grouped words by genre (domain-unspecified). Table 7-29 shows the ranking for Text Covering Efficiency (TCE) of the grouped words in each genre (domain-unspecified). These are based on the simplest

classification not graded by level but only by the four groups (common academic words, limited-academic-domain words, literary words and the others (general)) except for the low frequency words beyond 20K and Assumed Known Words (mostly proper nouns).

As shown in Table 7-28, learning common academic words (AWs) is twice as efficient as learning general words (GWs) in covering academic texts and newspapers but not non-academic texts. For example, the ratios of TCE (AW:GW) are 82:50 for TB (social science texts), 108:40 for JS-Tn (journal articles in technological natural sciences) and 80:48 for UYN (newspaper texts). This gap is much larger at the intermediate level or above. The average TCE of common academic words (JAWL Ⅰ and II) and non-common-academic words for academic texts is 145 and 16 respectively at the intermediate level (calculated from the figures in Table 7-30). Intermediate common academic words (JAWL Ⅰ and II) are 9 times as useful as general words for reading academic texts. The ratios of the two are 7, 6 and 8 times at the 6K-10K, 11K-15K and 16K-20K levels respectively. General words are as important as common academic words only at the basic level.

Table 7-28 also shows that domain-non-specified limited-academic-domain words are at the same level as general words on average if the words at different levels are calculated together. Literary words only have one-eighth the value of common academic words for academic texts and newspapers.

Table 7-29 Ranking for Text Covering Efficiency (TCE) of the Grouped Words in Each Genre (Not Graded by Level) *Domain-unspecified

		Corpus Code	MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
		Genre	Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
		Total Tokens (Million)	1.13	2.30	2.10	**32.82**	5.68	0.19	0.05	0.04	0.01	0.07	**2.90**	0.72	2.71
WIS	F-JLPT Level	Label	Number of Lexemes in VDRJ	Ranking for TCE of the Grouped Words in Each Genre											
1-20,000	L4-L1, Others	General	13,302	2	1	1	**1**	2	2	2	2	2	**2**	2	2
682-20,000	L3-L1, Others	AW	2,591	3	3	3	**2**	1	1	1	1	1	**1**	1	1
		LAD	2,542	4	4	4	**4**	3	3	3	3	3	**3**	3	3
		LW	1,616	1	2	2	**3**	4	4	4	4	4	**4**	4	4
20,001+	L2, L1, Others	21K+	91,104	6	6	6	**6**	6	6	6	5	5	**6**	5	5
--		AKW	30,821	5	5	5	**5**	5	5	5	6	6	**5**	6	6

*WIS: Word Rankings for International Students
*F-JLPT: The former Japanese Language Proficiency Test
*VDRJ: Vocabulary Database for Reading Japanese
*AW: Common Academic Words
*LAD: Limited-academic-domain words
*LW: Literary Words
*AKW: Assumed Known Words (mostly proper nouns)
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

Table 7-30 Text Covering Efficiency (TCE) of the Grouped Words by Level and Genre *Domain-unspecified

			Corpus Code	MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
			Genre	Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
			Total Tokens (Million)	1.13	2.30	2.10	**32.82**	5.68	0.19	0.05	0.04	0.01	0.07	**2.90**	0.72	2.71
WIS	F-JLPT Level	Level	Label	Number of Lexemes in VDRJ	TCE: Text Covering Efficiency = Expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.											
1-1,291	L4, L3	Basic	General	1,027	716.0	671.8	672.8	**640.3**	530.6	585.8	623.0	572.9	571.1	564.0	**551.1**	495.6	481.2
682-1,291	L3		AW	70	175.1	367.1	367.7	**430.2**	560.0	729.3	744.8	682.6	625.7	778.0	**654.1**	723.4	687.5
			LAD	78	47.6	84.5	65.0	**99.4**	222.6	162.0	139.1	251.2	105.1	80.0	**123.3**	91.2	93.9
			LW	142	474.1	201.5	248.0	**149.1**	55.0	74.1	79.6	87.9	68.4	93.5	**72.1**	44.1	46.3
1,292-5,000	L2	Inter.	General	1,478	35.0	32.8	27.4	**31.8**	33.4	21.6	14.1	27.5	10.4	10.3	**17.7**	13.9	10.6
			AW	1,101	11.8	38.3	41.8	**65.3**	138.9	134.1	132.6	134.6	138.8	127.0	**152.3**	169.3	178.8
			LAD	704	14.4	35.8	29.7	**49.3**	102.9	85.0	75.5	78.9	58.6	38.9	**80.2**	51.4	37.6
			LW	446	72.4	62.0	63.5	**41.8**	16.7	11.7	8.6	14.9	6.6	18.1	**13.9**	9.5	8.9
5,001-10,000		Adv. 1	General	3,070	4.4	7.2	7.1	**7.4**	7.7	4.9	2.4	6.8	2.2	2.7	**3.7**	3.2	2.9
			AW	664	1.4	4.0	4.9	**7.8**	16.5	16.1	12.5	12.1	38.9	38.6	**21.1**	31.8	35.9
			LAD	788	2.2	5.2	5.2	**9.3**	19.7	15.0	13.7	13.1	13.5	17.0	**21.0**	16.6	19.9
			LW	483	15.5	13.2	15.2	**8.5**	3.1	1.2	0.7	1.6	3.3	2.6	**1.7**	1.1	1.5
10,001-15,000	L1 Others	Adv. 2	General	3,681	1.6	3.2	2.9	**3.3**	3.0	2.1	0.8	2.0	1.4	1.4	**1.5**	2.0	1.5
			AW	431	0.5	1.7	1.8	**3.3**	7.0	6.0	3.9	5.1	8.5	16.4	**9.3**	15.6	14.5
			LAD	543	1.1	2.5	2.0	**4.4**	8.6	6.0	3.1	4.9	9.8	18.2	**11.6**	9.7	15.5
			LW	345	2.1	6.1	6.7	**4.1**	0.6	0.2	0.1	0.3	0.2	0.5	**0.4**	0.6	0.3
15,001-20,000		S-Adv.	General	4,046	0.8	1.9	1.9	**1.8**	1.5	0.9	0.4	1.0	0.6	0.7	**0.8**	1.2	0.9
			AW	325	0.4	1.2	1.2	**1.9**	3.7	3.0	2.7	2.9	2.9	14.9	**5.5**	8.0	12.5
			LAD	429	0.6	1.5	1.1	**2.6**	4.3	3.9	2.1	1.6	5.5	5.6	**8.2**	8.3	8.0
			LW	200	0.7	3.9	4.5	**2.6**	0.3	0.1	0.1	0.6	0.0	0.1	**0.2**	0.8	0.1
20,001+		21K+	21K+	91,104	0.1	0.2	0.2	**0.2**	0.2	0.1	0.1	0.1	0.4	0.4	**0.3**	0.3	0.5
--		AKW	AKW	30,821	0.6	0.8	0.4	**0.6**	0.4	0.1	0.1	0.3	0.1	0.2	**0.4**	0.2	0.1
1-5,000	L4-L1, Others	1K-05K	1K-05K	5,024	184.2	177.7	177.4	**176.7**	177.3	183.2	186.6	182.9	171.1	167.8	**176.6**	163.1	159.0
1-10,000	L4-L1, Others	1K-10K	1K-10K	10,024	94.7	92.6	92.5	**92.5**	94.0	95.6	96.2	95.6	90.2	88.9	**92.8**	86.2	84.6

*WIS: Word Rankings for International Students
*F-JLPT: The former Japanese Language Proficiency Test
*VDRJ: Vocabulary Database for Reading Japanese
*AW: Common Academic Words
*LAD: Limited-academic-domain words
*LW: Literary Words
*AKW: Assumed Known Words (mostly proper nouns)
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

On the other hand, literary words (LWs) are an efficient source for covering non-academic text. Interestingly, literary words provide the highest TCE for conversation but

not for literary works (Table 7-28 and 7-29) especially from the intermediate to 10K level; however, general words (GWs) and literary words are on average at the same level for the three non-academic test corpora (Table 7-28). Checking TCE by level, for reading literary works, general words are more important at the basic level; however, at the intermediate level or above, literary words are consistently twice as useful as general words (Table 7-30). Compared to these words, common academic words and limited-academic-domain words are less than half as valuable for non-academic texts.

It is also clear that natural science texts contain more low frequency words beyond the top 20,000 word (21K+) level. These words are not very high in ratio at around 0.4 TCE; however, many of these words will be technical terms which are essential for understanding the texts. The fact that natural science texts contain more low frequency words is more clearly shown in Table 7-30 and 7-31 where each group of words is graded into five levels. TCE figures of academic vocabulary (AWs and LADs) are greater than 5 even beyond 10K in academic texts, especially, they are high in journal articles at 8.0-15.6. Also, TCE of the top 5,000 and 10,000 words are also shown at the bottom of the Table 7-30. The figures tend to be low in natural science texts. This also proves the inclination to high-frequency words in natural science texts. As mentioned about common academic words in 7.4.2.2, the fact that natural science texts contain more low frequency words is seemingly common in other languages, whether in high school texts (Coxhead, Stevens, & Tinkle, 2010) or in highly technical journal articles.

Comparing the TCE figures in Table 7-30 with the figures in Table 7-14, we can see what ranking of a general word a domain-specific word is equivalent to. Table 7-30 shows that the TCE figures of common academic words at Adv. 2 level (10-15K) for journal articles are 15.6 and 14.6, which mean the common academic words at this level are as useful as general intermediate words in general texts.

Table 7-31 Ranking for Text Covering Efficiency (TCE) of the Grouped Words in Each Genre *Domain-unspecified

			Corpus Code	MC	BSB	UPC	BCCWJ	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	BCCWJ-T	JS-Bn	JS-Tn
			Genre	Conver-sation	Novels, Essays etc.	Essays, Novels etc.	Whole	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	Academic (Various)	Bn (Journal Articles)	Tn (Journal Articles)
			Total Tokens (Million)	1.13	2.30	2.10	32.82	5.68	0.19	0.05	0.04	0.01	0.07	2.90	0.72	2.71
WIS	F-JLPT Level	Level	Label	Number of Lexemes in VDRJ	Ranking for TCE of the Grouped Words in Each Genre											
1-1,291	L4, L3		General	1,027	1	1	1	1	2	2	2	2	2	2	2	2	2
682-1,291	L3	Basic	AW	70	2	2	2	2	1	1	1	1	1	1	1	1	1
			LAD	78	***5***	4	4	4	3	3	3	3	4	***5***	4	4	4
			LW	142	1	3	3	3	***6***	***6***	***5***	***5***	***5***	4	***6***	***6***	***5***
1,292-5,000		Inter.	General	1,478	6	8	8	8	7	7	7	7	***9***	***13***	***9***	***10***	***12***
			AW	1,101	***9***	6	6	5	**4**	**4**	**4**	**4**	**3**	**3**	**3**	**3**	**3**
			LAD	704	8	7	7	6	5	5	6	6	6	6	5	5	6
			LW	446	**4**	5	5	7	***9***	***10***	***10***	8	***12***	9	***10***	***12***	***13***
5,001-10,000		Adv. 1	General	3,070	10	10	10	12	12	***13***	***14***	11	***16***	***15***	***15***	***15***	***15***
			AW	664	***14***	***13***	***13***	11	10	**8**	9	10	**7**	**7**	**7**	**7**	**7**
			LAD	788	11	12	12	9	**8**	9	**8**	9	**8**	10	**8**	**8**	**8**
	L2		LW	483	**7**	9	9	10	***16***	***17***	***17***	***16***	***14***	***16***	***16***	***18***	***16***
10,001-15,000	L1 Others	Adv. 2	General	3,681	13	15	15	16	***17***	16	16	15	***17***	***17***	***17***	16	***17***
			AW	431	***20***	***18***	***18***	15	13	**12**	**11**	**12**	**11**	**11**	**12**	**9**	**10**
			LAD	543	15	16	16	13	**11**	**11**	**12**	13	**10**	**8**	**11**	**11**	**9**
			LW	345	**12**	**11**	**11**	14	***19***	***19***	***22***	***21***	***20***	***19***	***19***	***20***	***20***
15,001-20,000		S-Adv.	General	4,046	**16**	17	17	20	18	18	18	18	18	18	18	17	18
			AW	325	***21***	20	19	19	**15**	**15**	**13**	**14**	**15**	**12**	**14**	**14**	**11**
			LAD	429	19	19	20	17	**14**	**14**	**15**	17	**13**	**14**	**13**	**13**	**14**
			LW	200	17	**14**	**14**	18	***21***	***22***	19	19	***22***	***22***	***22***	19	***22***
20,000+		21K+	21K+	91,104	22	22	22	22	22	21	21	22	**19**	**20**	21	21	**19**
--		AKW	AKW	30,821	**18**	21	21	21	**20**	**20**	**20**	**20**	21	21	**20**	22	21

*TCE means the expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.

*WIS: Word Rankings for International Students

*F-JLPT: The former Japanese Language Proficiency Test

*VDRJ: Vocabulary Database for Reading Japanese

*AW: Common Academic Words

*LAD: Limited-academic-domain words

*LW: Literary Words

*AKW: Assumed Known Words (mostly proper nouns)

*Ha: Humanities & Arts

*Ss: Social Sciences

*Tn: Technological Natural Sciences

*Bn: Biological Natural Sciences

*Numbers in bold show the rankings higher than expected ranking i.e. 1-4 for basic, 5-8 for intermediate, 9-12 for Adv. 1, 13-16 for Adv. 2, 17-20 for S-Adv and 21-22 for 21K+ and AKW. On the other hand, Italic numbers show the rankings lower than expected ranking.

In Table 7-31, numbers in bold show the rankings higher than the expected ranking i.e. 1-4 for basic, 5-8 for intermediate, 9-12 for Adv. 1, 13-16 for Adv. 2, 17-20 for S-Adv. and 21-22 for 21K+ and AKW. On the other hand, italic numbers show the rankings lower than the expected ranking. These bold and italic figures show the relative importance which is not expected from the frequency rankings. Domain-specificity shown by these figures is much clearer in academic texts. Academic vocabulary (AWs and LADs) are very useful at all levels in academic texts while literary words are not useful.

For non-academic texts, this tendency is not clearly shown. Literary words are somewhat more useful for reading literary texts; however, it is not as clear as academic vocabulary for academic texts. Learning words by following the (adjusted) frequency ranking (Word Rankings for International Students or maybe other rankings introduced in

Chapter 3) may be efficient.

As it should be, the TCE rankings in BCCWJ, the corpus used for creating the word rankings and extracting domain-specific words, are all within the expected range (no bold or italic figures).

As mentioned in 7.4.3, newspaper texts are similar to academic texts, but contain more academic vocabulary (AWs and LADs) at the advanced level. Newspapers can be a good resource for learning common academic words and limited-academic-domain words for social sciences (See also Table 7-32 and 7-33).

In sum, general words are important for any genre at the basic level. Academic vocabulary is 6-9 times as useful as general words for reading academic texts and newspapers at the intermediate level or above. Literary words are twice as useful as general words for reading literary works at the intermediate level or above. Natural science texts contain more low-frequency words than other domains. Domain-specificity is stronger in academic texts than in literary texts.

(From here down blank.)

Table 7-32 Text Covering Efficiency (TCE) of the Grouped Words by Level and Genre (Detailed) *Domain-unspecified

