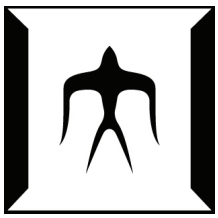
論文 / 著書情報
Article / Book Information

題目(和文)	
Title(English)	CONTEXTUALLY AWARE WRITING ASSISTANCE SYSTEM FOR JAPANESE
著者(和文)	HODOSCEKBor
Author(English)	Bor HODOSCEK
出典(和文)	学位:博士(工学), 学位授与機関:東京工業大学, 報告番号:甲第9334号, 授与年月日:2013年9月25日, 学位の種別:課程博士, 審査員:室田 真男,中山 実,野原 佳代子,山元 啓史,赤間 啓之
Citation(English)	Degree:Doctor (Engineering), Conferring organization: Tokyo Institute of Technology, Report number:甲第9334号, Conferred date:2013/9/25, Degree Type:Course doctor, Examiner:,,,,
学位種別(和文)	博士論文
Type(English)	Doctoral Thesis

CONTEXTUALLY AWARE WRITING ASSISTANCE SYSTEM FOR JAPANESE

Bor Hodošček



Superviser: Prof. Masao Murota
Examiners: Prof. Minoru Nakayama
Prof. Kayoko Nohara
Assoc. Prof. Hilofumi Yamamoto
Assoc. Prof. Hiroyuki Akama

Submitted to the

Department of Human System Science,
Graduate School of Decision Science and Technology,
Tokyo Institute of Technology

in partial fulfillment of the requirements for the degree of

Doctor of Engineering

August 8, 2013

PUBLICATIONS

Some of the ideas and figures presented in this dissertation have previously appeared in the following publications:

Research Papers

- Joyce, T., Hodošček, B., & Nishina, K. (2012). Orthographic representation and variation within the Japanese writing system: Some corpus-based observations. *Written Language & Literacy*, *15*(2) Special Issue on Units of Language – Units of Writing, 254–278. doi:10.1075/wll.15.2.01rob
  - §1.1, §3.5
- Hodošček, B. (2011). Word class ratios and genres in written Japanese: Revisiting the Modifier Verb Ratio. *Acta Linguistica Asiatica*, *1*(2), 53–62. Retrieved from http://revije.ff.uni-lj.si/ala/article/view/28/37
  - §4.1
- Hodošček, B., Abekawa, T., Bekeš, A., & Nishina, K. (2011). Assisting co-occurrence production in report writing: Evaluation of writing assistance tool Natsume. *Journal of Technical Japanese Education*, *13*, 33–40. doi:10.11448/jtje.13.33
  - §7.5.2

Research Articles (Project Reports)

- Hodošček, B. & Nishina, K. (2012b). Japanese learning support systems: Hinoki project report. *Acta Linguistica Asiatica*, *2*(3) Lexicography of Japanese as a Second/Foreign Language (Part 2), 95–124. Retrieved from http://revije.ff.uni-lj.si/ala/article/view/221
  - §3.1, §3.2, §3.3, §3.4, §7, §8

Conference Proceedings

- Hodošček, B., Abekawa, T., Murota, M., & Nishina, K. (2012, August). Readability of example sentences in writing assistance tool Natsume. (pp. 1–4). 5th international conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J). Nagoya. Retrieved from http://2012castelj.kshinagawa.com/proceedings/Poster/Poster8_Bor%20Hodo%C5%A1%C4%8Dekodoscek.pdf
  - §6