			Corpus Code	MC	BSB	UPC	**BCCWJ**	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	**BCCWJ-T**	JS-Bn	JS-Tn
			Genre	Conver-sation	Novels, Essays etc.	Essays, Novels etc.	**Whole**	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	**Academic (Various)**	Bn (Journal Articles)	Tn (Journal Articles)
			Total Tokens (Million)	1.13	2.30	2.10	**32.82**	5.68	0.19	0.05	0.04	0.01	0.07	**2.90**	0.72	2.71
WIS	F-JLPT Level	Label 1	Label 2	Number of Lexemes in VDRJ	TCE: Text Covering Efficiency = Expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.											
1-1,291	L4, L3	General	Basic	1,027	716.0	671.8	672.8	**640.3**	530.6	585.8	623.0	572.9	571.1	564.0	**551.1**	495.6	481.2
682-1,291	L3	AW	Basic+Aca4D	31	186.9	405.0	382.0	**525.5**	667.0	887.6	965.8	772.9	881.6	1177.6	**856.2**	1098.7	1069.0
			Basic+Aca3D	39	165.7	337.0	356.3	**354.5**	474.8	603.5	569.1	610.7	422.3	460.3	**493.5**	425.1	384.3
		LAD	Basic+Aca2D	45	41.0	73.9	58.8	**96.5**	273.6	195.0	178.3	302.1	47.9	113.4	**143.8**	113.1	119.2
			Basic+Aca1D_Ah	13	43.4	124.5	91.4	**122.7**	195.8	159.4	153.5	366.8	55.3	55.6	**102.4**	35.1	67.1
			Basic+Aca1D_Ss	6	20.7	47.3	38.5	**72.4**	229.5	141.9	62.6	126.5	12.0	4.5	**119.4**	67.6	20.4
			Basic+Aca1D_Tn	5	52.8	73.0	60.7	**92.0**	97.1	51.4	39.5	33.2	0.0	37.5	**90.6**	92.2	97.4
			Basic+Aca1D_Bn	9	101.8	110.9	78.1	**102.5**	70.9	75.6	28.5	34.3	583.4	22.3	**71.9**	77.8	53.1
		LW	Basic+Lit	142	474.1	201.5	248.0	**149.1**	55.0	74.1	79.6	87.9	68.4	93.5	**72.1**	44.1	46.3
1,292-5,000		General	Inter	1,478	35.0	32.8	27.4	**31.8**	33.4	21.6	14.1	27.5	10.4	10.3	**17.7**	13.9	10.6
		AW	Inter+Aca4D	559	13.8	47.5	56.2	**81.8**	155.7	173.8	174.8	181.7	197.8	198.4	**198.9**	241.1	271.0
			Inter+Aca3D	542	9.8	28.8	26.9	**48.4**	121.5	93.1	89.0	86.0	78.0	53.4	**104.2**	95.4	83.7
		LAD	Inter+Aca2D	391	13.7	33.3	24.9	**47.4**	113.2	89.0	79.7	89.3	79.8	44.7	**82.7**	59.4	48.2
			Inter+Aca1D_Ah	104	11.2	54.6	56.7	**54.3**	48.8	47.0	72.2	78.9	25.6	9.8	**49.8**	30.6	18.5
			Inter+Aca1D_Ss	111	16.5	31.5	26.4	**57.4**	168.1	167.6	114.7	81.0	4.5	4.1	**125.8**	30.2	15.3
			Inter+Aca1D_Tn	46	24.0	24.5	22.7	**43.4**	39.6	23.3	9.0	18.1	48.5	171.0	**58.0**	39.4	77.0
			Inter+Aca1D_Bn	52	12.9	35.7	24.8	**41.6**	50.4	9.9	25.1	49.7	89.9	10.8	**44.5**	89.1	9.1
		LW	Inter+Lit	446	72.4	62.0	63.5	**41.8**	16.7	11.7	8.6	14.9	6.6	18.1	**13.9**	9.5	8.9
5,001-10,000	L2	General	Adv.1	3,070	4.4	7.2	7.1	**7.4**	7.7	4.9	2.4	6.8	2.2	2.7	**3.7**	3.2	2.9
		AW	Adv.1+Aca4D	212	1.2	3.5	5.5	**7.8**	15.2	17.4	17.9	11.0	21.0	32.2	**23.5**	41.4	54.5
			Adv.1+Aca3D	452	1.4	4.3	4.7	**7.8**	17.1	15.5	9.9	12.6	47.3	41.6	**20.0**	27.3	27.2
		LAD	Adv.1+Aca2D	429	1.9	4.7	5.0	**8.6**	20.1	14.6	13.8	17.0	20.3	27.5	**19.5**	21.7	25.7
			Adv.1+Aca1D_Ah	104	2.9	7.1	8.6	**9.1**	6.7	4.4	3.6	5.9	4.1	2.1	**12.2**	2.8	9.0
			Adv.1+Aca1D_Ss	127	1.7	5.1	2.9	**11.1**	39.1	37.6	28.6	14.4	0.0	1.6	**38.0**	5.3	4.7
			Adv.1+Aca1D_Tn	60	3.2	3.9	4.7	**9.6**	7.1	2.7	2.0	3.6	1.2	14.5	**16.5**	23.3	49.3
			Adv.1+Aca1D_Bn	68	2.7	6.5	5.9	**10.6**	12.3	2.0	11.3	5.6	21.2	4.5	**16.1**	20.3	2.2
		LW	Adv.1+Lit	483	15.5	13.2	15.2	**8.5**	3.1	1.2	0.7	1.6	3.3	2.6	**1.7**	1.1	1.5
10,001-15,000	L1 Others	General	Adv.2	3,681	1.6	3.2	2.9	**3.3**	3.0	2.1	0.8	2.0	1.4	1.4	**1.5**	2.0	1.5
		AW	Adv.2+Aca4D	103	0.5	1.5	2.0	**3.2**	5.0	5.0	4.8	5.5	5.6	7.7	**10.1**	22.6	18.4
			Adv.2+Aca3D	328	0.5	1.7	1.8	**3.3**	7.6	6.3	3.7	4.9	9.4	19.2	**9.0**	13.4	13.3
		LAD	Adv.2+Aca2D	296	0.9	2.0	1.9	**3.9**	8.6	5.0	4.5	6.3	14.1	23.2	**10.2**	9.5	14.3
			Adv.2+Aca1D_Ah	71	1.1	3.4	3.4	**4.5**	1.7	1.0	1.1	1.7	0.0	0.8	**9.6**	2.2	9.9
			Adv.2+Aca1D_Ss	74	1.5	2.9	1.3	**5.5**	20.9	21.8	3.2	7.1	0.0	1.8	**21.2**	3.7	1.0
			Adv.2+Aca1D_Tn	48	0.8	1.6	2.3	**4.6**	3.4	1.3	0.0	1.0	0.0	58.6	**10.3**	17.0	67.9
			Adv.2+Aca1D_Bn	54	1.8	3.8	1.6	**5.1**	5.7	1.0	0.4	2.2	21.3	0.2	**10.5**	22.0	2.9
		LW	Adv.2+Lit	345	2.1	6.1	6.7	**4.1**	0.6	0.2	0.1	0.3	0.2	0.5	**0.4**	0.6	0.3
15,001-20,000		General	S-Adv	4,046	0.8	1.9	1.9	**1.8**	1.5	0.9	0.4	1.0	0.6	0.7	**0.8**	1.2	0.9
		AW	S-Adv+Aca4D	56	0.5	0.9	0.8	**1.9**	2.6	3.8	3.5	7.2	7.7	17.9	**6.4**	10.3	21.3
			S-Adv+Aca3D	269	0.4	1.2	1.3	**1.9**	3.9	2.8	2.5	2.0	1.9	14.3	**5.4**	7.6	10.7
		LAD	S-Adv+Aca2D	232	0.5	1.2	1.1	**2.2**	3.5	3.8	1.8	1.4	6.8	7.5	**6.5**	9.6	10.7
			S-Adv+Aca1D_Ah	60	0.8	2.6	0.8	**2.4**	2.5	0.5	0.7	1.6	0.0	0.0	**5.3**	1.9	2.9
			S-Adv+Aca1D_Ss	55	0.6	1.4	0.8	**3.6**	12.2	13.5	7.2	1.7	0.0	0.0	**18.7**	0.8	1.7
			S-Adv+Aca1D_Tn	29	0.6	1.4	1.7	**3.1**	1.4	0.4	0.7	0.0	19.8	14.8	**9.0**	7.9	19.4
			S-Adv+Aca1D_Bn	53	0.7	1.8	1.0	**3.1**	2.9	0.3	0.7	3.1	4.1	4.0	**7.6**	17.6	2.4
		LW	S-Adv+Lit	200	0.7	3.9	4.5	**2.6**	0.3	0.1	0.1	0.6	0.0	0.1	**0.2**	0.8	0.1
20,001+		21K+	21K+	91,104	0.1	0.2	0.2	**0.2**	0.2	0.1	0.1	0.1	0.4	0.4	**0.3**	0.3	0.5
--		AKW	AKW	30,821	0.6	0.8	0.4	**0.6**	0.4	0.1	0.1	0.3	0.1	0.2	**0.4**	0.2	0.1
1-5,000	L4-L1, Others	1K-05K	1K-05K	5,024	184.2	177.7	177.4	**176.7**	177.3	183.2	186.6	182.9	171.1	167.8	**176.6**	163.1	159.0
1-10,000	L4-L1, Others	1K-10K	1K-10K	10,024	94.7	92.6	92.5	**92.5**	94.0	95.6	96.2	95.6	90.2	88.9	**92.8**	86.2	84.6

*WIS: Word Rankings for International Students
*F-JLPT: The former Japanese Language Proficiency Test
*VDRJ: Vocabulary Database for Reading Japanese
*AW: Academic Words
*LAD: Limited-academic-domain words
*LW: Literary Words
*Aca: Academic Vocabulary (AW & LAD)
*4D/3D/2D/1D: 4-/3-/2-/1-domain words
*AKW: Assumed Know Words (mostly proper nouns)
*Ha: Humanities & Arts
*Ss: Social Sciences
*Tn: Technological Natural Sciences
*Bn: Biological Natural Sciences

Table 7-33 Ranking for Text Covering Efficiency (TCE) of the Grouped Words in Each Genre (Detailed) *Domain-unspecified

			Corpus Code	MC	BSB	UPC	BCCWJ	UYN	TB	MTT-Ss	TIS	MTT-Bn	MTT-Tn	BCCWJ-T	JS-Bn	JS-Tn
			Genre	Conver-sation	Novels, Essays etc.	Essays, Novels etc.	Whole	News-paper	Ss	Ss (Intro.)	Ss & Ha	Bn (Intro.)	Tn (Intro.)	Academic (Various)	Bn (Journal Articles)	Tn (Journal Articles)
			Total Tokens (Million)	1.13	2.30	2.10	32.82	5.68	0.19	0.05	0.04	0.01	0.07	2.90	0.72	2.71
WIS	F-JLPT Level	Label 1	Label 2	Number of Lexemes in VDRJ	Ranking for TCE of the Grouped Words in Each Genre												
1-1,291	L4, L3	General	Basic	1,027	1	1	1	1	2	3	2	3	3	2	2	2	2
682-1,291	L3	AW	Basic+Aca4D	31	3	2	2	2	1	1	1	1	1	1	1	1	1
			Basic+Aca3D	39	4	3	3	3	3	2	3	2	4	3	3	3	3
		LAD	Basic+Aca2D	45	9	7	9	7	4	4	4	5	***12***	6	5	5	5
			Basic+Aca1D_Ah	13	8	5	5	5	6	7	6	4	***10***	9	9	***15***	***10***
			Basic+Aca1D_Ss	6	***12***	***12***	***12***	***10***	5	8	***12***	7	***21***	***30***	7	***10***	***19***
			Basic+Aca1D_Tn	5	7	8	8	8	***11***	***13***	***13***	***15***	***40***	***13***	***10***	7	6
			Basic+Aca1D_Bn	9	5	6	6	6	***12***	***11***	***15***	***14***	2	***17***	***13***	9	***12***
		LW	Basic+Lit	142	2	4	4	4	***13***	***12***	***10***	9	9	7	***12***	***12***	***15***
1,292-5,000		General	Inter	1,478	10	15	13	18	18	18	18	16	***22***	***25***	***23***	***26***	***28***
		AW	Inter+Aca4D	559	15	11	11	**9**	**8**	**5**	**5**	**6**	**5**	**4**	**4**	**4**	**4**
			Inter+Aca3D	542	***19***	17	14	13	**9**	**9**	**8**	10	**8**	10	**8**	**6**	**7**
		LAD	Inter+Aca2D	391	16	14	16	14	10	10	**9**	**8**	**7**	11	11	11	14
			Inter+Aca1D_Ah	104	18	10	10	12	15	14	11	12	14	***26***	15	16	***21***
			Inter+Aca1D_Ss	111	13	16	15	11	**7**	**6**	**7**	11	***28***	***31***	**6**	17	***23***
			Inter+Aca1D_Tn	46	11	18	18	15	16	16	***22***	17	11	**5**	14	14	**8**
			Inter+Aca1D_Bn	52	17	13	17	17	14	***24***	16	13	**6**	***24***	16	**8**	***30***
		LW	Inter+Lit	446	**6**	**9**	**7**	16	***22***	***23***	***23***	***19***	***26***	***19***	***26***	***30***	***32***
5,001-10,000	L2	General	Adv	3,070	20	20	21	27	27	***28***	***32***	25	***32***	***33***	***40***	***36***	***34***
		AW	Adv+Aca4D	212	***31***	***30***	24	26	23	19	**17**	22	**17**	**14**	**18**	**13**	**11**
			Adv+Aca3D	452	***30***	26	26	25	21	20	21	21	**13**	**12**	20	**18**	**16**
		LAD	Adv+Aca2D	429	25	25	25	23	20	21	19	18	**18**	**15**	21	22	**17**
			Adv+Aca1D_Ah	104	22	21	20	22	***30***	***29***	***28***	27	***29***	***35***	27	***37***	***31***
			Adv+Aca1D_Ss	127	27	24	***31***	19	**17**	**15**	**14**	20	***40***	***37***	**17**	***34***	***33***
			Adv+Aca1D_Tn	60	21	27	27	21	***29***	***33***	***33***	***31***	***35***	22	24	19	**13**
			Adv+Aca1D_Bn	68	23	22	23	20	24	***35***	20	28	**16**	***29***	25	23	***38***
		LW	Adv+Lit	483	**14**	19	19	24	***36***	***37***	***40***	***38***	***31***	***34***	***41***	***42***	***40***
10,001-15,000	L1 Others	General	H_Adv	3,681	28	32	30	36	***37***	34	36	34	34	***38***	***42***	***39***	***41***
		AW	H_Adv+Aca4D	103	43	40	33	***37***	32	**26**	**25**	29	**27**	**27**	31	**20**	**22**
			H_Adv+Aca3D	328	42	***38***	36	35	28	**25**	**27**	30	**23**	**18**	34	**27**	**25**
		LAD	H_Adv+Aca2D	296	33	35	35	33	**26**	**27**	**26**	**26**	**20**	**16**	30	31	**24**
			H_Adv+Aca1D_Ah	71	32	31	29	31	***41***	***39***	35	37	***40***	***39***	32	***38***	29
			H_Adv+Aca1D_Ss	74	29	33	***39***	28	**19**	**17**	30	**24**	***40***	36	**19**	35	***42***
			H_Adv+Aca1D_Tn	48	34	***39***	32	30	35	36	***47***	***42***	***40***	**8**	29	**25**	**9**
			H_Adv+Aca1D_Bn	54	**26**	29	38	29	31	***38***	41	33	**15**	43	28	**21**	36
		LW	H_Adv+Lit	345	**24**	**23**	**22**	32	***44***	***44***	***46***	***45***	***38***	***41***	***44***	***45***	***45***
15,001-20,000		General	S_Adv	4,046	**35**	**36**	**34**	45	42	40	42	41	**36**	40	43	41	43
		AW	S_Adv+Aca4D	56	44	45	45	43	39	**30**	**29**	**23**	**24**	**20**	37	**28**	**18**
			S_Adv+Aca3D	269	***46***	43	40	44	**33**	**32**	**31**	**35**	**33**	**23**	38	**33**	**26**
		LAD	S_Adv+Aca2D	232	45	44	41	42	**34**	**31**	**34**	40	**25**	**28**	**36**	**29**	**27**
			S_Adv+Aca1D_Ah	60	**36**	**34**	44	41	40	41	39	39	40	***46***	39	40	**35**
			S_Adv+Aca1D_Ss	55	40	42	43	**34**	**25**	**22**	**24**	**36**	40	***46***	**22**	43	39
			S_Adv+Aca1D_Tn	29	39	41	37	39	43	42	38	47	**19**	**21**	**33**	**32**	**20**
			S_Adv+Aca1D_Bn	53	38	37	42	38	38	43	37	**32**	**30**	**32**	**35**	**24**	37
		LW	S_Adv+Lit	200	37	**28**	**28**	40	***46***	***47***	43	43	40	45	***47***	44	***47***
20,000+		21K+	21K+	91,104	47	47	47	47	47	46	**45**	46	**37**	**42**	46	46	**44**
--		AKW	AKW	30,821	**41**	46	46	46	**45**	**45**	44	**44**	**39**	**44**	**45**	47	46

*TCE means the expected number of tokens of a lexeme in the tested group in a one-million-token text in the target domain.

*WIS: Word Rankings for International Students

*F-JLPT: The former Japanese Language Proficiency Test

*VDRJ: Vocabulary Database for Reading Japanese

*AW: Academic Words

*Aca: Academic Vocabulary (AW & LAD)

*LAD: Limited-academic-domain words

*4D/3D/2D/1D: 4-/3-/2-/1-domain words

*LW: Literary Words

*AKW: Assumed Know Words (mostly proper nouns)

*Ha: Humanities & Arts

*Ss: Social Sciences

*Tn: Technological Natural Sciences

*Bn: Biological Natural Sciences

*Numbers in bold show the rankings higher than expected ranking i.e. 1-9 for basic, 10-18 for intermediate, 19-27 for Adv. 1, 28-36 for Adv. 2, 37-45 for S-Adv and 46-472 for 21K+ and AKW. On the other hand, Italic numbers show the rankings lower than expected ranking.

7.4.5.2.2 Efficient learning order of words

Text Covering Efficiency (TCE) by more detailed grouping of words and the ranking for these groups in each genre are shown in Table 7-32 and 7-33.

In Table 7-32, it is easy to see that 1-domain words in a particular domain provide much higher TCE figures in that particular domain. For example, 1-domain words for technological natural sciences at the S-Adv. level (16K-20K) provides 19.4 TCE in JS-Tn (journal articles in technological natural sciences) while 1-domain words for the other three domains only provide 1.7-2.9 TCE in JS-Tn.

The key for answering the main research questions is shown in Table 7-33. That is, the most efficient learning order of words is to follow the order of Text Covering Efficiency (TCE) in the target genre. For example, if a learner aims to be able to read Japanese newspapers, the most efficient learning order of words must be the order shown in the column of UYN in Table 7-33. Within each group of words, it must be efficient to follow the adjusted frequency rankings of VDRJ introduced in Chapter 3. When we want to compare the efficiency of grouped words, we can look at Table 7-32. If the comparison between different genres is not necessary, domain-specified analysis will provide more accurate information.

How can we apply this results and method to teaching and learning? If a group of learner are working or will work on a specific genre/major, the TCE order in the target genre/major can be applied directly to the group of learners. If not, as discussed in 7.1.1.3, vocabulary learning should go from a wider to narrower range of domains according to the learners' level of study, namely first year, undergraduate major and postgraduate studies. In a preparatory (or maybe first year) curriculum for tertiary education, common academic words must be very useful. After entering a university, if the major is already limited within humanities social sciences or natural sciences, then limited-academic domain words will

also be useful. From the viewpoint of teaching vocabulary, grouping learners at different stages of curriculum will lead to a more efficient way.

TCE is a simple, convenient and strong predictor of learning efficiency in gaining text coverage. This index is not necessarily limited to this study. If a learner or a teacher aims to learn/teach a specific domain of texts, TCE can be calculated if s/he has a set of target texts which reflect the learners' needs. As introduced in 7.4.1.2, only three figures below are needed to calculate TCE, i.e. 1) Number of tokens of the tested group of words in the target text, 2) Number of lexemes of the tested group of words, 3) Number of total tokens in the target text. One strong point of TCE is that it enables us to compare the efficiency quantitatively. We can estimate how many times as efficient learning a group of words will be as learning another group. TCE is not influenced by the text size. It is comparable across genres and/or levels as it just shows the expected number of tokens of a lexeme (which can be another unit such as a word family or type depending on the purpose) of the tested group of words.

Of course, the efficiency here only means efficient gain of text coverage in a text; therefore, other factors must also be considered. Such factors include the complexity of orthographical and phonological forms, meaning and grammatical function of the words, which contribute to learnability. Nevertheless, words in the target texts reflect social needs. Even if the words in the texts are difficult, learners need to understand the words to understand the texts.

7.4.5.2.3 How does learner's language background possibly affect the understanding of texts?

Table 7-34 shows the proportion of word origins (counted by lexemes) of different groups of words in the top 20,000 words.

Table 7-34 Proportion of Word Origins (Counted by Lexemes) by Different Groups of Words in the Most Frequent 20,000 Words

Word Origin / Word Tier	Label	Number of Lexemes	Japanese (%)	Chinese (%)	Western & Other (%)	Mixed (%)	Proper Nouns (%)	Unknown & Signs (%)	Total (%)
General	General	13,302	38.4	45.3	10.8	3.2	1.5	0.8	100.0
Academic	AW	2,591	15.0	**75.2**	7.0	1.9	0.4	0.5	100.0
Limited-academic-domain	LAD	2,542	12.4	**69.1**	**13.7**	1.7	2.2	1.0	100.0
Literary	LW	1,616	**71.7**	21.8	2.5	3.1	0.3	0.6	100.0
Overlap	--	-27	74.1	22.2	0.0	3.7	0.0	0.0	100.0
Total (*)	--	20,024	34.7	50.3	10.0	2.8	1.4	0.8	100.0

*Including 24 compound numerals (01K+)

Origins of academic and literary words are considerably clearly separated. Japanese-origin words are significantly dominant at 71.7% in literary words while Chinese-origin words are significantly dominant at 75.2% and 69.1% in common academic words (AWs) and limited-academic-domain words (LADs) respectively. LADs contain more Western-origin words (Gairaigo, e.g., エンジン 'enjin' (engine), ボランティア 'borantia' (volunteer)) at 13.7% which is almost double the proportion for common academic words at 7.0%. Western-origin words tend to be used as technical terms in particular domains.

Chinese learners of Japanese should have a large advantage in understanding words used in academic texts. Not only the proportion of Chinese-origin words is high in academic vocabulary, but also semantic gaps with these cognates between Japanese and Chinese are relatively small, since a large amount of academic vocabulary is so-called 'new Sino-Japanese words' (新漢語 'shin-kango') created with Chinese-origin word parts relatively lately by Japanese academics in the Meiji era (1868-1912), exported to China by Chinese students who studied in Japan, and spread over China (Suzuki, 1981).

The different proportions of different word origins will directly lead to different degrees of learning burden depending on the learner's first language. As discussed in 4.5, this is a serious problem in curriculum design for teaching Japanese as a second language. The gap is larger in academic and literary texts than in general texts.

7.5 Implications and remaining issues

Academic vocabulary has a relatively clear domain-specificity. That is, academic vocabulary is frequent in academic texts but not in general texts. This suggests that the understanding of these words can be a key or a barrier for academic success (Corson, 1985, 1997). In other words, understanding of these words may be a predictor of academic success. The relationship between the lexical knowledge of academic vocabulary and general academic performance is an important topic to explore. This is not necessarily limited to second language learners, but can contribute to learners with any language background including first language learners (Townsend & Collins, 2008).