- Hodošček, B. & Nishina, K. (2012a, March). BCCWJ ni okeru shuttenjōhō to topikku oyobi rejisutā to no kankei [Comparison of metadata with topic and register in the BCCWJ]. In *Dai ikkai kōpasu nihongo waākushoppu yokōshū [Proceedings of the First Workshop on Japanese Corpus Linguistics]* (pp. 339–342). Dai ikkai nihongo kōpasu wākushoppu [First Workshop on Japanese Corpus Linguistics]. Tokyo, Japan
  - §5
- Hodošček, B. & Nishina, K. (2011b, August). On the treatment of register in writing assistance systems. (Vol. 2, pp. 522–523). International Conference on Japanese Language Education 2011. Tianjin, China
  - §8.2
- Hodošček, B. & Nishina, K. (2011a, March). Learning effect on academic Japanese expression usage with writing support system Natsume. (pp. 1–2). 13th forum of the society for technical japanese education. Tohoku University, Sendai, Japan. Retrieved from http://stje.kir.jp/download/13STJE_discussion.pdf
  - §7.5.1
- Abekawa, T., Hodošček, B., & Nishina, K. (2012, August). Nihongo sakubun shien shisutemu Natsume ni okeru aratana youhou no kumikomi. International Conference on Japanese Language Education. Nagoya, Japan
  - §7
- Yagi, Y., Hodošček, B., & Nishina, K. (2012, March). BCCWJ to gakushūsha sakubun kōpasu o riyōshita nihongo sakubun shien [Japanese writing assistance using the BCCWJ and a learner corpus]. In *Dai ikkai kōpasu nihongogaku wākushoppu yokōshū [Proceedings of the First Workshop on Japanese Corpus Linguistics]*. Dai ikkai nihongo kōpasu wākushoppu [First Workshop on Japanese Corpus Linguistics]. Tokyo, Japan
  - §8.2.3
- Abekawa, T., Hodošček, B., & Nishina, K. (2011b, August). Japanese writing support system Natsume using genre information. (pp. 774–775). International Conference on Japanese Language Education. Tianjin, China
  - §7
- Abekawa, T., Hodošček, B., & Nishina, K. (2011a, March). Go no kyōki o kōritsuteki ni kensaku dekiru nihongo sakubun shien shisutemu Natsume no shōkai [Introduction to efficient collocation search in Japanese writing assistance system Natsume]. (Vol. 17, pp. 595–598). Proceedings of The

17th Annual Meeting of The Association for Natural Language Processing. Toyama: The Association for Natural Language Processing

  - §7

- Abekawa, T., Hodošček, B., & Nishina, K. (2010). Collocation search refinements in the natsume writing support system. In *Tokutei ryōiki 'nihongo kōpasu' hēsē 21 nendo kōkai wākushoppu (kenkyū sēka hōkokukai) yokōshū* (pp. 243–244)

  - §7

Book Chapters

- Hodošček, B. (2013). Kōpasu no shūshū, setsumei, janru; jikken to bunseki [Corpus collection, explanation and genre; Experiment and analysis]. In Y. Sunakawa (Ed.), *Kōza nihongo kōpasu [Japanese corpus textbook series]* (Chap. 5, Vol. 5, Vols. 8). Asakura Publishing Co., Ltd.

  - §7

- Hodošček, B. (2012). Sakubun sien to rejisutā [Writing assistance and register]. In K. Nishina, M. Kamada, H. Cao, T. Utashiro, & T. Muraoka (Eds.), *Nihongo gakusyūsien no kōtiku: gengokyōiku kōpasu sisutemu kaihatu [Constructing Japanese Language Learning: Language education, corpus and system development]* (3, pp. 275–287). Tokyo, Japan: Bonjinsha

  - §7

Master's Thesis

- Hodošček, B. (2010). *Development of a register-based writing assistance system for academic Japanese* (Master's thesis, Tokyo Institute of Technology)

  - §2.1

We have seen that computer programming is an art,
because it applies accumulated knowledge to the world,
because it requires skill and ingenuity, and especially
because it produces objects of beauty.

Knuth, D. E. (2007). Computer programming as an art. In *ACM Turing award lectures* (p. 1974). ACM

ACKNOWLEDGEMENTS

This thesis could not have come about without the help of many people. Foremost, I wish to thank my supervisors, Professor Emeritus Kikuko Nishina and Professor Masao Murota, for their support and guidance. The most significant contributions that made this research possible were the work of the informally named "Asunaro Group" led by Kikuko Nishina. I thank Takeshi Abekawa, Yutaka Yagi, Taizan Suzuki, and Liang Fu for their work on the Asunaro, Natsume, and Nutmeg systems. This work would not be possible without them.

Next, I would like to thank Terry Joyce for considerable help with proofreading this thesis.