A vocabulary-conscious curriculum should be designed and incorporated in Japanese programs depending on the learners' needs and language backgrounds. As Chinese-origin words account for three quarters of academic vocabulary[96], if a curriculum is for academic purposes, an extra treatment for non-Chinese-background learners is particularly required, especially in reading and writing. As discussed in 2.5.2, autonomous mode for learning vocabulary will be necessary particularly when the learners' needs and language backgrounds are various. Especially, limited-academic-domain words and literary words are important for some learners but may not be so important for other learners.

It is also important to study how we can exploit these domain-specific word lists for classroom teaching. We need to figure out good ways for teaching common academic words as they are highly abstract. Lists can be used for checking gaps in learner knowledge at least.

The gap between Chinese-background learners (CBLs) and non-CBLs will be less in basic conversation and reading literary works than in reading academic texts; however, especially the levels beyond 10K, literary words will also play an important role, as the literary words are not common in daily-life conversation at the low-frequency level.

[96] This seems a similar feature to the status of Graeco-Latin words in English (Corson, 1985, 1997).

There are some limitations with this study. Some of them are issues with the extraction and analysis of the domain-specific words, others are issues with specific word lists.

First, the unit of analysis is limited to the lexeme. As is often the case with vocabulary studies, multi-word units (MWU) are not considered in this study. Some MWU should have higher frequency than lexemes and perform like a word in the texts. This is one of the future research topics. Also, individual Kanji (word parts except for affixes) are not considered in this chapter. As discussed in previous chapters (especially in Chapter 6), the Japanese language has many (semi-)transparent compounds composed of a limited number of Kanji. Therefore, it may be useful to explore Kanji tiers and how they are related to the word tiers. It is not done here as it would be too complicated; however, the idea will be the same as the conclusion of Chapter 6. As I discussed with Kanji used for common academic words in 7.2.3.5, many Kanji are recycled but some of them are not. Learning words in a sentence or a wider context should be the basic way of learning Japanese vocabulary; however, considering the complexity of Kanji orthography, the possibility of semantic inference from word parts, the importance of Kanji as components of compounds, a 'bottom-up' way by learning individual Kanji with the compound words along with the top-down way should also be an efficient method of learning Japanese vocabulary. In this sense, a 'Japanese academic Kanji list' and a 'Japanese literary Kanji list' may also be of some value.

Second, as is often the case with corpus studies, homographs and polysemy (figurative usages of words) are not considered for this study. If an academic usage of a word is derived from a metaphorical usage of a daily-life word, it is not likely to be extracted. For example, the word 注ぐ 'sosogu' (pour) is used as a verb for liquids as well as 力 'chikara' (power), 情熱 'jounetsu' (passion), 心血 'shinketsu' (heart and soul), 精力 'seiryoku' (energy) or 愛情 'aijou' (affection) as a frequent metaphorical usage. It also has

academic usages such as “信濃川が日本海に注ぐ” (The Shinano river flows into the Japan Sea). In these cases, it is hard to extract it as a common academic word even if it is high-frequency in an academic field.

Third, limited-academic-domain words (LADs) will be less valid and reliable than common academic words as the corpus is not designed for academic purposes. In particular, there are not enough tokens in some academic fields such as pharmacy or dentistry. LADs are not technical terms; however, the extracted words may be biased for LADs if the size of sub-sections is not large enough. It is desirable to have a more substantial academic corpus.

Fourth, related to the previous point, which level of words is worth being made into a word list, is still not clear. I believe JAWLⅠ and II are good lists, yet, I am still not totally sure for the other groups of words. TCE shows the usefulness of these words; however, if the separate lists contain thousands of words, learners may be discouraged by them. Careful steps will be required for supplying the lists to leaners. The groups of words are surely useful for the word tier analysis to clarify the lexical features of genres and to assess the value of grouped words for a genre, though.

Fifth, JAWL (Japanese Common Academic Word List) contains a few inappropriate words at low-frequency levels, e.g. 同校 ‘doukou’ (the aforementioned school), 四面 ‘shimen’ (the four sides, all sides), ユア ‘yua’ (your), そり ‘sori’ (sleigh, sledge), ずる ‘zuru’ (cheating, foul play) and でんぷん ‘dempun’ (starch). This is probably due to the error of word-segmentation or the set level for the cut-off point. Leech, Rayson, & Wilson (2001) set the cut-off point of log-likelihood ratio as 3.8 because it is the border for significance with $p < .05$. I was afraid that some important words are missing from the list; however, some words should be removed from the list by checking the usage by a concordance.

Sixth, the grading of JAWL may be somewhat arbitrary, especially for 3-domain words. It is not easy to tell if the domain-specified analysis is appropriate for 3-domain

words; however, the TCE should be compared between 4-domain, 3-domain words and 2-domain words by domain-specified analysis to decide if, for example, intermediate 3-domain words are more important than advanced 4-domain words.

Seventh, as I have already mentioned, literary words are common in conversation at least up to the 10K level. Elaborating literary words is also a future research topic.

7.6 Conclusion of Chapter 7

In this chapter, after reviewing relevant previous studies and proposing specific research questions, I extracted common academic words, limited-academic-domain words and literary words first, and then examined their features, distribution and Text Covering Efficiency to evaluate their usefulness in different genres. To decide the most efficient learning order of words as well as clarifying lexical features of different genres, word tier analysis was proposed and conducted in 7.5.

The most important claim i.e. the answer to the main research questions in this thesis is in this chapter. That is,

1) The most efficient learning order of words can be decided by Text Covering Efficiency (TCE) proposed in 7.4.1.2, which is the expected number of tokens of a lexeme of a tested group of words in a test corpus which reflect the learner needs. The greater the TCE, the more words in the target text likely to be covered by a lexeme of the grouped words. TCE can be compared with a general word frequency per million (as shown in Table 7-14).

TCE also provides a good analysis for clarifying lexical features of different text domains. Main specific findings based on these analyses of extracted domain-specific words and text domains include:

2) 2,541 common academic words (4-domain and 3-domain words) at nine levels in the Japanese Common Academic Word List (JAWL) provide remarkably higher text coverage and TCE in academic texts than other types of words. They also provide higher coverage and TCE in academic texts than in non-academic texts.
3) JAWL Ⅰ (559 words, intermediate) is the most important common academic word list. The words provide high text coverage and TCE in any type of academic text.
4) Academic vocabulary (common academic words (AWs) and limited-academic-domain words (LADs)) at advanced levels do not provide high text coverage; however, they provide much higher TCE for academic texts than other types of words.
5) Academic vocabulary contains more nouns, verbal nouns, affixes and archaic words than other types of words.
6) Many of the common academic words are used for managing academic information. The meanings of common academic words are highly abstract. Limited-academic-domain words (2-domain words and 1-domain words) have more concrete meanings.
7) Some combinations of the two domains for the 2-domain words show a particular semantic field, i.e. 'humanities and arts' × 'social sciences' = 'history' (especially political history), 'social sciences' × 'technological natural sciences' = 'industry' and 'social sciences' × 'biological natural sciences' = 'social security, medical and nursing'.
8) Only 27 literary words overlap with academic vocabulary. The 27 words account for 1.7% of literary words and 0.5% of academic vocabulary.
9) Academic texts show high TCE for academic vocabulary but low TCE for literary words. In contrast to that, literary texts show a moderately high TCE for literary words but low TCE for academic vocabulary. This means that academic and literary texts have totally different lexical features. Domain specificity is stronger in academic texts than in literary texts.

10) Literary words are the words for describing human actions and feelings vividly and effectively, as they contain numerous words for body parts and body actions, many modal adverbs, interjections and words for metaphorical expressions.
11) Literary words from the basic to 10K level also provide high coverage and TCE for conversation; however, literary words at 11-15K level or above only provide higher TCE for literary texts but not for conversation.
12) Origins of academic and literary words are considerably clearly separated; 3/4 of literary words originate in Japanese while 3/4 of academic vocabulary originates in Chinese. LADs contains more Western-origin words (Gairaigo)
13) Newspaper texts have similar lexical features to social science texts. Newspaper texts will be a good resource for learning academic vocabulary.
14) Natural science texts have more low-frequency words.

The most important tables for this thesis are Tables 7-29, 7-30, 7-31 and 7-32 as they show the expected learning efficiency of different groups of words, specific learning order of grouped words, and different lexical features of different genres.

Chapter 8 Analysing a Japanese reading text as a vocabulary learning resource by lexical profiling and indices

8.1 Introduction

From Chapter 3 to Chapter 7, I investigated the efficient learning order of vocabulary mainly by analysing words and characters. The findings, databases and word lists can be exploited by learners, teachers or researchers. In this chapter, as one use of the vocabulary database (VDRJ) and the word lists, I will show a method for analysing a reading text from a teacher's or a material developer's viewpoint.

Specifically, I will discuss how we can control the vocabulary of a reading text to maximize the vocabulary learning effect. If a text is too easy for a learner, there will be few words to learn in the text. On the other hand, if a text contains too many unknown words, no inference is likely to occur, let alone learning. The goal for this chapter is to show methods to assess a (Japanese) reading text as a vocabulary learning resource by exploiting lexical profiling and indices. I will also propose a systematic way to control the vocabulary load of a text for learners to read.

The research questions for this chapter are:

1) How can we assess a reading text as a resource for vocabulary learning? How can it be expressed in numbers to allow us to make comparisons between different texts?
2) How can a reading text be modified as a resource for vocabulary learning?

The main points are as follows.

The simplest way to rewrite a reading text (with 2000 words or less) for a better resource for vocabulary learning is 1) Delete or replace one-timers (or the words whose occurrences are less than the set level) at the lowest frequency level in the text, or 2) Make

the one-timers occur more in the text by adding words or replacing other words with the one-timer.

For this attempt, I propose an index entitled LEPIX for Lexical Learning Possibility Index for a Reading Text. By taking steps 1) and 2) shown above, the LEPIX figure will be improved. These methods make it possible to predict and compare the efficiency of second language vocabulary learning with a reading text.

8.2 Significant research

There are some similar previous ideas and attempts for assessing a reading text as a lexical learning resource and/or proposing a systematic way to rewrite a text (Cobb, 2007; Ghadirian, 2002; I. S. P. Nation & Deweerdt, 2001). There are many arguments about the usefulness of and methods for text modification including simplification of vocabulary and grammar, mainly in English studies. Studies which take a relatively negative position to simplification include Honeyfield (1977), Yano, Long, & Ross (1994) and Young (1999). Most of their arguments are based on the measure of reading comprehension but not the measure of vocabulary gain. There are also some recent studies including Gardner & Hansen (2007) and Nation & Deweerdt (2001) which are positive to simplification. They claim a couple of reasons to justify the merits of simplification; however, I just focus on one point to support their argument. That is, when a learner is able to understand enough words, they can read an unsimplified text, therefore, any material which contributes to vocabulary gain is useful. As Nation & Deweerdt (2001) claim, many issues are not the matter of reading texts but the matter of course design.

Despite some arguments about the value and method of simplification, numerous graded readers are widely exploited in learning and teaching English for both first language and second language learners. This also leads to studies on examining the usefulness of extensive reading (e.g. Elley & Mangubhai, 1981, 1983) and factors of incidental

vocabulary learning (e.g. Laufer & Hulstijn, 2001; Waring & Takaki, 2003; Webb, 2008). In Japanese studies, there are also some attempts to develop extensive reading programmes (Hitosugi & Day, 2004) materials by controlling lexical and grammatical items (Mikami & Harada, 2011). However, no integrated index for the possibility of vocabulary learning is shown in previous studies. Also, for Japanese reading texts, there needs to be a method for controlling Kanji as well as vocabulary; however, there seem few studies on the issue for second language learners.

In this chapter, I will focus on controlling vocabulary and Kanji in the target text from a lexical learning perspective but not from other aspects such as readability. The suggested application of this study will not be limited to developing extensive reading materials, but will be extended to developing course material used for classroom teaching.

The term 'lexical profiling' used for this chapter is basically the same idea as Lexical Frequency Profiling (LFP) which is defined as "the percentage of words …… at different vocabulary frequency levels" (Laufer, 1994, p 23). Laufer used this term as an index for assessing a learner's composition. I apply this concept for assessing a reading text based on the simple definition as shown above[97].

8.3 Assumptions for developing a new index: LEPIX

There are four important assumptions for developing the LEPIX (Lexical Learning Possibility Index for a Reading Text).

The first assumption is about the required level of text coverage. That is, words which are assumed known to the reader must be within a certain level. (For details, see the related studies introduced in 2.2.2.)

The second assumption is about the minimum occurrences of target words (lexemes). That is, among the words assumed unknown, words which occur more frequently than a

[97] This is also a similar idea to the 'word tier analysis' introduced in Chapter 7.

certain times can be the learning target words (Hulstijn, Hollander, & Greidanus, 1996; Waring & Takaki, 2003; Webb, 2007).

The third assumption is on the number of lexemes (or word families). That is, the text where the more (types of) target lexemes occur is a better text for vocabulary learning. Note that the second assumption is about the number of tokens of the target lexemes while the third assumption is about the number of lexemes.

The fourth assumption is on the density of the target words (lexemes). That is, the text where the target lexemes occur at a higher proportion is a better text as a vocabulary learning resource.

8.4 Method for calculating LEPIX

In order to calculate LEPIX, baseword lists are needed for lexical profiling. VDRJ[98] baseword lists are used for this purpose. When analysing Japanese texts, it is also necessary to set a certain level of known characters (Kanji) as well as vocabulary. In order to control the Kanji level, CDJ[99] is used. The software tool AntWordProfiler Ver. 1.200W (Anthony, 2009) is used for lexical profiling.

The steps to calculate LEPIX are as follows.

1) To identify the lexical level of the text by lexical profiling, set the threshold level of (assumed) known words.

In this study, the levels are:

A) 98% for an extensive reading text

B) 95% for instructional material

I call these levels Lexical Level of Text 98 (LLT98) and Lexical Level of Text 95 (LLT95)

[98] See Chapter 3 for details.

[99] See Chapter 5 for details.

for convenience. These levels are set as a trial in reference to Hu & Nation (2000) and Laufer & Ravenhorst-Kalovski (2010); however, these can be changed depending on the situation.

2) To identify the target words, set the minimum occurrences of target words. 6-10 occurrences are required for learning a word incidentally through reading (e.g. Waring & Takaki, 2003); however, a word is not learned by reading one short text. Therefore, I set the minimum occurrences of target words as below.

   A) Twice or more for an extensive reading text

      Set occurrences will depend on the text length.

   B) Twice for a short instructional material

3) Count T which is the number of lexemes (or types) of the target words.

4) Calculate (W*100)/N where:

   W is the number of tokens of the target words.

   N is the total number of tokens of the text.

5) Calculate LEPIX (Lexical Learning Possibility Index for a Reading Text) by simply multiplying the factors of III & IV.

(LEPIX) I = T*(W*100)/N = (T*W*100)/N

8.5 A sample analysis of text by LEPIX

8.5.1 A sample modification of a text

Below is a sample original text and its modified text. The set known words level is

set at 95% instructional material used for classroom teaching. This level should agree with the learners' level which is ideally measured by a vocabulary test sampled from the same database used for modifying the text. Letters highlighted in bold Gothic in the original text are to be changed. The correspondent modified expressions are also highlighted in bold Gothic in the modified text. Underlined words are the target words in the both texts. Subscripts A-C attached to the underlined words mean the types of treatment shown below.

A: Target words changed from assumed known words due to the change of Lexical Level of Text (LLT)

B: Target words changed from non-target words by adding occurrences to one-timer

C: Newly added target words by replacing original expressions with new expressions

(From here down blank. See next page for the sample text.)

Sample Text (original) *highlighted by the author of this thesis

人知のシミュレーションが人工A知能だとすれば、コンピュータのなかに「知をあつかうBメカニズム」を**作り込まなければ**ならない。

ところでコンピュータとは、要するに〈A記号処理**マシン**〉である。だからこの場合の〈知〉とは、「A記号で表された知」ということになる。A記号といっても色々あるが、人工A知能が得意なのは、いわゆる言語A記号である。たとえば、「今は五月だ」「五月は春だ」「B**楓**の葉は、春と夏には**緑色**、秋には**赤色**である」などという**のがその守備範囲ということになる**。

ところでこういった例は、少しばかり**興ざめではなかろうか**？　というのは、〈知〉とは、単なる知識**の断片**ではなく、それらを**包括し**、B横断しながら世界に光を当てていく精神の**ダイナミズム**のように思えるからである。〈知〉は**イマジネーションの能力**を持たなければならない。さらに〈知〉は、スポーツのような身体の**所作にうめこまれている**、**明言化されない暗黙知**の領域をもカバーしなければならない。それこそが、知**の知たるゆえん**ではないだろうか？

残念ながら、現在の人工A知能技術は、この期待に応えるすべを知らない。**それはいまだに、図像さえ自由自在には扱えないのである**。英語や日本語などの〈自然言語〉を操作するだけでも**四苦八苦な**のである。

（出典：Nishigaki, T.（西垣 通）. *Hijutsu-toshite-no AI shikou*『秘術としてのＡＩ思考』（AI thinking as a secret technique).）

Many of low-frequency one-timers beyond the 95% coverage level (contained in the bold Gothic expressions in the original text) are to be changed as they are not likely to be learned according to the assumption. As a result, the Lexical Level of Text (LLT) moves down. In this sample case, it changed from 10K to 05K. In other words, the original text requires a vocabulary size of around 10,000 known words while the modified one only requires 5,000 words. Table 8-1 shows the treatment of low-frequency words in the sample texts.

Sample Text (modified)

人間のC頭脳**を**C模倣**して作ったもの**が人工A知能だとすれば、コンピュータの中に「知をあつかうBメカニズム」を**ていねいに作っていかなければ**ならない。しかしそこへの道はまだC程遠い。

コンピュータとは、要するに〈A記号処理のBメカニズム〉である。だからこの場合の知とは、「A記号で表された知」ということになる。A記号といってもいろいろあるが、人工A知能が得意なのは、いわゆる言語A記号である。例えば、「今は五月だ」「五月は春だ」「Bカエデの葉は、春と夏には**緑**、秋には**赤**である」などという**人工言語的表現は処理しやすいのである**。

しかし、こういった例は、少しばかり**つまらないのではないだろうか**？というのは、知とは、一つ一つの知識**がバラバラに存在するの**ではなく、それらを**一つにまとめたり**、B横断したりしながら、世界に光を当てていく精神の**力強い働き**のように思えるからである。知は**想像力**を持たなければならない。さらに知は、スポーツのような身体の**動きの中にある**、**はっきりとした言葉にならない知**の領域もカバーしなければならない。Bカエデといえば私たちが紅葉を見て感じる気持ちまでB横断的にカバーしなければならないのだ。それこそが、知**を知として成り立たせているもの**ではないだろうか。

残念ながら、現在の人工A知能技術は、この期待に応えるすべを知らない。人間のC頭脳のC模倣にはまだC程遠いレベルだ。英語や日本語などの〈自然言語〉を操作するだけでも**非常に苦労している**のである。

Table 8-1 Treatment of Low-frequency Words in the Sample Texts

Word Level in WIS	Lexeme	Frequency in Original	Cumulative Text Coverage	Frequency in Modified	Cumulative Text Coverage	Treatment
IS_05K	知	9	*88.7*	9	*94.1*	
IS_05K	紅葉	0	*88.7*	1	*94.4*	
IS_05K	**記号**	**4**	***90.2***	**4**	***95.6***	A
IS_06K	**マシン**	1	*90.5*	0	*95.6*	Deleted or Replaced
IS_06K	**メカニズム**	**1**	*90.9*	**2**	*96.2*	B
IS_06K	**横断**	**1**	*91.3*	**2**	*96.8*	B
IS_06K	**緑色**	1	*91.6*	0	*96.8*	Deleted or Replaced
IS_07K	**断片**	1	*92.0*	0	*96.8*	Deleted or Replaced
IS_07K	**自在**	1	*92.4*	0	*96.8*	Deleted or Replaced
IS_07K	**頭脳**	**0**	*92.4*	**2**	*97.3*	C
IS_08K	**包括**	1	*92.7*	0	*97.3*	Deleted or Replaced
IS_08K	**暗黙**	1	*93.1*	0	*97.3*	Deleted or Replaced
IS_08K	**楓**	**1**	*93.5*	**2**	*97.9*	B
IS_08K	**模倣**	**0**	*93.5*	**2**	*98.5*	C
IS_08K	**知能**	**3**	***94.5***	**3**	***99.4***	A
IS_08K	**程遠い**	**0**	*94.5*	**2**	*100.0*	C
IS_09K	**守備**	1	*94.9*	0	*100.0*	Deleted or Replaced
IS_10K	**シミュレーショ**	1	***95.3***	0	*100.0*	Deleted or Replaced
IS_10K	**埋め込む**	1	*95.6*	0	*100.0*	Deleted or Replaced
IS_11K	**明言**	1	*96.0*	0	*100.0*	Deleted or Replaced
IS_11K	**赤色**	1	*96.4*	0	*100.0*	Deleted or Replaced
IS_16K	**所作**	1	*96.7*	0	*100.0*	Deleted or Replaced
IS_17K	**図像**	1	*97.1*	0	*100.0*	Deleted or Replaced
IS_19K	**八苦**	1	*97.5*	0	*100.0*	Deleted or Replaced
IS_19K	**四苦**	1	*97.8*	0	*100.0*	Deleted or Replaced
IS_20K	**ダイナミズム**	1	*98.2*	0	*100.0*	Deleted or Replaced
IS_21K+	**イマジネーショ**	1	*98.5*	0	*100.0*	Deleted or Replaced
IS_21K+	**人知**	1	*98.9*	0	*100.0*	Deleted or Replaced
IS_21K+	**作り込む**	1	*99.3*	0	*100.0*	Deleted or Replaced
IS_21K+	**由縁**	1	*99.6*	0	*100.0*	Deleted or Replaced
IS_21K+	**興醒め**	1	*100.0*	0	*100.0*	Deleted or Replaced

*WIS: Word Ranking for International Students

*Explanation of Treatment

A: Changed from an assumed known word to a target word due to the change of Lexical Level of Text (LL

B: Changed from a non-target word to a target word by adding occurrences to one-timers

C: A newly added target word by replacing original expressions with new expressions

Kanji frequency level also needs to be controlled. In the sample case above, the Lexical Level of Text for 95% coverage (LLT 95) is set at 05K after the modification of the text. 5,000 words are almost covered by the former Japanese Language Proficiency Test Level 2; therefore, the Kanji level can be set at 1,000 (10C in the Character Database of Japanese (CDJ)) or maybe slightly more basic to around 800 depending on the learners' readiness level.

The frequencies in the original and modified texts stay the same for the words with the A-type treatment (e.g. 記号 'kigou' (sign) at 05K). These words are actually not changed at all. Just because the Lexical Level of Text for 95% coverage (LLT 95) changed, these words naturally become the target words. If these words are one-timers, some treatment is required to make the text a better resource for vocabulary learning.

The frequencies in the original and modified texts increased from 1 or 0 to 2 for the words with the B-type and C-type treatment (e.g. メカニズム 'mekanizumu' (mechanism) at 06K with B-type treatment and 頭脳 'zunou' (brain, head) at 07K with C-type treatment). These words are at a higher level than the Lexical Level of Text (LLT) after the modification. They became target words by adding occurrences instead of being replaced or deleted. Many other words are deleted as they are not likely to be learned if they stay the same. We need to think about whether low-frequency one-timers in the target text should be kept, replaced or deleted. If we decide to keep a one-timer, we need to add occurrences of the word to the set minimum occurrences so that the word is more likely to be learned.

What are the LEPIX and relevant statistical figures with these sample texts? How do these change after the modification? Table 8-2 is the comparison of the figures between the original and the modified text.

Table 8-2 LEPIX and Relevant Statistical Figures in the Original and Modified Sample Texts

Item	Original Text	Modified Text
Text Length (= Total Number of Token) (**N**)	275	339
Total Number of Lexemes	118	130
Number of Tokens over 95% Text Coverage	14	19
Number of Lexemes over 95% Text Coverage	14	8
95% Text Coverage Level = Lexical Level of the Text (**LLT95**)	***10K***	***05K***
Minimum Occurrences of Target Words over 95% Text Coverage	2	2
Number of Target Tokens over 95% Text Coverage (**W95**)	0	19
Number of Target Lexemes over 95% Text Coverage (T95)	0	8
Density of Target Words (%) (**W95*100/N**)	0.0	5.6
Average Occurrences of a Target Lexeme (**W95/T95**)	0.0	2.4
Lexical Learning Possibility Index for a Reading Text over 95% Text Coverage (**LEPIX95**) (**(T95*W95*100)/N**)	**0.0**	**44.8**
Number of Tokens over 98% Text Coverage	6	7
Number of Lexemes over 98% Text Coverage	6	3
98% Text Coverage Level = Lexical Level of the Text (**LLT98**)	***20K***	***08K***
Minimum Occurrences of Target Words over 98% Text Coverage	2	2
Number of Target Tokens over 98% Text Coverage (**W98**)	0	7
Number of Target Lexemes over 98% Text Coverage (T98)	0	3
Density of Target Words (%) (**W98*100/N**)	0	2.1
Average Occurrences of a Target Lexeme (**W98/T98**)	0.00	2.3
Lexical Learning Possibility Index for a Reading Text over 98% Text Coverage (**LEPIX98**) (**(T98*W98*100)/N**)	**0.0**	**6.2**

LEPIX: Lexical Learning Possibility Index for a Reading Text
The formula for LEPIX (*I*) is as follows.
I = T*(W*100)/N = (T*W*100)/N

The number of tokens in the text increased in the modified text while the number of lexemes decreased. As a result, the number of target lexemes which meet the required minimum occurrences increased drastically from 0 to 8 for 95% coverage and from 0 to 3 for 98% coverage. LEPIX figures improved from 0.0 to 44.8 and 6.2 for 95% (LEPIX 95) and 98% coverage (LEPIX 98) respectively. $T_{95/98}$ the represents number of target lexemes which refer to how many opportunities those are to meet different types of lexemes. ($W_{95/98}$*100)/N represents the density of target words (%) which is expected to predict the possibility of acquisition or consolidation of the target words per a unit of length. LEPIX (*I*) is a product of these two. It is expected to represent how good a text is for learning vocabulary.

8.5.2 Analysis of a text for learning domain-specific words

When we want to teach a specific group of words such as technical vocabulary in a specific field, then how can we calculate LEPIX? The basic idea is the same; however, the method for identifying the target words is different. The steps are shown below.

1) The target domain is set up at first (e.g. economics)
2) The domain-specific words included in the text are identified by checking the list of the domain-specific words
3) The levels of the identified domain-specific words included in the text are checked by lexical profiling to see how many unknown domain-specific words are contained in the text
4) The indices are calculated

The sample analyses of two modified economics texts are shown in Table 8-3. Except for the method for identifying target words, there is no difference in calculating LEPIX. $LEPIX_{95}$ for the two sample texts are 12.9 and 6.7 which are lower than the sample modified text shown in Table 8-2, just because some non-technical words are not identified as target words. If a teacher or a material developer aims to teach vocabulary in a limited domain, it will be harder to gain a high LEPIX figure without finding a text which contains high proportion of target domain-specified vocabulary at the target learners' level.

Table 8-3 LEPIX and Relevant Statistical Figures in Two Sample Modified Texts (Technical Words as Target Words)

Text Number	#1	#2
Text Length (= Total Number of Token) (**N**)	1193	2823
Total Number of Lexemes	250	690
Target Domain	Economics	
Number of Tokens over 95% Text Coverage	60	142
Number of Lexemes over 95% Text Coverage	24	87
95% Text Coverage Level = Lexical Level of the Text (**LLT$_{95}$**)	***04K***	***08K***
Number of ***Technical Word*** Tokens over 95% Text Coverage	25	35
Number of ***Technical Word*** Lexemes over 95% Text Coverage	10	15
Number of ***Technical Word*** Tokens over 95% Text Coverage (**W$_{95t}$**)	22	27
Number of ***Technical Word*** Lexemes over 95% Text Coverage (T95t)	7	7
Density of ***Technical*** Target Words (%) (**W$_{95t}$*100/N**)	1.84	0.96
Average Occurrences of ***Technical*** Target Words (**W$_{95t}$/T$_{95t}$**)	3.14	3.86
Lexical Learning Possibility Index for a Reading Text over 95% Text Coverage (**LEPIX 95$_t$**) **((T95$_t$*W95$_t$*100)/N)**	**12.9**	**6.7**
Number of Tokens over 98% Text Coverage	12	52
Number of Lexemes over 98% Text Coverage	8	37
98% Text Coverage Level = Lexical Level of the Text (**LLT$_{98}$**)	***09K***	***12K***
Number of ***Technical Word*** Tokens over 98% Text Coverage	7	9
Number of ***Technical Word*** Lexemes over 98% Text Coverage	4	6
Number of ***Technical Word*** Tokens over 95% Text Coverage (**W$_{98t}$**)	5	5
Number of ***Technical Word*** Lexemes over 95% Text Coverage (T98t)	2	2
Density of ***Technical*** Target Words (%) (**W$_{98t}$*100/N**)	0.42	0.18
Average Occurrences of ***Technical*** Target Words (**W$_{98t}$/T$_{98t}$**)	2.50	2.50
Lexical Learning Possibility Index for a Reading Text over 98% Text Coverage (**LEPIX 98$_t$**) **((T98$_t$*W98$_t$*100)/N)**	**0.8**	**0.4**

8.6 How does the text length distort LEPIX figures?

There is one weak point with LEPIX, that is, LEPIX figures cannot be compared if the text lengths are too different. As some previous studies (e.g. Richards & Malvern, 1997) point out, the number of types ('lexemes' in this case) and tokens are generally not in proportion even if the texts come from a single domain.

Table 8-4 LEPIX and Relevant Statistical Figures for Differently-sized Texts

Text Number	#5-1	#4-3	#3-1	#4-1	#8-2	#6-1	#1-3	#2-2	#9-2	#1-2	#8-1	#1-1	#3-3	#2-1	#2-3	#3-2	#9-1	#4-2	#7-1	#6-2	#7-2	#5-2	#9-3	M	SD
Text Length (= Total Number of Token) (N)	504	616	959	1055	1092	1193	1210	1317	1416	1418	1455	1592	1717	1785	1959	2035	2241	2342	2361	2823	2964	3754	4344	1832.7	928.0
Total Number of Lexemes	226	246	358	296	282	250	335	409	383	406	400	540	530	560	528	621	533	535	555	690	628	849	923	481.9	180.8
Number of Tokens over 95% Text Coverage	26	31	50	53	64	60	61	68	71	71	80	83	86	91	99	102	113	118	120	142	149	227	297	98.3	60.1
Number of Lexemes over 95% Text Coverage	24	19	37	43	39	24	37	48	53	51	32	73	82	62	67	84	77	82	68	87	71	138	138	62.4	31.0
95% Text Coverage Level = Lexical Level of the Text (**LLT95**)	***07K***	***08K***	***08K***	***04K***	***05K***	***04K***	***06K***	***09K***	***06K***	***07K***	***10K***	***06K***	***07K***	***08K***	***07K***	***07K***	***06K***	***05K***	***06K***	***08K***	***08K***	***06K***	***07K***		
Number of Target Tokens over 95% Text Coverage (**W95**)	4	14	20	17	36	47	33	33	27	25	58	15	7	33	45	27	55	50	68	81	99	115	184	47.5	40.1
Number of Target Lexemes over 95% Text Coverage (T95)	2	2	7	7	11	11	9	13	9	5	10	5	3	4	13	9	19	14	16	26	21	26	25	11.6	7.3
Density of Target Words (%) (**W95*100/N**)	0.8	2.3	2.1	1.6	3.3	3.9	2.7	2.5	1.9	1.8	4.0	0.9	0.4	1.8	2.3	1.3	2.5	2.1	2.9	2.9	3.3	3.1	4.2	2.4	1.0
Average Occurrences of Target Words (**W95/T95**)	2.0	7.0	2.9	2.4	3.3	4.3	3.7	2.5	3.0	5.0	5.8	3.0	2.3	8.3	3.5	3.0	2.9	3.6	4.3	3.1	4.7	4.4	7.4	4.0	1.6
Lexical Learning Possibility Index for a Reading Text over 95% Text Coverage (**LEPIX 95**) (**(T95*W95*100)/N**)	1.6	4.5	14.6	11.3	36.3	43.3	24.5	32.6	17.2	8.8	39.9	4.7	1.2	7.4	29.9	11.9	46.6	29.9	46.1	74.6	70.1	79.6	105.9	32.3	27.6
Number of Tokens over 98% Text Coverage	11	13	21	22	24	24	25	27	30	30	30	32	36	36	40	41	45	48	50	57	61	76	87	37.7	18.5
Number of Lexemes over 98% Text Coverage	10	13	16	15	17	9	12	21	22	21	16	29	34	35	36	39	34	40	25	38	33	66	72	28.4	15.9
98% Text Coverage Level = Lexical Level of the Text (**LLT98**)	***13K***	***10K***	***18K***	***11K***	***11K***	***09K***	***18K***	***18K***	***11K***	***12K***	***15K***	***12K***	***13K***	***13K***	***11K***	***11K***	***11K***	***09K***	***16K***	***12K***	***13K***	***11K***	***12K***		
Number of Target Tokens over 98% Text Coverage (**W98**)	2	0	8	11	12	19	15	10	11	11	18	4	3	2	8	4	20	14	30	31	39	18	23	13.61	9.92
Number of Target Lexemes over 98% Text Coverage (T98)	1	0	3	4	5	4	2	4	3	2	4	1	1	1	4	2	9	6	5	12	11	8	8	4.35	3.23
Density of Target Words (%) (**W98*100/N**)	0.4	0.0	0.8	1.0	1.1	1.6	1.2	0.8	0.8	0.8	1.2	0.3	0.2	0.1	0.4	0.2	0.9	0.6	1.3	1.1	1.3	0.5	0.5	0.7	0.4
Average Occurrences of Target Words (**W98/T98**)	2.0	0.0	2.7	2.8	2.4	4.8	7.5	2.5	3.7	5.5	4.5	4.0	3.0	2.0	2.0	2.0	2.2	2.3	6.0	2.6	3.5	2.3	2.9	3.2	1.6
Lexical Learning Possibility Index for a Reading Text over 98% Text Coverage (**LEPIX 98**) (**(T98*W98*100)/N**)	0.4	0.0	2.5	4.2	5.5	6.4	2.5	3.0	2.3	1.6	4.9	0.3	0.2	0.1	1.6	0.4	8.0	3.6	6.4	13.2	14.5	3.8	4.2	3.9	3.8

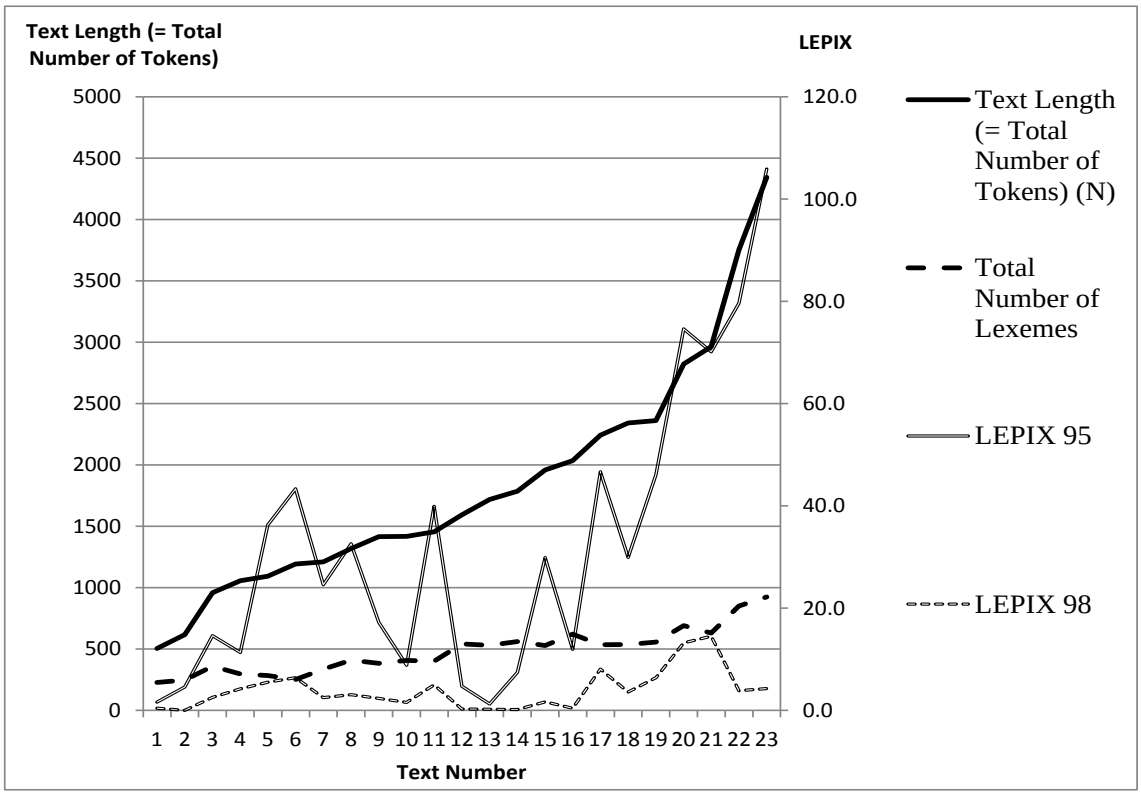
* Data from passages in a textbook (Shinya & Matsushita, 1994) which are mostly authentic but slightly modified for advanced learners of Japanese

** Minimum Occurrences of Target Words over 95%/98% Text Coverage = 2

I tried other ways to correct this flaw, for example, using logarithms and deleting one-timers[100]. However, the results were not ideal. Also, if the formula gets too complicated, it seems unrealistic to calculate and harder to interpret. Therefore, I decided to use the formula shown in 8.4.

Now the question is: how can differently-sized texts be compared? Graph 8-1 shows the total number of tokens/lexemes and LEPIX from twenty three differently-sized (504-4,344 tokens) texts (Text number 1 to 23 in Table 8-4). The texts are from Shinya & Matsushita (1994).