I am also indebted to Hilofumi Yamamoto, whose invaluable comments and support led to several ideas present in this thesis, and to Minoru Nakayama, Kayoko Nohara, and Hiroyuki Akama for their patience and feedback during the writing of this thesis.

Frequent research meetings with Andrej Bekeš, Irena Srdanović, and Kikuko Nishina also provided a place for discussions that influenced some of the ideas presented within the thesis.

I would also like to thank the members of Murota Lab, as well as the Nishina Lab members before them, for their warm feedback and support during the writing of this thesis.

I would like to thank all the research participants who gave so willingly of their time to the system evaluations.

And finally, I gratefully appreciate the financial support provided by MEXT that made it possible to focus on my studies and complete this thesis.

Bor Hodošček

August 8th, 2013

Tokyo

CONTENTS

LIST OF FIGURES

LIST OF TABLES

ACRONYMS

AdjN	Adjective-Noun
API	Application Programming Interface
BCCWJ	Balanced Corpus of Contemporary Written Japanese
CALL	Computer-Assisted Language Learning
CI	Confidence Interval
JASSO	The Japan Student Services Organization
JLPT	Japanese Language Proficiency Test
KWIC	KeyWord In Context
L2	Second Language
LDA	Latent Dirichlet Allocation
LUW	Long Unit Word
MD	Multi-Dimensional
MEXT	Japanese Ministry of Education, Culture, Sports, Science and Technology
MVR	Modifier Verb Ratio
NAIST	Nara Advanced Institute of Science and Technology
NDC	Nippon Decimal Classification
NINJAL	National Institute for Japanese Language and Linguistics
NLP	Natural Language Processing
NPAdj	Noun-Particle-Adjective
NPV	Noun-Particle-Verb
ROC	Receiver Operating Characteristic
SFL	Systemic Functional Linguistics
STJC	Scientific and Technical Japanese Corpus

SVM Support Vector Machine

SUW Short Unit Word

XML eXtensible Markup Language

Part I

BACKGROUND

1 INTRODUCTION

Writing is not only an important form of communication but is one of the core skills of higher education. The challenges of academic writing in Japanese as a second language (L2) are compounded by the complexities of the Japanese writing system and the difficulties of acquiring the specialized knowledge of academic writing style necessary to be an effective communicator in ones field of specialization. Individual learners can rely on domain-specific dictionaries, style guides, or extensive study of documents from their field, but these activities are often separated from the act of writing on a computer. Although the use of spellcheckers and online grammar correction services is becoming increasingly ubiquitous, the target of these services is often the general public, so they seldom support the specialized needs of L2 writers in an academic setting. For example, learners using a search engine are able to look for examples on how to use words and expressions, but the general purpose nature of many such tools limits the amount of assistance they can provide, especially in highlighting differences in writing style. More recently, with access to large quantities of language data available at low monetary and computational cost, a number of methods have been proposed to utilize these large quantities of language data[1] for the purposes of writing assistance. In this thesis, I propose the use of recently available Japanese language corpora to realize writing assistance systems for L2 writers attending Japanese institutions of higher education, where they are required to write reports and papers in Japanese. I will show that:

1. Linguistic resources such as corpora are effective for enhancing collocation usage in writing.
2. It is possible to develop a model of writing style (register) that can discriminate between correct writing styles using corpora containing diverse registers.

1.1 THE JAPANESE WRITING SYSTEM

Japanese is written in a mixture of script styles that reflect the language's history of borrowing and innovation (Kess & Miyamoto, 1999; Gottlieb, 2008; Tranter, 2008; Joyce, 2011). Modern Japanese writing is comprised of four principal component scripts: morphographic 漢字 /kanji/ "Chinese characters", the two "kana" syllabographic sets of ひらがな /hiragana/ and カタカナ /katakana/,

1 Referring to the gigabyte level and beyond.

and supplementary alphabetic ローマ字 /rōmaji/ "Roman alphabet", Arabic numerals, and various other symbols. This section will introduce these scripts and their interactions within the writing system as a whole.