Graph 8-1 Total Number of Tokens/Lexemes and LEPIX from Texts with *500-4,300* Tokens



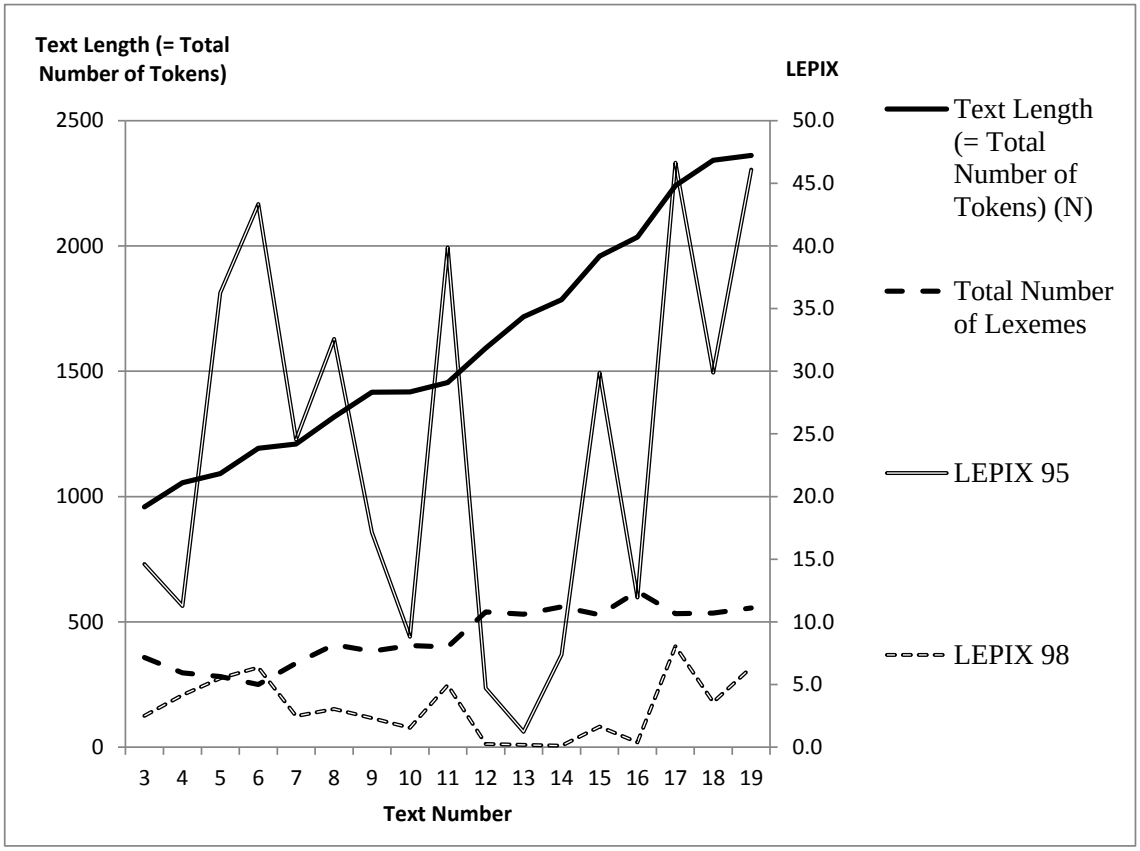
The graph shows that LEPIX $_{95}$ figures correlates with the text length, which should not be.

[100] The way Richards & Malvern (1997) proposed was not applicable to this study as the purpose of the measurement is different. It requires a set of data from one source (a child) at different times.

LEPIX $_{95}$ strongly correlates with the text length at r = .84 (p < .001). LEPIX $_{98}$ also correlates with text length at r = .44 (p < .05). LEPIX is designed to be compared between differently-sized texts to indicate its lexical learning possibility per a unit of text length. In the cases shown above, the length of the longest text is 8 times the shortest one.

After making several attempts with different combinations of texts, I found that differently-sized texts seem to allow comparison if the ratio between the longest and the shortest is within approximately 1:2. Graph 8-2 shows the total number of tokens/lexemes and LEPIX from seventeen differently-sized (959-2,361 tokens) texts (Text number 3 to 19 in Table 8-4).

Graph 8-2 Total Number of Tokens/Lexemes and LEPIX from Texts with *900-2,400* Tokens



Within this range, LEPIX figures fluctuate even when the text length goes up. The correlation coefficient between LEPIX $_{95}$ and the text length is low and not significant at r = .22 (*n.s.*). The correlation coefficient between LEPIX $_{98}$ and the text length is low and not

significant at $r = .06$ (*n.s.*). LEPIX seem to allow comparison when the text length is less than double the other. Ideally, texts should be the same length.

8.7 Remaining Issues

There are some remaining issues. First, if a repeatedly-used essential key word in the text is at the lowest frequency level, the index doesn't work well, because the keyword cannot be deleted but is only counted as one lexeme with many tokens. The number of tokens beyond 95% coverage is fixed by the text length. Therefore, if a word at the target level has many tokens, the number of lexemes will be limited. As a result, LEPIX will not be high. In this case, words within 95% coverage will also be target words, and the Lexical Level of Text will also move to a slightly more basic level. If a word is repeatedly used, the learning effect will also be reduced. For example, the effect of 4 occurrences may have double the effect of 2 occurrences, yet, 20 occurrences will not have double the effect of 10 occurrences. If it is not appropriate to reduce the occurrence of the repeatedly-used word at the target level, setting a cap for the maximum target word occurrence per unit of length for calculating LEPIX may be a solution. It makes the procedure and calculation more complicated, though.

Second, minimum occurrences of target words will differ according to the text length. Twice will be enough for a short text as instructional material, but the minimum occurrence level is not clear for a longer extensive reading text. There are several studies on the minimum occurrence level for incidental vocabulary learning. The results do not agree with each other (Hulstijn et al., 1996; Rott, 1999; Waring & Takaki, 2003; Webb, 2007). For instructional material, there seem few studies, maybe because it depends on the method of teaching.

Third, LEPIX needs validation through empirical study. The possible independent variables to be examined are the set lexical level (95%, 98% or other levels), use

(classroom instruction/extensive reading outside the classroom), minimum occurrence level, and possibly text length. The dependent variable should be the level of acquisition of target words. The formula could be amended by this study.

Last but not least, there are many other factors which have an impact on learning through reading. If the simplification is poorly done, it would deteriorate the text by influencing other factors. How these related factors should be controlled together for modifying a text is a topic for future research.

8.8 Conclusion of Chapter 8

In this chapter, as a sample use of the vocabulary database (VDRJ), I proposed a method of rewriting a reading text to make a better resource for vocabulary learning based on some assumptions. To express the possibility of lexical learning effect numerically, I proposed an index entitled Lexical Learning Possibility Index for a Reading Text (LEPIX). Sample modification and analyses were shown, followed by some issues with using LEPIX. These methods will make it possible to predict and compare the efficiency of second language vocabulary learning with a reading text.

To make a better modification on a short reading text (with 2,000 words or less) as a resource for learning vocabulary, there are two main techniques. 1) Delete or replace one-timers (or the words whose occurrences are less than the set level) at the lowest frequency level in the text, or 2) Make one-timers occur more in the text by adding words or replacing other words with the one-timer. By doing so, the LEPIX figure will be improved. That should mean the text becomes a better resource for vocabulary learning.

Chapter 9 Conclusion

9.1 Important findings

In this thesis, I explored the most efficient learning and teaching order for (Japanese) words as well as how it varies according to the target domain. The most important two chapters will be Chapters 3 and 7. The most efficient order between the groups of Japanese words and a universal method for deciding the most efficient order between the groups of words were shown in Chapter 7. The most efficient order within each group should follow the rankings in the database developed in Chapter 3. Below are the overall flow of this thesis and findings.

After the introduction in Chapter 1, I reviewed relevant previous studies in terms of the rationale for this research. In 2.2, I first confirmed the importance of the word in language processing, especially in reading, and then discussed the idea that text coverage can be the index for learning efficiency, how high a coverage is needed for reading comprehension followed by the cognate effect in processing vocabulary. In 2.3, after briefly introducing the features of the Japanese writing system, I surveyed relevant studies on Japanese in terms of text coverage, the relationship between word origins or parts of speech and register variations. In 2.4, for ordering words in the database, I discussed the importance of dispersion and investigated possible adjusted frequency measures which are combinations of frequency and dispersion. In 2.5, I discussed possible applications of the vocabulary database and word lists to learning, teaching and research from the viewpoints of learner, teacher/course designer and researcher.

Chapter 3 to Chapter 8 are the body of this thesis. In Chapter 3, I developed the Vocabulary Database for Reading Japanese (VDRJ) based on the Balanced Contemporary Corpus of Written Japanese (BCCWJ) 2009 monitor version (NINJAL, 2009). I showed that Juilland's *U* which is a product of frequency and dispersion, is the most suitable index

for the purpose of this study. I also proved that the three different word rankings (the Word Ranking for International Students, the Word Ranking for General Learners and the Word Ranking for General Written Japanese) are valid for the different purposes by examining text coverage in different target corpora. Specific findings in Chapter 3 are as follows.

1) The adjusted frequency measures of *U*, U_{DP} and SFI do not make a significant difference on overall rankings of words.
2) *U* is more sensitive to the salience of frequency of a single domain than U_{DP} and SFI.
3) The result of the multidimensional scaling shows that the ten sub-sections in BCCWJ can be divided into the three categories of the Internet Q&A forum sites (IF), literary works (LW) and the other eight (AD). IF and LW vocabulary will fit the basic and daily-life needs better than AD, while AD contains more academic and formal words than the other two.
4) The word ranking by Juilland's *U* (WWJ) shows that the balanced Contemporary Corpus of Contemporary Japanese (BCCWJ) 2009 monitor version has a formal and written nature.
5) The word rankings WIS/WGL made from VDRJ will work better for learners and teachers than the former Japanese Language Proficiency Test (F-JLPT) word lists since the WIS/WGL provide higher text coverage than F-JLPT lists.
6) The best learning order of words will be different depending on the purpose. WIS will fit for students or academics better than WGL, while WGL will work better for conversation than WIS. WWJ will only fits learners who do not need to learn daily conversation but only need to read (and write) Japanese.

In Chapter 4, based on VDRJ, I investigated lexical features of texts in different media and genres. I claimed that the distribution of word origins and some parts of speech

can be indices for formality/informality. I also investigated the distribution of Chinese cognates in Japanese. Specific findings in Chapter 4 are as follows.

7) Book texts are less biased compared to magazines and newspapers.
8) The POS distribution is a strong index of informality/formality to identify register variations on a continuum. In every genre in VDRJ, the more the proportion for the seven POS including particles and adverbs, the less the proportion for the four POS including suffixes and verbal nouns will be, and vice versa.
9) The number and proportion of assumed known Chinese cognates for Chinese-background learners (CBLs) are estimated to be 30% of the most frequent 5,000 words (i.e. 1,500 words).

In Chapter 5, based on BCCWJ, I developed the Character Database of Japanese (CDJ) and reported the distribution of Japanese characters. Specific findings in Chapter 5 are as follows.

10) The distribution of Japanese characters is not as uneven as words.
11) The character ranking KWJ (the Ranking for Kanji in Written Japanese) show higher correlations with F-JLPT (the former Japanese Language Proficiency Test) Kanji lists and Grades (the Japanese primary school Kanji grades) than frequencies in newspaper texts. KIS (the Ranking for Kanji for International Students) and KGL (the Ranking for Kanji for General Learners) show even higher correlations with F-JLPT Kanji lists and Grades than KWJ.
12) KIS and KGL provide higher text coverage than F-JLPT Kanji lists.
13) The best order of learning Kanji will be different depending on the purpose. KIS will fit for students or academics better than KGL, while KGL will work better for

conversation texts than KIS. KWJ will only fit learners who do not need to learn daily conversation but only need to read (and write) Japanese.

14) The proportions of character types in different genres are considerably different. The proportion of Hiragana or Kanji may be able to be an index for informality.

In Chapter 6, I discussed the discrepancy between the learning order of words and characters. Based on the account of the relationship between text coverage by words and by characters, I argued that the learning burden of Japanese vocabulary may not be as heavy as generally perceived because Japanese vocabulary is not as diverse as shown in the distributed coverage data and because a limited number of Kanji reach the required level of text coverage by words. Specific findings in Chapter 6 are as follows.

15) 63% of the Balanced Contemporary Corpus of Written Japanese (2009 monitor version) texts are covered without Kanji (but more than half of them are function words).
16) To attain 95% coverage, 1,000 Kanji are required; however, some important words are not covered by the most frequent 1,000 Kanji. To cover those words, several hundred other Kanji will be required.
17) Most of high-frequency and mid-frequency Japanese words are composed of limited number of Kanji, therefore, the burden of learning Japanese vocabulary may not be heavy as expected from the text coverage studies, once the learner knows:
    a) the most frequent 1,000 to 1,500 characters.
    b) forms, meanings and compounding rules of Kanji.
    c) metaphors of Kanji compounds.
    d) different readings (e.g. On-reading and Kun-reading) of each Kanji.

In Chapter 7, I first extracted common academic words, limited-academic-domain

words and literary words, and then, evaluated how different vocabulary use is according to genre by examining text coverage and a newly developed index called Text Covering Efficiency (TCE). TCE is the expected number of tokens of a lexeme of a group of words (per million) in a target text. TCE represents the expected return per unit of text length from learning a group of words. TCE is the most important criterion for judging the most efficient learning order of words. I also discussed different lexical features of different text genres by examining TCE. Specific findings in Chapter 7 are as follows.

18) The most efficient learning order of words can be decided by Text Covering Efficiency (TCE) proposed in 7.4.1.2, which is the expected number of tokens of a lexeme of a tested group of words in a test corpus which reflects the learners' needs. The greater the TCE, the more words in the target text likely to be covered by a lexeme of the grouped words. TCE can be compared with a general word frequency per million.
19) 2,541 common academic words (4-domain and 3-domain words) at nine levels in Japanese Common Academic Word List (JAWL) provide remarkably higher text coverage and TCE in academic texts than other types of words. They also provide higher coverage and TCE in academic texts than in non-academic texts.
20) JAWL Ⅰ (559 words, intermediate) words provide high text coverage and TCE in all types of academic texts.
21) Common academic words (AWs) and limited-academic-domain words (LADs) at advanced levels do not provide high text coverage; however, they provide much higher TCE for academic texts than other types of words.
22) Academic vocabulary contains more nouns, verbal nouns, affixes and archaic words than other types of words.
23) Many of the common academic words are used for managing academic information. The meanings of common academic words are highly abstract. Limited-academic-

domain words (2-domain words and 1-domain words) have more concrete meanings.

24) Some combinations of the two domains for the 2-domain words show a particular semantic field, i.e. ‘humanities and arts’ × ‘social sciences’ = ‘history’ (especially political history), ‘social sciences’ × ‘technological natural sciences’ = ‘industry’ and ‘social sciences’ × ‘biological natural sciences’ = ‘social security, medical and nursing’.
25) Only 27 literary words overlap with academic vocabulary. The 27 words account for 1.7% of literary words and 0.5% of academic vocabulary.
26) Academic texts show high TCE for academic vocabulary but low TCE for literary words. In contrast to that, literary texts show moderately high TCE for literary words but low TCE for academic vocabulary. This means that academic and literary texts have totally different lexical features. Domain specificity is stronger in academic texts than in literary texts.
27) Literary words are the words for describing human actions and feelings vividly and effectively. They contain numerous words for body parts and body actions, many modal adverbs, interjections and words for metaphorical expressions.
28) Literary words from the basic to 10K level also provide high coverage and TCE for conversation; however, literary words at 11-15K or above only provide higher TCE for literary texts but not for conversation.
29) Origins of academic and literary words are considerably clearly separated; 3/4 of literary words originate in Japanese while 3/4 of academic vocabulary originate in Chinese. LADs contain more Western-origin words (Gairaigo).
30) Newspaper texts have similar lexical features to social science texts.
31) Natural science texts have more low-frequency words.

In Chapter 8, based on VDRJ, CDJ and the domain-specific word lists, I proposed a method for simplifying a text to make it as efficient a resource as possible for vocabulary

learning. I also developed an index called the Lexical Learning Possibility Index for a Reading Text (LEPIX) to evaluate how efficient a text could be for vocabulary learning. Specific findings in Chapter 8 are as follows.

32) By calculating a newly developed index Lexical Learning Possibility Index for a Reading Text (LEPIX), a reading text can be assessed as a vocabulary learning resource. The LEPIX figure will be improved by 1) replacing one-timers with other words or 2) making one-timers occur more in the text.
33) LEPIX should not be used for comparing a text with another text twice its length.

18) in Chapter 7 is the most important finding as a method for deciding the most efficient learning order of grouped words. Other findings in Chapter 7 specifically refer to the efficient order in learning Japanese vocabulary.

9.2 Implications for language learning and teaching

I am going to mention implications for learning and teaching before referring to methodological and theoretical implications since this study focuses on a practical question: In what order learners should learn Japanese vocabulary? The implications are twofold, one is more or less universal to any language and the other is specific to Japanese.

Practical implications universal to any language are as follows.

1) The method for identifying the most efficient learning order of words. The requirements for such research include a corpus which reflects learner needs, word profiling software such as AntWordProfiler (Anthony, 2009), and a word frequency list, if available, to list all the possible target words. (If the language does not have a space between words, word-segmentation is necessary.) Taking account of domains where

the learner(s) work, group words in an appropriate manner first and count all the lexemes (word families or types) in each group and tokens of each group in the target text, and then Text Covering Efficiency (TCE) can be calculated. A new approach proposed by this study is to learn words in the TCE order as an efficient way for gaining text coverage in the target domain. Within each group, following the frequency order will largely be efficient.

2) For learning texts in a domain with high domain-specificity, learning domain-specific words will be efficient in gaining text coverage, especially at the intermediate level or above. This was specifically examined and made clearer in some academic domains and literary texts in Japanese by checking TCE in this study.

Practical implications specific to Japanese are given below.

3) Among the genres in this study, learning grouped words in the TCE order (Table 7-31) will be most efficient. Within each group, follow the order of the Word Ranking for International Students or the Word Ranking for General Learners depending on the purpose.
4) In particular, learners with academic purposes are expected to gain a high return by learning common academic words (AW) in the Japanese Common Academic Word List (JAWL) Ⅰ and II after learning basic words.
5) For learners who have decided their major, learning limited-academic-domain words (LADs) is also an efficient way, especially for natural science students. (This seems to be also true for other languages (Coxhead & Hirsh, 2007).
6) For reading Japanese newspapers, learning common academic words (AW) and limited-academic-domain words (LADs) in social sciences is particularly efficient. Conversely, for learning these words, newspapers are a good resource.

7) For learning kanji, word-oriented learning such as learning compounding rules and metaphors is particularly important. Especially, learning different words with a particular Kanji and making links between them, including linking between the On-reading (Chinese-origin) and Kun-reading (Japanese-origin) with a Kanji, seems essential. Without these, vocabulary learning burden will increase.
8) The efficient learning orders of words and characters largely agree with each other; however, some low-frequency Kanji are used for high-frequency words while some high-frequency Kanji are not used for high-frequency words. Kanji learning order should be reconsidered by taking account of these cases.

Things I mentioned above may not be practical enough. One of the most direct and practical uses of the outcomes from this study will be the use of VDRJ, the Vocabulary Database for Reading Japanese. This is not the result of the research questions but a product created in the process of the research; however, as I reviewed in 2.5, there are various practical uses for learning, teaching and researching Japanese vocabulary with word lists which can be derived from databases in various different ways.

Firstly, VDRJ is convenient for searching and grouping Japanese words by many different types of criteria. When you teach or learn a Japanese sentence pattern which requires a particular type of words, you can search the group of words quite easily with VDRJ. Part of speech (POS), word origin, reading of the word, frequency will be the frequently used criteria for grouping words. For example, when you teach nominal adjectives (or *Na*-adjectives) for describing situations, possible words for teaching can be searched and ordered in frequency or importance. Many of those nominal adjectives are Chinese-origin (e.g. 健康な 'kenkou-na' (healthy)) or Western-origin words, most of which have different levels of difficulty for learners with different language backgrounds. These words can be sorted out with word origins by sorting or filtering function of the

database, in conjunction with other criteria such as frequency. Some nominal adjectives (*na*-adjectives) ending with *-i* often causes confusion to elementary learners as they look like an adjective (*i*-adjective). Possible confusing words can also be easily searched by the database.

Secondly, the database can derive various different baseword lists for lexical profiling using a word profiler (e.g. AntWordProfiler; Anthony, 2009). This makes possible to check the vocabulary load of material texts and to analyse learner vocabulary use. One example is shown in Chapter 8. This is a detailed analysis of texts; however, it is commonly easy to check the vocabulary level of a text using the baseword lists (compiled in the accompanying CD) and a morphological analyser with a dictionary (e.g. MeCab (Kudo, 2009a) with UniDic (Den et al., 2009)) . This is extremely useful. When teachers use some authentic materials for advanced or intermediate learners, you can order the texts by lexical load of the texts. If you check the learner's vocabulary level by a vocabulary test where the test items are sampled from the same database as the database used for checking the vocabulary load of the texts, it will be easy to judge whether the text is at an appropriate level for the learner or not. It has made possible to answer the questions: Where are the 95% and 98% text coverage points in the target text? What words will be the target words to learn in the text? The analysis of texts can also be applied to the analysis of learner language as well. This is exactly the idea for checking the productive knowledge of learner vocabulary by Lexical Frequency Profiling (Laufer, 1994). With VDRJ, learner vocabulary can also be checked by different word origins and frequency levels.

Thirdly, as is overlapped with the first and the second points, the database contributes to learning and teaching vocabulary in a specific domain. Domain-specific words can be identified in some domains using the database. As shown in Chapter 7, academic vocabulary and literary vocabulary are already extracted and marked in the database. Also, the database shows the standardized frequency in each of the ten domains shown in Table

3-2, 3-4 and 3-5, so that the words can be reordered by the frequency in any domain among the ten domains. The same things can be done on Kanji by the Character Database of Japanese developed in Chapter 6, though the Kanji domain-specificity, as discussed in Chapter 6, is not as distinct as vocabulary.

Fourthly, any word list discussed above can also be served as a self-check list for learners. As mentioned in 2.5.2, previous studies show that self-directed vocabulary learning is important for learning a language in general. Word lists will contribute to this point.

Last but not least, the database is useful for developing language tests, especially vocabulary tests. Vocabulary tests are not only useful for judging whether a text is at a suitable level for a particular group of learners, but also useful for self-checking the vocabulary level by learners themselves. Matsushita (2012) has already developed a Japanese vocabulary size test for reading Japanese based on VDRJ. Matsushita (2011b) claims the usefulness of the feedback from the vocabulary test and how the feedback should be based on the trial version of the Japanese vocabulary size test.
Akiyama & Matsushita (2012) have developed a computer-adaptive version of the Japanese vocabulary size test. One of the strengths of the computer-adaptive test (CAT) is that it can be repeatedly used by a testee as it provides different test items based on the testee's answers to estimate the ability. Developing a web-based CAT is expected for self-checking the vocabulary level from time to time to see the progress which facilitates learner autonomy.

These five points on the use of the database and word lists will also be this study's major contribution for learning, teaching and researching Japanese vocabulary.

9.3 Methodological and theoretical implications

The theoretical implications are also twofold. The theoretical implications universal

to any language are as follows. (Numbers follow the previous section.)

9) Juilland's *U*, which is a product of frequency (*F*) and dispersion (*D*), tends to downgrade words with uneven distribution to a single domain more than other adjusted frequency measures, as Juilland's *D* is more sensitive to the salience of frequency of a single domain than other dispersion measures. Therefore, it is suitable to adjust the ranking for unevenly distributed words with sampling bias.
10) According to the purpose, word rankings can be developed and improved by weighting frequencies from different genres.
11) The method for extracting domain-specific words and the word tier analysis by Text Covering Efficiency (TCE) are applicable to any language. TCE is a simple and powerful index by which differently-sized texts in different genres can be compared and the results are easy to interpret.
12) The Lexical Learning Possibility Index for a Reading Text (LEPIX) still needs to be improved; however, the method for simplifying a text and assessing it with LEPIX is applicable to any language.

The theoretical implications specific to Japanese are shown below.

13) Ito (2002) claims that the proportion of Japanese-origin words is a better index for register variation than the proportion of Chinese-origin words; however, it is not always true. For measuring formality, the proportion of Chinese-origin words will be a better index.
14) Kabashima's law (Kabashima, 1955, 1981) is not always true. The proportion of conjunctions will indicate the level of formality as well as verbal nouns and affixes.

As introduced and discussed in 2.3.4 and 4.4.4, these are the findings which refer to how lexical features of texts are related to register variations.

9.4 Directions for further research

There are various directions for further research. First, the vocabulary databases (VDRJ and CDJ) themselves should be further refined. There still remain incorrect data with wrongly-analysed items in VDRJ. Corrections were made to the top 20,000 words; however, there will still be incorrect items. Items beyond them should be further refined. For users' convenience, the databases should be improved by attaching frequent example words, phrases and sentences hopefully with a concordance function on the web-site with a user-friendly interface.

Second, for analysing vocabulary load and lexical features of texts, it is particularly desirable to develop a system which calculates indices such as dispersion, adjusted frequency, TCE and LEPIX automatically by setting a target text and relevant baseword lists. It requires collaboration with researchers and technical staff in information science.

Third, researching vocabulary use in spoken Japanese is necessary. For creating VDRJ, I had to make a compromise for setting basic words by partly adopting the former Japanese Language Proficiency Test word lists, just because we do not have any good spoken corpus for creating a frequency list. Building up a spoken corpus which reflects the language use in learners' domains based on needs analysis is indispensable for developing teaching Japanese as a second language.

Fourth, developing vocabulary tests such as a vocabulary size test and validation of them are needed[101]. By measuring learners' vocabulary knowledge and checking the results with lexical analysis of target texts, we are able to design a curriculum which suits learners'

[101] In fact, I developed Vocabulary Size Test for Reading Japanese and collected data for validation; however, I could not include the outcomes in this thesis due to various reasons. I would like to validate and improve the test for future use.

levels better.

Fifth, developing a cognate database is useful for investigating the cognate effect in more depth. Various types of information on differences and similarities can be included in it. Diagnostic tests to measure the learners' sensitivity and usability of their first language knowledge will also be useful for Chinese-background and English-background learners.

Sixth, developing more domain-specific word lists and refining them is desirable. Technical term lists in more specific domains will be needed. Literary words can also be elaborated by classifying literary texts into different literary genres. It is also important to explore how these domain-specific words function in a text.

Last but not least, specific applications of these databases and word lists to learning and teaching should be further explored. Otherwise, these databases and word lists are useless.

Learning a second language is like swimming in the Vocabulary Sea. I would like to continue to build islands and bridges, and throw a rope with an emergency ring in the sea for learners so they will not drown.

References

*For readers' convenience, Japanese/Chinese authors' names in Roman alphabet are put first followed by their original names in the brackets. e.g., Nozaki, H. (野崎浩成)

*If a Japanese/Chinese organization is the author or editor and has its English name, it is put first followed by its original name in the brackets.

e.g., Japan Student Services Organization (独立行政法人　日本学生支援機構)

*If a title of a Japanese/Chinese article, journal or book has its English title; it is put in round brackets. If not, I translate the title and put it in square brackets.

Agency for Cultural Affairs (文化庁). (1978). *中国語と対応する漢語 [Chinese-origin Words Correspondent to Chinese Words]*. Tokyo: National Printing Bureau of the Ministry of Finance (大蔵省印刷局).

Agency for Cultural Affairs （文化庁）. (2010). 改訂常用漢字表 [Revised list of common Japanese Kanji]. Downloaded from http://www.bunka.go.jp/oshirase_other/2010/kaitei_jyoyokanji_nyusyu.html

Akimoto, M. （秋元美晴）. (2002). *よくわかる語彙 [Understanding Vocabulary]*. Tokyo: Alc （アルク）.

Akimoto, M., & Oshio, K. （秋元美晴・押尾和美）. (2008). 新しい日本語能力試験のための語彙表・漢字表作成中間報告　－新語彙表 Ver. III の完成まで－ (An interim report on developing new word lists and Kanji lists for the new Japanese Language Proficiency Test: Up to the competion of the new word list version III). *日本語学 [Japanese Linguistics]*, *27*(10), 36–49.

Akiyama, M., & Matsushita, T. (秋山 實・松下達彦）. (2012). 潜在ランク理論に基づくコンピュータ適応型テストシステムの開発と日本語語彙サイズテストへの適用　―シミュレーションによる評価― [Developing computer-adaptive test based on the latent-rank theory and its application to a Japanese vocabulary size test: Evaluation by simulation]. Presented at the The 16th Annual Conference of the Japan Language Testing Association, Senshu University, Tokyo.

Amano, S., & Kondo, T. （天野成昭・近藤公久）. (1999). *日本語の語彙特性　第1期 (Lexical Features of Japanese, 1st Period)*. Tokyo: Sanseido （三省堂）.

Amano, S., & Kondo, T. （天野成昭・近藤公久）. (2000). *日本語の語彙特性　第2期 (Lexical Features of Japanese, 2nd Period)*. Tokyo: Sanseido （三省堂）.

Anthony, L. (2007). *AntConc Version 3.2.1 (text analysis tool)*. Downloaded from http://www.antlab.sci.waseda.ac.jp/software.html

Anthony, L. (2009). *AntWordProfiler Version 1.2w (word profiler)*. Downloaded from http://www.antlab.sci.waseda.ac.jp/software.html

Arakawa, K. (荒川清秀). (1979). 中国語と漢語　－文化庁「中国語と対応する漢語」の評を兼ねて [Chinese langauge and Chinese-origin words: Review of "Chinese-origin Words

Correspondent to Chinese Words" by Agency for Cutural Affairs]. *愛知大学文学論叢 [Literature Forum of Aichi University]*, *62*, 1–28.

Araya, T. （荒屋 勤）. (1983). 日中同形語 (Chinese cognates in Japanese). *大東文化大学紀要 人文科学 [Bulletin of Daito Bunka University: Humanities]]*, *21*, 17–29.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics with R*. Cambridge: Cambridge University Press.

Balota, D. A., & Spieler, D. H. (1998). The utility of item level analyses in model evaluation: a reply to Seidenberg and Plaut. *Psychological Science*, *9*(3), 238–240.

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography*, *6*(4), 253–279.

Beck, I. L., & McKeown, M. G. (1985). Teaching vocabulary: Making the instruction fit the goal. *Educational Perspectives*, *23*(1), 11–15.

Beck, I. L., McKeown, M. G., & Kucan, L. (2002). *Bringing Words to Life: Robust Vocabulary Instruction*. Solving problems in the teaching of literacy. New York: Guilford Press.

Beglar, D. (2010). A Rasch-based validation of the Vocabulary Size Test. *Language Testing*, *27*(1), 101–118. doi:10.1177/0265532209340194

Bernhardt, E. (2005). Progress and procrastination in second language reading. *Annual Review of Applied Linguistics*, *25*, 133–150. doi:10.1017/S0267190505000073

Biber, D. (1988). *Variation Across Speech and Writing*. Cambridge [England] ; New York: Cambridge University Press.

Biber, D. (1995). *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge ; New York: Cambridge University Press.

BLI (Research Institute for Language Learning and Teaching, Beijing Language Institute) (北京语言学院语言教学研究所). (1986). *现代汉语频率词典 (A Word Frequency Dictionary for Modern Chinese)*. Beijing: 北京语言学院出版社 (Beijing Language Institute Press).

Brysbaert, M., & New, B. (2009). Moving beyond Kucera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods*, *41*(4), 977.

Butler, Y. G. （バトラー後藤裕子）. (2010). 小中学生のための日本語学習語リスト（試案） (A list of Japanese academic vocabulary for elementary and junior high school students in Japan). *母語・継承語・バイリンガル教育(MHB)研究 (Studies in Mother Tongue, Heritage Language, and Bilingual Education)*, *6*, 42–58.

Carroll, J. B. (1970). An alternative to Juilland's usage coefficient for lexical frequencies, and a proposal for Standard Frequency Index (SFI). *Computer Studies in the Humanities and Verbal Behavior*, *3*(2), 61–65.

Carroll, J. B. (1971). Statistical analysis of the corpus. *Word Frequency Book* (p xxi–xl). New York: Houghton Mifflin, Boston American Heritage.

Carroll, J. B., Davies, P., & Richman, B. (1971). *Word Frequency Book*. New York: American Heritage.

Castellví, M. T. C. (2003). Theories of terminology Their description,prescription and explanation. *Terminology*, *9*(2), 163–199.

Chen, Y. (陳 毓敏). (2009). 中国語学習者の日本語の漢字語習得研究のための新たな枠組みの提案　－意味使用の一般性と意味推測可能性を考慮して－ (A new framework for acquisition of Japanese kanji compounds targetting Chinese learners of Japanese: in consideration of genaral semantic usage and semantic inferability). *日本語科学 (Japanese Linguistics)*, *25*, 105–117.

Chen, Z., & Henning, G. (1985). Linguistic and cultural bias in language proficiency tests. *Language Testing*, *2*, 155–163.

Chikamatsu, N. (1996). The effects of L1 orthography on L2 word recognition: A study of American and Chinese learners of Japanese. *Studies in Second Language Acquisition*, *18*(4), 403–432. doi:10.1017/S0272263100015369

Chikamatsu, N., Yokoyama, S., Nozaki, H., Long, E., & Fukuda, S. (2000). A Japanese logographic character frequency list for cognitive science research. *Behavior Research Methods, Instruments, & Computers*, *32*(3), 482–500. doi:10.3758/BF03200819

Chiu, H. （邱學瑾）. (2002). 漢字圏・非漢字圏日本語学習者における漢字熟語の処理過程 : 意味判断課題を用いた形態・音韻処理の検討 (Processing orthography and phonology in semantic decision tasks: Processing of Japanese Kanji words by learners of Japanese as a second language. *教育心理学研究 (The Japanese Journal of Educational Psychology)*, *50*(4), 412–420.

Chomsky, N. (1965). *Aspects of the theory of syntax*. Special technical report (Massachusetts Institute of Technology. Research Laboratory of Electronics) (Vol. no. 11). Cambridge, Mass.: M.I.T. Press.

Chujo, K. （中條和光）. (1983). 日本語単文の理解過程–文理解ストラテジ-の相互関係. *心理学研究*, *54*(4), p250–256.

Chujo, K., & Utiyama, M. (2006). Selecting level-specific specialized vocabulary using statistical measures. *System*, *34*, 255–269.

Chung, T. M. (2003a). *Identifying technical terms*. Unpublished PhD dissertation, Victoria University of Wellington.

Chung, T. M. (2003b). A corpus comparison approach for terminology extraction. *Terminology*, *9*(2), 221–245.

Chung, T. M., & Nation, P. (2003). Technical vocabulary in specialised texts. *Reading in a Foreign Language*, *15*(2), 103–116.

Cobb, T. (1996). *From concord to lexicon: development and test of a corpus-based lexical tutor* (PhD thesis). Concordia University,, Montreal.

Cobb, T. (2000). One size fits all? Francophone learners and English vocabulary tests. *Canadian Modern Language Review*, *57*(2), 295–324.

Cobb, T. (2007). Computing the vocabulary demands of L2 reading. *Language Learning and Technology*, *11*(3), 38–63.

Cohen, A. D. (1990). *Language Learning: Insights for Learners, Teachers, and Researchers*. New York, NY: Newbury House Publishers.

Corson, D. J. (1985). *The Lexical Bar*. Oxford: Pergamon Press.

Corson, D. J. (1997). The learning and use of academic English words. *Language Learning*, *47*(4), 671–718.

Coxhead, A. (1998). *An Academic Word List*. LALS Occasional Publication Number 18. Wellington: Victoria University of Wellington.

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

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Francaise de Linguistique Appliquee*, *12*(2), 65–78.

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why might secondary science textbooks be difficult to read? *New Zealand Studies in Applied Linguistics*, *16*(2), 37–52. doi:Article

Daulton, F. E. (1998). Japanese loanword cognates and the acquisition of English vocabulary. *The Language Teacher*, *22*(1), 17–25.

Daulton, F. E. (2004). *Gairaigo -- The Built-in Lexicon? -The Common Loanwords in Japanese Based-on High-frequency English Vocabulary and Their Effect on Language Acquisition* (Unpublished Doctoral Dissertation). Victoria University of Wellington, Wellington, New Zealand.

de Groot, A. M. B., & Keijzer, R. (2000). What is hard to learn is easy to forget: the roles of word concreteness, cognate statuus, and word frequency in foreign-language vocabulary learning. *Language Learning*, *50*(1), 1–56.

Deming, W. E. (1994). *The New Economics for Industry, Government, Education* (2nd ed.). Cambridge, MA: Massachusetts Institute of Technology, Center for Advanced Educational Services.

Den, Y., Ogiso, T., Ogura, H., Yamada, A., Minematsu, N., Uchimoto, K., & Kioso, H. (伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵). (2007). コーパス日本語学のための言語資源　ー形態素解析用電子化辞書の開発とその応用ー (The development of an electric dictionary for morpholigical analysis and its application to Japanese corpus linguistics). *日本語科学 (Japanese Linguistics)*, *22*, 101–123.

Den, Y., Yamada, A., Ogura, H., Koiso, H., & Ogiso, T. （伝 康晴・山田 篤・小磯花絵・小木曽智信）. (2009). *UniDic (digitized dictionary for morphological analysis) 1.3.11*. Downloaded from http://www.tokuteicorpus.jp/dist/

Den, Y. （伝康晴）, Yamada, A. （山田篤）, Ogura, H. （小椋秀樹）, Koiso, H. （小磯花絵）, & Ogiso, T. （小木曽智信）. (2009). *UniDic*. Downloaded from http://www.tokuteicorpus.jp/dist/

Dijk, T. A. van, & Kintsch, W. (1983). *Strategies of discourse comprehension*. New York: Academic Press.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist.*, *19*(1), 61–74.

Eaton, H. S. (1940). *An English-French-German-Spanish Word Frequency Dictionary*. New York: Dover Publications.

Elley, W. B., & Mangubhai, F. (1981). The long-term effects of a book flood on children's language growth. *Directions*, *7*, 15–24.

Elley, W. B., & Mangubhai, F. (1983). The Impact of Reading on Second Language Learning. *Reading Research Quarterly*, *19*(1), 53–67. doi:10.2307/747337

Folse, K. (2011). Applying L2 Lexical Research Findings in ESL Teaching. *TESOL Quarterly*, *45*(2), 362. doi:10.5054/tq.2010.254529

Fudano, H., & Fukasawa, N. （札野寛子・深澤のぞみ）. (1995). 理工系学生を対象とした実験・研究に必要な日本語指導のための語彙表現研究　ー「科学技術基礎日本語」教材開発に向けてー [A study on vocabulary for science students' experiments and research: Towards developing a learning material for the 'basic Japanese for science and technology']. *平成 7 年度 日本語教育学会春季大会 予稿集 [Proceedings for the Conference of the Society for Teaching Japanese as a Foreign Language, Spring 1995]* (p 186–191).

Fukao, Y. （深尾百合子）. (2001). 「専門日本語教育研究」の現状と展望 [Studies on teaching technical Japanese: Present and future]. *2001 年度 日本語教育学会秋季大会予稿集 [Proceedings for the Conference of the Society for Teaching Japanese as a Foreign Language, Autumn 2001]* (p 233–234).

Gairns, R., & Redman, S. (1986). *Working with Words*. Cambridge: Cambridge University Press.