1.1.1 *Morphographic Kanji*

The first component of the Modern Japanese writing system, kanji, was in fact borrowed from China as early as the 1st century CE (Shibatani, 1990). It was adopted with the Chinese language of the period and used as the written language, but retained Chinese syntax, orthography, and pronunciation. The gradual change towards using kanji to write Japanese language led to the creation of the dual system of 音読み /on'yomi/ "Sino-Japanese pronunciations" and 訓読み /kun'yomi/ "Native-Japanese pronunciations" associated with kanji today.

One factor that contributed to the change was the invention of special conventions for reading order and the pronunciations of Chinese texts, known as 訓読 /kundoku/. This greatly facilitated the influx of Chinese words into the Japanese language (Miller, 1967), where imitations of the original Chinese pronunciation were pronounced in the *kundoku* style.

For example, the kanji 龍 representing a dragon was pronounced as /ryū/ in the Sino-Japanese approximation. For many other words that already existed in the Japanese language, the Japanese *kun'yomi* native pronunciation was preferred. For example, the Chinese character 藤 is pronounced in Native-Japanese as /fuji/ and refers to the wisteria flower. And in some other cases, the two readings coexist, providing us with two or more different ways to read one character. In fact, the aforementioned dragon character 龍 was simplified to the character 竜 and which is more commonly pronounced in the Native-Japanese /tatsu/, although both kanji and pronunciations are still in common use, though they are used in different circumstances.

The number of kanji used has historically been high, but through several government reforms beginning in the middle of the 20th century, the proscribed kanji list has ranged from the 1,850 kanji introduced in the first reform (the 当用漢字表 /tōyō kanjihyō/ list in 1946), to the most recent standard list (the 常用漢字表 /jōyō kanjihyō/ "List of characters for general use", last modified in 2010) that contains 2,136 kanji. A subset of the jōyō kanji list consisting of 1,006 教育漢字 /kyōiku kanji/ "education kanji" is also the proscribed standard list of characters taught during the six years of elementary school.

In Modern Japanese, kanji are primarily used to write content words from many word classes, including nouns, verbs, adjectives, and adverbs. As the number of kanji is large, and the *jōyō* kanji list is typically learned by the end of high school, kanji can in some ways be associated with reading difficulties. Such

See §6 for more information

difficulties are even more pronounced for L2 learners of Japanese[2] who are still in the process of learning kanji.

1.1.2 *Syllabographic Hiragana and Katakana*

The two native syllabaries of hiragana and katakana were developed in the 9th century, although both were independently developed. Their basis as a phonetic transcription is found in the 万葉集 /man'yōshū/ (759 CE), an anthology of Japanese verse[3]. While it is possible to write all sounds of the Japanese language in either hiragana or katakana, their use did not displace kanji[4] completely, leading to the present mixed system of writing in Modern Japanese.

Hiragana and katakana characters represent syllables or morae of equal duration, and are both comprised of 46 basic characters. For example, the syllable /na/ is represented as な in hiragana and ナ in katakana. Additionally, using the 濁音 /dakuon/ "voiced" (゛) and 半濁音 /handakuon/ "semi-voiced" (゜) diacritics, the set of characters expands to 71 for both scripts. For example, combining the syllable か /ka/ with the voiced diacritic ゛ results in the character が /ga/.

Though both syllabaries can represent all Japanese writing, their use is markedly different. Hiragana are mostly used for grammatical elements and inflections. They are also used in place of kanji in texts that are meant to be read by readers who have not yet learned all *jōyō* kanji, typically children and pupils in compulsory education. Katakana, on the other hand, are predominantly used to write foreign loan words, scientific names, and onomatopoeia[5].

1.1.3 *Alphabetic Rōmaji and Arabic Numerals*

The most recent borrowing of a foreign script is the rōmaji script, which was first introduced to Japan around 1548 through Japanese Catholics translating Catholic books using Portuguese orthography. The prevailing transliteration conventions for Modern Japanese are the ヘボン式 /hebonshiki/ "Hepburn system" (used here), proposed by the American missionary James Curtis Hepburn (1815–1911), and the 訓令式 /kunreishiki/ "Cabinet Ordinance System", which is the official transliteration method ordained by the Japanese government in 1954. Rōmaji are commonly used for supplementary glosses in public transportation, and for advertising and mass media.

---

2 The situation is somewhat different for Chinese learners of Japanese, as other issues such as semantic deviance between Japanese and Chinese characters exist.

3 Indeed, characters employed in the *man'yōshū* are referred to as 万葉仮名 /man'yōgana/ (Miller, 1967; Shibatani, 1990).

4 Women traditionally forbidden from learning kanji were the first (and last) wholesale adopters of hiragana during the Heian period (794–1185).

5 Tranter (2008) includes discussion of more novel uses.

Table 1: Usage domains for each component of the Japanese writing system. In the case of the mixed examples, bold characters are used for kanji to distinguish them from hiragana symbols.

	Example	**Common usage**
Kanji	漢字, 混じり文	Nouns, verb stems
Hiragana	ひらがな	Inflectional and grammatical elements
Katakana	カタカナ	Foreign names, loan words, scientific species names, onomatopoeia
Rōmaji	MOTTAINAI, MEXT	Signs, advertisements
Arabic num.	42, 2013	Numbers
Symbols	。, 「, 」, (^^)	Punctuation, emoji (emoticons)

Arabic numerals are frequently used with horizontally-arranged text, especially on the Internet and in academic correspondence, while numbers represented with kanji are still very common for vertical text, such as that found in literary works.

1.1.4 *A Multi-Script Writing System*

The mixed-script style of writing Modern Japanese is known as 漢字かな混じり文 /kanji-kana-majiribun/, which can be translated as "mixed kanji and kana writing". Though the term implies mixing, the four components of this system — kanji, hiragana, katakana, and rōmaji — are used in complementary ways. → *Table 1*
These separate fields of use can thus be expected to influence the writing process in a great way.

1.2 COMPUTER-AIDED WRITING ASSISTANCE

This section briefly notes various methods for computer-aided writing assistance with a particular focus on those targeting the Japanese language. The main types of writing assistance can be categorized according to the types of errors they seek to prevent: spelling, grammatical, stylistic, or semantic errors (Naber, 2003).

Online dictionaries and, especially, the Japanese language input environment (IME) exert considerable influences on the styles of many writers, but they provide little help in automatically checking for problems in meaning, grammar, or style. Grammar checkers have predominantly been limited in scope; a good example is Chantokun, a recent system that only looks for misuses of Japanese particles. Style checkers mostly help to identify common style incongruities, such → *§8.3*
as the distinction between the "dearu-cho" and "desu/masu-cho" forms of writing.

Table 2: Error types and common applications providing assistance.

Error type	Tools and systems employing these techniques
Spelling	Word Processor, Web Browser, IME (Japanese)
Grammar	Word Processor (limited support in MS Word)
Style	"After the Deadline" (Mudge, 2010; "After the Deadline - Spell, Style, and Grammar Checker for WordPress, Firefox, TinyMCE, jQuery, and CKEditor," 2013), Leacock et al. (2009)

Overall, there is an increasing trend to take advantage of large-scale language data for these tasks, as is the case with the "After the Deadline - Spell, Style, and Grammar Checker for WordPress, Firefox, TinyMCE, jQuery, and CKEditor" (2013) system as well as the Microsoft Research ESL Reading Assistant[6] (Mudge, 2010; Leacock, Gamon, & Brockett, 2009).

1.3 JAPANESE SECOND LANGUAGE EDUCATION

It is important to distinguish between different forms of Japanese L2 education, according to learner location, availability of language instructors, and access to native Japanese speakers.

According to a 2009 report from the Japan Foundation, there are over three and a half million people learning Japanese outside Japan ("Japan Foundation: Survey Report on Japanese Language Education Abroad," 2009). Fortunately, access to quality educational materials has become easier with the advent of the Internet. However, the situation for learners with specialized language needs, such as those who are pursuing a degree at a Japanese institution of higher education, has, regrettably, not improved as much.