Gardner, D., & Hansen, E. C. (2007). Effects of lexical simplification during unaided reading of English informational texts. *TESL Reporter*, *40*(2), 27–59.

Ghadirian, S. (2002). Providing controlled exposure to target vocabulary through the screening and arranging of texts. *Language Learning and Technology*, *6*(1), 147–164.

Gilquin, G., & Gries, S. T. (2009). Corpora and experimental methods: A state-of-the-art review. *Corpus Linguistics and Linguistic Theory*, *5*(1), 1–26. doi:10.1515/CLLT.2009.001

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. *International Journal of Corpus Linguistics*, *13*(4), 403–437. doi:10.1075/ijcl.13.4.02gri

Gries, S. T. (2010). Dispersions and adjusted frequencies in corpora: further explorations. *Language & Computers*, *71*(1), 197–212.

Gu, Y., & Johnson, R. K. (1996). Vocabulary learning strategies and language learning outcomes. *Language Learning*, *46*(4), 643–679.

Gu, Y. P. (2003). Vocabulary learning in a second language: Person, task, context, and strategies. *TESL-EJ*, *7*(2), 1–31.

Harlan, P. （パトリック・ハーラン）Translated by T. M. (2011). ゼロからの日本語学習と僕の好きな日本のカルチャー (Learning Japanese from zero, and the Japanese culture I like). Cited from http://www.wochikochi.jp/topstory/2011/04/packun.php

Hatasa, Y. A. (1992). *Transfer of the Knowledge of Chinese Characters to Japanese* (Unpublished Doctoral Dissertation). University of Illinois at Urbana-Champaign, Illinois.

Hida, Y., & Ro, G. （飛田良文・呂玉新）. (1987). *日本語・中国語意味対照辞典 [Contrastive Dictinary for Meanings of Chinese Cognates in Japanese]*. Tokyo: Nan'un-do （南雲堂）.

Hirsh, D., & Nation, P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language*, *8*(2), 689–696.

Hirsh, David. (2004). *A functional representation of academic vocabulary*. Unpublished Doctoral Thesis, Victoria University of Wellington, School of Linguistics and Applied Language Studies, Wellington.

Hishinuma, T. (菱沼 透). (1984). 中国の標準字体と日本の通用字体 [The standard character form in Chinese and the common character forms in Japanese]. *日本語学 [Japanese Linguistics]*, *3*(3), 32–40.

Hitosugi, C. I., & Day, R. R. (2004). Extensive reading in Japanese. *Reading in a Foreign Language*, *16*(1), 20–39.

Honeyfield, J. (1977). Simplification. *TESOL Quarterly*, *11*(4), 431–440.

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

Hulstijn, J. H. (2001). Intentional and incidental second-language vocabulary learning: a reappraisal of elaboration, rehearsal and automaticity. P. Robinson (Ed.), *Cognition and Second Language Instruction*. Cambridge: Cambridge University Press.

Hulstijn, J. H., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: the influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *Modern Language Journal*, *80*(3), 327–339.

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

Institute of JUSE (日本科学技術研修所). (2010). ２つの相関係数の差の検定について [The test for the gap between two correlation coefficients]. Cited from http://www.i-juse.co.jp/statistics/xdata/faq-11665.pdf

Ishiwata, T. （石綿敏雄）. (1970). 日本語研究の問題　ー計量語い論 [Issues with studies in Japanese language: Quantitative lexicology]. *講座日本語と日本語教育*, 早稲田大学語学教育研究所, *6*, 84–100.

Ito, M. (伊藤雅光). (2002). 語彙の量的性格 [Quantitative characteristics of vocabulary]. *朝倉日本語講座 4: 語彙・意味 [Asakura lecture series on Japanese language 4: vocabualry and meaning]*. (p 29–53). Tokyo: Asakura Publishing.

Iwabuchi, E. （岩淵悦太郎）. (1970). *現代日本語：ことばの正しさとは何か [Modern Japanese: What is the Rightness of Language?]*. Tokyo: Chikumashobo （筑摩書房）.

Japan Foundation, & Association of International Education, Japan （国際交流基金・日本国際教育協会） (Ed.). (2002). *日本語能力試験出題基準【改訂版】 [The Standards for Japanese Language Proficiency Test]*. Tokyo: Bonjinsha （凡人社）.

JASSO (Japan Student Services Organization). (2010). *International Students in Japan 2010*. Cited from http://www.jasso.go.jp/statistics/intl_student/data10_e.html#no7

Jiang, N. (2000). Lexical representation and development in a second language. *Applied Linguistics, 21*(1), 47–77.

JSPS (Japan Society for the Promotion of Science). (2010a). List of categories, areas, disciplines and research fields. Cited from http://www.jsps.go.jp/english/e-grants/data/09_2010/05_1_e.pdf

JSPS (Japan Society for the Promotion of Science). (2010b). Appendix table of keywords. Cited from http://www.jsps.go.jp/english/e-grants/data/09_2010/05_1_e.pdf

Juilland, A. G., Brodin, D. R., & Davidovitch, C. (1970). *Frequency Dictionary of French Words*. The Romance languages and their structures. First ser. The Hague: Mouton.

Juilland, A. G., & Chang-Rodrigues, E. (1964). *Frequency Dictionary of Spanish Words*. London: Mouton & Co.

Kabashima, T. (樺島忠夫). (1955). 類別した品詞に見る規則性 [Regularity in classified part of speech]. *国語国文 [Japanese Language and Literature], 250*, 385–387.

Kabashima, T. (樺島忠夫). (1981). *日本語はどう変わるか　ー語彙と文字ー [How Will the Japanese Language Change? Vocabulary and Characters]*. Tokyo: Iwanami Shoten (岩波書店).

Kai, M. （甲斐睦朗）. (2000). *日本語基本語彙　ー文献解題と研究 (Study of Vocabulary Lists of Basic Japanese Vocabulary: Commentary and Research)*. (NLRI (The National Language Research Institute) (国立国語研究所), Ed.). Tokyo: Meiji Shoin （明治書院）.

Kai, M. （甲斐睦朗）. (2002). 現代日本語の基本語彙 [Basic vocabulary of modern Japanese]. Y. Hida & T. Sato （飛田良文・佐藤武義） (Eds.), *語彙 [Vocabulary]*, 現代日本語講座　[Modern Japanese Linguistics Course] (Vol. 4, p 25–45). Tokyo: Meiji Shoin （明治書院）.

Kano, C. （加納千恵子）. (1994). 漢字教育のためのシラバス案 (A proposed syllabus for Kanji teaching). *筑波大学留学生センター日本語教育論集 (Journal of Japanese Language Teaching, Interenational Student Center, University of Tsukuba), 9*, 41–50.

Kashiwano, W., Maruyama, T., Inamasu, S., Tanaka, Y., Akimoto, Y., Sano, H., Ooyauchi, Y.他. (2009). *「現代日本語書き言葉均衡コーパス」における収録テキストの抽出手順と事例 [Procedure and Examples of the Extraction of Texts in the Balanced Corpus of Contemporary Written Japanese]*. 特定領域研究「日本語コーパス」データ班 (General Headquarters, Priority-Area Research "Japanese Corpus"). Tokyo: NINJAL (National Institute for Japanese Language).

Kato, T. (加藤稔人). (2005). 中国語母語話者による日本語の漢語習得　ー他言語話者との習得過程の違いー (Acquisition of Japanese kanji compaounds by Chinese native speakers: differences in the acquisition process from speakers of other languages). *日本語教育 (Journal of Japanese Language Teaching), 125*, 96–105.

Kawamura, Y. （川村よし子）. (2006). 日本語学習者のための基本語選定の一試案 [A proposal for selecting fundamental vocabulary for learners of Japanese]. *第 11 回ヨーロッパ日本語教育シンポジウム (The 11th Europen Symposium on Japanese Language Education)*.

Kawamura, Y., Kitamura, T., & Hobara, R. (1997). Reading Tutor (リーディング・チュー太). Cited from http://language.tiu.ac.jp/index_e.html

Kayamoto, Y. （茅本百合子）. (1995). 同一漢字における中国語音と日本語音の音読みの類似度に関する調査 (Similarities and differences between readings of Chinese characters and On-readings of Japanese Kanji). *広島大学日本語教育学科紀要 [Bulletin of the Department of Teaching Japanese as a Second Language, Hiroshima University]*, *5*, 67–75.

Kayamoto, Y. （茅本百合子）. (2000). 日本語を学習する中国語母語話者の漢字の認知　－上級者・超上級者の心内辞書における音韻情報処理－ (Processing phonological information: Recognition of Japanese characters by advanced- and superior-level native speakers of Chinese). *教育心理学研究 (Japanese Journal of Educational Psychology)*, *48*, 315–322.

Kayamoto, Y. （茅本百合子）. (2002). 語彙判断課題と命名課題における中国語母語話者の日本語漢字アクセス (Lexical access to Japanese Kanji by native speakers of Chinese: Evidence from lexical decision and naming tasks). *教育心理学研究 (Japanese Journal of Educational Psychology)*, *50*(4), 436–445.

Kim, E. （金愛蘭）. (2011). 20 世紀後半の新聞語彙における外来語の基本語化 (Shift of loanwords to basic words in Japanese newspapers published in the second half of the 20th century). *Handai Nihongo Kenkyu （Studies in Japanese Language, Osaka University, 阪大日本語研究), Separate 3*, 1–175.

Kin, J. （金若静）. (1987). *同じ漢字でも [Even if the Same Kanji are Used]*. Tokyo: Gakuseisha （学生社）.

Kin, J. （金若静）. (1990). *続・同じ漢字でも [Even if the Same Kanji are Used, Second Series]*. Tokyo: Gakuseisha （学生社）.

Kindaichi, H. （金田一春彦）. (1981). *日本語の特質 [Features of Japanese Language]*. Tokyo: NHK Publishing （日本放送出版協会）.

Kindaichi, H. （金田一春彦）. (1988). *日本語　新版 [The Japanese Language: New version]*. Tokyo: Iwanami Shoten （岩波書店）.

Koda, K. (1989). The Effects of Transferred Vocabulary Knowledge on the Development of L2 Reading Proficiency. *Foreign Language Annals*, *22*(6), 529–540. doi:10.1111/j.1944-9720.1989.tb02780.x

Kojic-Sabo, I., & Lightbown, P. M. (1999). Student's Approaches to Vocabulary Learning and Their Relationship to Success. *The Modern Language Journal*, *83*(2), 176–192.

Komiya, C. （小宮千鶴子）. (1995). 専門日本語教育の専門語　－経済の基本的な専門語の特定を目指して－ [Technical terms for teaching technical Japanese: Aiming at identifying basic technical terms for economics]. *日本語教育 (Journal of Japanese Language Teaching)*, *86*, 81–92.

Komiya, C. （小宮千鶴子）. (2005). 理工系留学生のための化学の専門語–高校教科書の索引調査に基づく選定 (Basic chemistry vocabulary for international students: A selection culled from indices of high school textbooks). *専門日本語教育研究 (Journal of technical Japanese education)*, (7), 29–34.

Komori, K. （小森和子）. (2005). 第二言語としての日本語の文章理解における第一言語の単語認知処理方略の転移 : 視覚入力と聴覚入力の相違を中心に (The transfer of L1 cognitive orthographic strategies into the text comprehension of Japanese as L2 : Processing differences between visual information and phonological Information). *横浜国立大学留学生センター紀要 (Journal of International Student Center, Yokohama National University), 12*, 17–39.

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.

Kroll, J. F., & Stewart, E. (1994). Category Interference in Translation and Picture Naming: Evidence for Asymmetric Connections Between Bilingual Memory Representations. *Journal of Memory and Language*, *33*(2), 149–174. doi:10.1006/jmla.1994.1008

Kudo, T. (工藤 拓). (2009a). *MeCab (morphological analyzer) 0.98*. Downloaded from http://mecab.sourceforge.net/

Kudo, T. (工藤 拓). (2009b). 日本語解析ツール MeCab, CaboCha の紹介 [Introduction of MeCab and CaboCha, the analysis tools of Japanese]. Cited from http://chasen.naist.jp/chaki/t/2009-09-30/doc/mecab-cabocha-nlp-seminar-2009.pdf

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? Lauren, C. and M. Nordman (Eds.), *Special Language: From Humans Thinking to Thinking Machines*. Clevedon: Multilingual Matters.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? P. J. L. Arnaud & H. Bejoint (Eds.), *Vocabulary and Applied Linguistics* (p 126–132). London: Macmillan.

Laufer, B. (1994). The lexical profile of second language writing: does it change over time? *RELC Journal*, *25*(2), 21–33.

Laufer, B., & Hulstijn, J. (2001). Incidental vocabulary acquisition in a second language: the construct of task-induced involvement. *Applied Linguistics*, *22*(1), 1–26.

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.

Laufer, B., & Shmueli, K. (1997). Memorizing new words: Does teaching have anything to do with it? *RELC Journal*, *28*(1), 89 –108. doi:10.1177/003368829702800106

Leech, G., Rayson, P., & Wilson, A. (2001). *Word Frequencies in Written and Spoken English*. Harlow: Pearson Education.

Levelt, W. J. M. (1989). *Speaking*. ACL-MIT Press series in natural-language processing. MIT Press.

Levelt, W. J. M. (1993). The architecture of normal spoken language use. G. Blanken, J. Dittmann, H. Grimm, J. C. Marshall, & C. W. Wallesch (Eds.), *Linguistic disorders and pathologies: An international handbook* (p 1–15). Berlin: de Gruyter.

Long, E., & Yokoyama, S. (2005). Tex genre and Kanji frequency. *Corpus Studies on Japanese Kanji*, Glottometrics (Vol. 10). Tokyo, Japan/Lüdenscheid, Germany: Hituzi Syobo/RAM-Verlag.

Lotto, L., & de Groot, A. M. B. (1998). Effects of learning method and word type on acquiring vocabulary in an unfamiliar language. *Language Learning, 48*(1), 31.

Lu, B. （鲁宝元）. (2000). 汉日同形词对比研究与对日汉语教学 [Contrastive study on Chinese cognates in Japanese and teaching Chinese to Japanese students]. International exchange institute, Beijing University of Foreign Studies (北京外国语大学国际交流学院) (Ed.), *汉日语言研究文集* (Vol. 3). Beijing: Beijing Publishing （北京出版社）.

Lyne, A. A. (1985). *The Vocabulary of French Business Correspondence: Word Frequencies, Collocations and Problems of Lexicometric Method*. Genève: Slatkine.

Machida, S. (2001). Japanese text comprehension by Chinese and non-Chinese background learners. *System, 29*(1), 103–118. doi:10.1016/S0346-251X(00)00048-8

Maruyama, T. （丸山岳彦）. (2009a).「現代日本語書き言葉均衡コーパス」モニター公開データ（2009 年度版）書誌情報・サンプル情報・著者情報について [On the information of books, samples and authors for the Balanced Corpus of Contemporary Written Japanese 2009 monitor version]. *「現代日本語書き言葉均衡コーパス」2009 年モニター版 [Balanced Corpus of Contemporary Written Japanese 2009 monitor version]*. Tokyo: NINJAL (National Institute for Japanese Language).

Maruyama, T. （丸山岳彦）. (2009b).「現代日本語書き言葉均衡コーパス」モニター公開データ（2009 年度版）サンプリング方法について [Sampling method for the Balanced Corpus of Contemporary Written Japanese 2009 monitor version]. *「現代日本語書き言葉均衡コーパス」2009 年モニター版 [Balanced Corpus of Contemporary Written Japanese 2009 monitor version]*. Tokyo: NINJAL (National Institute for Japanese Language).

Matsunaga, S. (1999). The role of Kanji knowledge transfer in acquisition of Japanese as a foreign language. *世界の日本語教育. 日本語教育論集 (Japanese-language Education around the Globe), 9*, 87–100.

Matsushita, T. (2012).「日本語を読むための語彙量テスト」の開発」[Developing the 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 the International Conference on Japanese Language Education (ICJLE) Nagoya 2012, Vol. 1]*, 310.

Matsushita, T. （松下達彦）. (2009). マクロに見た常用漢字語の日中対照研究　―データベース開発の過程から― [A macro study of meanings and usages of the common Japanese Kanji vocabulary in contrast to Chinese: findings from the process of development of a database]. *桜美林言語教育論叢 [Obirin Forum of Language Education], 5*, 117–131.

Matsushita, T. （松下達彦）. (2010). 日本語を読むために必要な語彙とは？ －書籍とインターネットの大規模コーパスに基づく語彙リストの作成－ [What words are essential to read Japanese? Making word lists from a large corpus of books and internet

forum sites]. *2010 年度日本語教育学会春季大会予稿集 [Proceedings for the Conference of the Society for Teaching Japanese as a Foreign Language, Spring 2010]*.

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. （松下達彦）. (2011b). 自律的な語彙学習を促す語彙テストのフィードバック [Vocabulary test feedback for facilitating autonomous vocabulary learning]. *平成 23 年度日本語教育学会第３回研究集会　予稿集 [Proceedings for 2011 3rd Research Meeting of the Society for Teaching Japanese as a Foreign Language]* (p 49–52).

Matsushita, T. （松下達彦）. (2011c). 日本語の学術共通語彙（アカデミック・ワード）の抽出と妥当性の検証 [Extracting and validating the Japanese Academic Word List]. *[2011 年度 日本語教育学会春季大会 予稿集 [Proceedings of the Conference for Teaching Japanese as a Foreign Language, Spring 2011]* (p 244–249).

Matsushita, T., Taft, M., & Tamaoka, K. （松下達彦・Ｍａｒｃｕｓ Ｔａｆｔ・玉岡賀津雄）. (2004). 中国語「単語」を知っていることは日本語漢字語の発音学習に役立つか？ [Is it useful to know Chinese 'words' to learn Kanji pronunciation?]. *平井勝利教授退官記念　中国学・日本語学論文集 [Collected papers on Sinology and Japanese linguistics in memory of retirement of Prof. Katsutoshi Hirai]* (p 578–590). Tokyo: 白帝社 [Hakutei-sha].

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, *88*(5), 375–407. doi:10.1037/0033-295X.88.5.375

MEXT (Ministry of Education,Culture,Sports,Science & Technology in Japan) （文部科学省）. (1989). 学年別漢字配当表 [The graded Kanji lists for Japanese primary schools]. Downloaded from http://www.mext.go.jp/b_menu/shuppan/sonota/990301b/990301d.htm

Mikami, K., & Harada, T. （三上京子・原田照子）. (2011). 多読による付随的語彙学習の可能性を探る :日本語版グレイディド・リーダーを用いた多読の実践と語彙テストの結果から(Exploring the Possibility of Incidental Vocabulary Acquisition through Extensive Reading : From results based on extensive reading and vocabulary tests of Japanese graded readers). *国際交流基金日本語教育紀要*, *7*, 7–23.

Mizumoto, T., & Ikeda, R. （水本光美・池田隆介）. (2003). 導入教育における「基礎専門語」の重要性–環境工学系留学生のための語彙調査と分析から (The importance of basic technical Japanese in introductory education for specific purposes: Based on a survey and an analysis of vocabulary for interenational students majoring in environmental engineering). *専門日本語教育研究 (Journal of technical Japanese education)*, (5), 21–28.

Mizumoto, T., Ikeda, R., Hirayama, Y., Fukuda, H., Sun, L., & Lee, S.-W. （水本光美・池田隆介・平山義則・福田展淳・孫連明・李丞祐）. (2005). カタカナ語を含む専門用語の特徴–環境工学系「純粋専門語」の調査と分析 (Characteristics of Katakana-technical Japanese: Survey and analysis of "Core-technical Japanese" for environmental engineering studies). *専門日本語教育研究 (Journal of technical Japanese education)*, (7), 35–40.

Mogi, T., Yamaguchi, M., Maruyama, T., & Tanaka, M. （茂木俊伸・山口昌也・丸山岳彦・田中牧郎）. (2005). 語種辞書「かたりぐさ」の開発と月刊雑誌の語種構成分析 (Development of a word origin type dictionary Katarigusa and analysis of proportion of word origin type in monthly magazines). *言語処理学会第 11 回年次大会発表論文集 (Proceedings for 11th Annual Conference of the Association for Natural Language Processing)*, 341–344.

Mori, K. （森清）(Original E., & Japan Library Association (revised edition). (1995). *日本十進分類法 (Nippon Decimal Classification)* (9th ed.). Tokyo: 日本図書館協会 (Japan Library Association).

Mori, Y. (1998). Effects of first language and phonological accessibility on Kanji recognition. *The Modern Language Journal*, *82*(1), 69–82.

Morioka, K. （森岡健二）. (1984). 形態素論　－語基の分類－ (Morphology: Classification of word base). *上智大学国文学科紀要 [Bulletin of School of Japanese Language and Literature, Sophia University]*, *1*, 129–181.

Muller, C. (1965). Fréquence, dispersion et usage: à propos des dictionnaires de fréquence. *CdeL*, *7*(2), 33–42, cited in Lyne (1985, p. 125).

Muraoka, T., Kagehiro, Y., & Yanagi, T. （村岡貴子・影廣陽子・柳智博）. (1997). 農学系 8 学術雑誌における日本語論文の語彙調査-農学系日本語論文の読解および執筆のための日本語語彙指導を目指して- (Vocabulary analysis of Japanese papers in eight agricultural science journals: For the teaching of technical vocabulary to foreign students majoring in agricultural science). *日本語教育 (Journal of Japanese Language Teaching)*, *95*, 61–72,176–177.

Muraoka, T., & Yanagi, T. （村岡貴子・柳智博）. (1995). 農学系学術雑誌の語彙調査　－専門分野別日本語教育の観点から－ [A survey of vocabulary in academic journals of agriculture: From the viewpoint of teaching Japanese for technical domains]. *日本語教育 (Journal of Japanese Language Teaching)*, *85*, 80–89.

Nagano, T. （長野 正）. (1995). *日本語の音声表現: スピーチ・コミュニケーション [Spoken Expression in Japanese: Speech Communication]*. Tokyo: Tamagawa University Press （玉川大学出版部）.

Nakano, H., & Nomura, M. （中野洋・野村雅昭）. (1979). 日本語の形態素解析 (An analysis of Japanese morpheme). *情報処理 (Information Processing)*, *20*(10), 857–864.

Nation, I. S. P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. P. Bogaards & B. Laufer (Eds.), *Vocabulary in a Second Language: Selection, Acquisition, and Testing* (p 3–13). Amsterdam: John Benjamins.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82.

Nation, I. S. P., & Deweerdt, J. (2001). A defence of simplification. *Prospect*, *16*(3), 55–67.

Nation, I. S. P., & Heatley, A. (2002). *Range*. LALS, Victoria University of Wellington, New Zealand. Downloaded from http://www.victoria.ac.nz/lals/staff/paul-nation.aspx

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (p 6–19). Cambridge: Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). *Researching and analysing vocabulary*. Boston: Heinle Cengage Learning.

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

Nation, P., & Wang, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language*, *12*(2), 355–380.

New, B., Brysbaert, M., Veronis, J., & Pallier, C. (2007). The use of film subtitles to estimate word frequencies. *Applied Psycholinguistics*, *28*(4), 661–677.

NINJAL (The National Institute for Japanese Language) (国立国語研究所). (2009). 現代日本語書き言葉均衡コーパス 2009 年モニター版 (Balanced Corpus of Contemporary Written Japanese 2009 monitor version). Unpublished (available by application).

Nishimura, Y. （西村由起子）. (2010). 話し言葉、書き言葉、そしてオンライン言語をめぐって　ー日本語全体像を捉える試みへのパイロットリサーチー (On variation across speech, writing and language online: A pilot research toward an 「overall」 approach to Japanese). *特定領域研究「日本語コーパス」平成 21 年度公開ワークショップサテライトセッション予稿集* (p 73–84). Tokyo: 国立国語研究所コーパス開発センター [The Center for Corpus Development, the National Institute for Japanese Langauge].

NLRI (The National Language Research Institute) (国立国語研究所). (1962). *現代雑誌九十種の用字・用語　第一分冊　総記および語彙表 (Vocabulary and Chinese Characters in Ninety Magazines of Today: (Volume I) General Description & Vocabulary Frequency Tables)*. Tokyo: Shuuei Shuppan (秀英出版).

NLRI (The National Language Research Institute) (国立国語研究所). (1963). *現代雑誌九十種の用字・用語　第二分冊　漢字表 (Vocabulary and Chinese Characters in Ninety Magazines of Today: (Volume II) Kanji Frequency Tables)*. Tokyo: Shuuei Shuppan (秀英出版).

NLRI (The National Language Research Institute) (国立国語研究所). (1964). *現代雑誌九十種の用字・用語　第三分冊　分析 (Vocabulary and Chinese Characters in Ninety Magazines of Today: (Volume III) Analysis of the Results)*. Tokyo: Shuuei Shuppan (秀英出版).

NLRI (The National Language Research Institute) (国立国語研究所). (1970). *電子計算機による新聞の語彙調査(I) (Studies on the vocabulary of Modern Newspapers, Volume I)*. Tokyo: Shuuei Shuppan （秀英出版）.

NLRI (The National Language Research Institute) (国立国語研究所). (1976). *現代新聞の漢字 (A Study of Uses of Chinese Characters in Modern Newspapers)*. Tokyo: Shuuei Shuppan （秀英出版）.

NLRI (The National Language Research Institute) (国立国語研究所). (1984). *日本語教育のための基本語彙調査 (A Study of Fundamental Vocabulary for Japanese Language Teaching)*. Research Report. Tokyo: Shuuei Shuppan （秀英出版）.

NLRI (The National Language Research Institute) (国立国語研究所). (2006). 現代雑誌 200 万字言語調査語彙表 (The vocabulary lists from the language survey of contemporary magazines with two million running characters). Downloaded from http://www.kokken.go.jp/katsudo/seika/goityosa/index.html

Noguchi, H. (野口裕之). (2008). 試験結果の分析 (Analyses of the test results). 国際交流基金・日本国際教育支援協会 (The Japan Foundation and Japan Educational Exchanges and Services) (Ed.), *平成 17 年度日本語能力試験　分析評価に関する報告書 (Report on the analysis and evaluation of the Japanese-Language Proficiency Test 2005)* (p 45–111). Tokyo: Bonjinsha (凡人社).

Nozaki, H., Yokoyama, S., Isomoto, Y., & Yoneda, J. （野崎浩成・横山詔一・磯本往雄・米田純子）. (1996). 文字使用に関する計量的研究: 日本語教育支援の観点から (A quantitative research on character usages: from the viewpoint of support for teaching Japanese). *日本教育工学雑誌 (Japan Journal of Educational Technology)*, *20*(3), 141–149.

Ogiso, T. (小木曽智信). (2009). *茶まめ "Chamame" Version 1.71 (graphical user interface for MeCab)*. Downloaded from https://www.tokuteicorpus.jp/dist/modules/system/modules/menu/main.php?page_id=1&op=change_page

Ogura, H., Koiso, H., Fujiike, Y., & Hara, Y. (小椋秀樹・小磯花絵・富士池優美・原裕). (2009). *「現代日本語書き言葉均衡コーパス」形態論情報規定集　改定版 [The Rule Book of Morphological Information for the Balanced Corpus of Contemporary Written Japanese]*. 国立国語研究所内部報告書. NINJAL (National Institute for Japanese Language).

Oka, M. （岡 益巳）. (1992). 非漢字圏の留学生のための日本経済基本用語表 [Basic terms of the Japanese economy for non-Kanji background students]. *岡山大学経済学会雑誌 (Okayama Economic Review)*, *23*(4), 191–229.

Oxford, R., & Crookall, D. (1990). Vocabulary learning: a critical analysis of techniques. *TESL Canada Journal*, *7*(2), 9–30.

Paivio, A., & Desrochers, A. (1980). A dual-coding approach to bilingual memory. *Canadian Journal of Psychology/Revue canadienne de psychologie*, *34*(4), 388–399. doi:10.1037/h0081101

Paribakht, S. (2005). The influence of first language lexicalization on second language lexical inferencing: A study of Farsi-speaking learners of English as a foreign language. *Language Learning*, *55*(4), 701–748.

Prince, P. (1996). Second language vocabulary learning: the role of context versus translations as a function of proficiency. *Modern Language Journal*, *80*(4), 478–493.

Quackenbush, H., & Oso, M. （カッケンブッシュ寛子・大曾美恵子）. (1990). *外来語の形成とその教育 [Formation and teaching of Loanwords from Western languages in Japanese]*. Tokyo: Printing Bureau, Ministry of Finance （大蔵省印刷局）.

Read, J. (1988). Measuring the Vocabulary Knowledge of Second Langauge Learners. *RELC Journal*, *19*(2), 12–25. doi:10.1177/003368828801900202

Reicher, G. M. (1969). Perceptual recognition as a function of meaningfulness of stimulus material. *Journal of Experimental Psychology*, *81*(2), 275–280. doi:10.1037/h0027768

Richards, B. J., & Malvern, D. D. (1997). *Quantifying lexical diversity in the study of language development*. The New Bulmershe Papers. Reading: University of Reading.

Richards, J. C. (1974). Word lists: problems and prospects. *RELC Journal*, *5*(2), 69–84.

Richards, J. C. (2001). *Curriculum Development in Language Teaching*. Cambridge, UK: Cambridge University Press.

Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning*. Second language acquisition (Clevedon, England) (Vol. 21). Multilingual Matters.

Rosengren, I. (1971). The quantitative concept of language and its relation to the structure of frequency dictionaries. *Études de linguistique appliquée (Nouvelle Série)*, *1*, 103–27.

Rott, S. (1999). The Effect of Exposure Frequency on Intermediate Language Learners' Incidental Vocabulary Acquisition and Retention Through Reading. *Studies in Second Language Acquisition*, *21*(04), 589–619. doi:null

Saito, T. （斎藤匡史）. (1988). 日中同形語をめぐって [On Chinese cognates in Japanese]. *研究論叢　東亜大学学術研究所 [Research Forum, Research Center, University of East Asia]*, *13*(1), 147–171.

Sanaoui, R. (1995). Adult learners' approaches to learning vocabulary in second languages. *Modern Language Journal*, *79*(1), 15–28.

Sato, M. （佐藤政光）. (1999). 日本語学習者の語彙習得に関する調査研究　－(1)基本語彙の問題点 (On the acquisition of Japanese vocabulary: (1) Several problems with fundamental vocabulary). *明治大学人文科学研究所紀要 (Memoirs of the Institute of Cultural Sciences, Meiji University)*, *44*, 169–180.

Schmitt, N. (2000). *Vocabulary in language teaching*. Cambridge, UK: Cambridge University Press.

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. doi:10.1111/j.1540-4781.2011.01146.x

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88.

Shinya, T., & Matsushita, T. （新屋映子・松下達彦） (Eds.). (1994). *日本語上級読解演習 国際学アラカルト [International Studies, A la Carte: Reading Seminar Texts for Advanced Learners of Japanese]*. Tokyo: Obirin University [桜美林大学], internally published textbook supported by Unique Educational Research Grant, The Promotion and Mutual Aid Corporation for Private Schools of Japan [日本私学振興財団「特色ある教育研究」].

Sone, H. （曾根博隆）. (1988). 日中同形語に関する基礎的考察 [A basic study on Chinese cognates in Japanese]. *明治学院論叢 (The Meiji-Gakuin review), 424*, 61–96.

Sumi, T. （角 知行）. (2010). 学術基本用語集作成の試み (An attempt for making a basic academic word list). *アカデミック・ジャパニーズ・ジャーナル (Academic Japanese Journal), 2*, 11–21.

Sutarsyah, C., Nation, P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based study. *RELC Journal*, *25*(2), 34–50.

Suzuki S. （鈴木修次）. (1981). *日本漢語と中国: 漢字文化圏の近代化 [Sino-Japanese Words and China: Modernization of Kanji Culture Area]*. Tokyo: Chuo Koronsha （中央公論社）.

Swan, M. (1997). The influence of the mother tongue on second language vocabulary acquisition and use. N. Schmitt & M. McCarthy (Eds.), *Vocabulary: Description, Acquisition and Pedagogy* (p 156–180). Cambridge: Cambridge University Press.

Tajino, A., Dalsky, D., & Sasao, Y. (2009). Academic vocabulary reconsidered: An EAP curriculum-design perspective. *Journal of Teaching English as a Foreign Language and Literature*, *1*(4), 3–21.

Tajino, A., Terauchi, H., Sasao, Y., & Maswana, S. (田地野 彰・寺内 一・笹尾洋介・マスワナ紗矢子). (2007). 総合研究大学における英語学術語彙リスト開発の意義　－EAP カリキュラム開発の観点から－ (The development of academic words lists at a multi-disciplinary university in Japan: A fundamental step in EAP curriculum design). *京都大学高等教育研究 (Kyoto University Researches in Higher Education), 13*.

Takano, S., & Wang, B. （高野繁男・王宝平）. (2002). 日中現代漢語の層別　―日中同形語に見る― [Tiers of modern Chinese-origin vocabulary in Japanese and Chinese: A study on Chinese cognates]. Institute for Humanities Research, Kanagawa University（神奈川大学人文学研究所） (Ed.), *日中文化論集 [Papers on Japanese and Chinese Cultures]* (p 118–139). Tokyo: 勁草書房 [Keiso Shobou].

Tamamura, F. (玉村文郎). (1984). *語彙の研究と教育（上） [Studies and Education on Vocabulary Vol.1]*. (NLRI (The National Language Research Institute) (国立国語研究所), Ed.). Tokyo: Printing Bureau, Ministry of Finance （大蔵省印刷局）.

Tamamura, F. （玉村文郎）. (1987). 日本語教育基本 2570 語 [Basic 2570 words for teaching Japanese as a second language]. *日本語の語彙・意味 (2) [Japanese Vocabulary and Meaning]*, NAFL Institute 日本語教師養成通信講座 [Training Course of Teachers of Japanese as a Second Language]. アルク (Alc).

Tamaoka, K. (2004). The 4th edition database for the 1,945 basic Japanese kanji. Downloaded from http://www.lang.nagoya-u.ac.jp/~ktamaoka/down_en.htm

Tamaoka, K., & Matsushita, T. （玉岡賀津雄・松下達彦）. (1999). 中国語系日本語学習者による日本語漢字二字熟語の認知処理における母語の影響 [First language effects on the cognitive processing of Japanese two-kanji compounds by Chinese learners of Japanese]. 第 4 回国際日本語教育・日本研究シンポジウム「アジア太平洋地域における日本

語教育・日本語研究」 [The 4th International Symposium on Japanese Language Teaching and Japanese Studies].

Tamaoka, K., Miyaoka, Y., & Matsusita, T. （玉岡賀津雄・宮岡弥生・松下達彦）. (2004). Inter-language activations and inhibitions in cognitive word processing by bilinguals in the Chinese and Japanese languages. *In the Proceedings of 6th International Conference of the Japanese Society for Language Sciences (JSLS 2004)* (p 43–48).

Terajima, H. （寺嶋弘道）. (2010). BCCWJ により検証した日本語教育語彙の検証－経営学を専攻する日本語学習者を対象にしてー (Surveying and selecting specialised vocabulary for Japaense learners majoring business administration). *特定領域研究「日本語コーパス」平成 21 年度公開ワークショップサテライトセッション予稿集* (p 107–115). Tokyo: 国立国語研究所コーパス開発センター [The Center for Corpus Development, the National Institute for Japanese Langauge].

The Japanese-language Institute, Japan Foundation （国際交流基金日本語国際センター）. (1995). *日本語かな入門　英語版 [Intoroduction to Japanese Kana, English version]*. Tokyo: 凡人社 (Bonjinsha).

Thorndike, E. L., & Lorge, I. (1944). *The Teacher's Word Book of 30,000 Words*. New York: Teachers College Columbia University.

Townsend, D., & Collins, P. (2008). Academic vocabulary and middle school English learners: an intervention study. *Reading and Writing*, *22*(9), 993–1019. doi:10.1007/s11145-008-9141-y

Toyoda, E. (2007). Enhancing autonomous L2 vocabulary learning focusing on the development of word-level processing skills. *The Reading Matrix*, *7*(3), 13–34.

Toyoda, E., & McNamara, T. (2011). Character recognition among English‐speaking L2 readers of Japanese. *International Journal of Applied Linguistics*, *21*(3), 383–406. doi:10.1111/j.1473-4192.2011.00285.x

Utiyama, M., & Isahara, H. (2003). Reliable Measures for Aligning Japanese-English News Articles and Sentences. *ACL-2003* (p 72–79).

Utiyama, M., & Takahashi, M. (2003). *English-Japanese Translation Alignment Data*. Downloaded from http://www2.nict.go.jp/x/x161/members/mutiyama/align/index.html

Vander Beke, G. E. (1932). *French Word Book*. New York: Publications of American and Canadian Committees on Modern Languages, Vol. XV. Cited in Lyne (1985).

Vermeer, A. (2004). The relation between lexical richness and vocabulary size in Dutch L1 and L2 children. P. Bogaards & B. Laufer (Eds.), *Vocabulary in a Second Language: Selection, Acquisition, and Testing* (p 173–189). Amsterdam: John Benjamins.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a medical academic word list. *English for Specific Purposes*, *27*(4), 442–458.

Ward, J. (1999). How large a vocabulary do EAP Engineering students need? *Reading in a Foreign Language*, *12*(2), 309–323.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language*, *15*(2), 130–163.

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65.

Webb, S. (2008). The effects of context on incidental vocabulary learning. *Reading in a Foreign Language*, *20*(2), 232–245.

West, M. (1953). *A General Service List of English Words*. London: Longmans, Green & Co.

Xiao, R., Rayson, P., & McEnery, T. (2009). *A Frequency Dictionary for Mandarin Chinese: Core Vocabulary for Learners*. New York: Routledge.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–229.

Yamada, A. （山田 篤）. (2008). *NumTrans Version 0.5*.

Yamada, A., & Koiso, H. （山田 篤・小磯花絵）. (2008). NumTrans マニュアル. Mounted on the user interface software 茶まめ "Chamame" (Ogiso, 2009).

Yamazaki, M., & Onuma, E. （山崎 誠・小沼 悦）. (2004). 現代雑誌における語種構成 (The proportion of word origin types in contemporary magazines). 言語処理学会第 10 回年次大会発表論文集 (Proceedings for 10th Annual Conference of the Association for Natural Language Processing).

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.

Yokoyama, S., Sasahara, H., Nozaki, H., & Long, E. （横山詔一・笹原宏之・野崎浩成・エリク＝ロング）. (1998). *新聞電子メディアの漢字――朝日新聞 CD-ROM による漢字頻度表―― [Kanji in digitized newspapers: A Kanji frequency list made from the Asahi CD-ROM]*. 三省堂 (Sanseido).

Young, D. J. (1999). Linguistic simplification of SL reading material: effective instructional practice? *Modern Language Journal*, *83*(3), 350–366.

Zipf, G., K. (1949). *Human Behavior and the Principle of Least Effort: An Introduction to Human Ecology*. New York: Hafner.