The Japan Student Services Organization (2013) reports that there are 138,075 international students in Japan; another report by the Japanese Ministry of Education, Culture, Sports, Science and Technology (MEXT) Agency For Cultural Affairs (2011) reports that there are 40,799 international students studying Japanese in Japan. While recognizing that not all international students, whose primary research activity is conducted in English, need to master writing in Japanese, still the number of students who study Japanese while in Japan is not insignificant. However, all students pursuing specialized study at institutions of higher education within Japan can undoubtedly benefit from mastering some of the following skills:

- read Japanese textbooks
- write reports and papers

6 Although it operation has discontinued.

- listen to lectures and take notes
- present at conferences and seminars

Because it is often not practical to tailor the Japanese language class to meet the specialized language needs of various learners from different fields of specialization, an alternative is needed (Oshima, 2009; Muraoka, Chinami, & Nishina, 2009). One way of approaching the problem is to provide Computer-Assisted Language Learning (CALL) systems for use online. CALL systems can supplement the language learning provided to learners and assist them in studying authentic materials from their research domains. For example, the reading assistance system Asunaro includes coursework for students in Japanese scientific and technical Japanese in addition to reading assistance features (Nishina et al., 2004).

1.4 COLLOCATIONS

The term collocations, as used in this thesis, is defined as frequently co-occurring patterns of words. Knowledge of collocations is regarded as being essential to achieving high levels of L2 proficiency (Pawley & Syder, 1983). Specifically, collocations are important because they offer more contextual information about a word than can be found in conventional dictionaries. Indeed, Nation (2001) likens collocations to the "word properties" that Römer (2006) argues interact with orthography, grammatical behavior, meaning, association, frequency, and style.

Collocations are usually extracted from corpora using computer programs and the strength of the association is measured by some statistical measure such as a χ^2- or t-test. More recently, with the availability of morphological and dependency grammar analyzers, it is possible to compute more robust and comprehensive collocation rankings with less effort, if the collocations are further constrained to combinations of certain word classes or grammatical relations. These techniques offer a complement to the filtering for "less-interesting" functional and common words that are commonly employed with naive n-gram extraction approaches. In this thesis, collocations are extracted based on the grammatical relations between them.

Appendix §10.4 contains definitions for several collocation measures

See §3.5 for more information

The application of collocations to language learning and teaching has a long history, beginning with Sinclair (1987)'s COBUILD project, which led to the development of the COBUILD English dictionary. Johns (1991)'s proposal[7] of data-driven learning (DDL) situates the role of the learner as a discoverer of language and the teacher as a guide who helps the learner to develop more effective strategies for learning the rules of the language.

7 See Johns (2002) for a later conceptualization.

1.5 PROBLEM STATEMENT

L2 Japanese language learners enrolled at Japanese institutions of higher education, where they are required to communicate and write in Japanese, face formidable challenges in not only learning the day-to-day language skills that they need to successfully communicate with other students and teachers, but also in learning to write in the specific register of Japanese academic discourse. Although some textbooks that specifically cover the Japanese academic register for L2 learners currently exist[8], they are inherently limited in what they can cover and, consequently, can only feature small subsets of the specialized vocabulary or expressions of a particular domain. In contrast, the availability of large-scale language resources grants access to a wide variety of topic and registers. The effective use of such resources holds the promise of supplanting static learning and reference resources with dynamic systems capable of covering a wider range of writing and information needs.

1.6 RESEARCH QUESTIONS

The aim of this research is to assist L2 Japanese language learners in writing better reports and research articles. To that aim, I make use of corpora to develop context-aware writing assistance systems that enable learners to use the right word in the right context when writing Japanese compositions.

In evaluating the validity of this approach, it is essential to ask two important research questions.

- *Research Question 1*

  Is it possible to utilize linguistic resources, such as corpora, to realize more effective systems for computer-assisted writing?

- *Research Question 2*

  Is it possible to provide writing style guidance based on corpora alone?

1.7 APPROACH

The approach of this thesis can be summarized in the following two points:

1. Undertake a linguistic theory-informed and quantitative exploration of language data.

8 One relevant example is Nishina, Doi, and Takano (2007).

2. Conduct evaluation experiments with student participants to assess the efficacy of the developed system, to detect deficiencies, and make appropriate modifications to the system design.

In order to address Research Question 1, this research first presents the necessary foundation for modeling context within Japanese writing using register- and topic-diverse corpora. First, context is modeled in terms of register, topic, and readability. With a register model corresponding to the functional or stylistic aspects of writing, a topic model corresponds to the content part of writing, and a readability model can function to reflect the degree of field of experience overlap between a writer and a reader. With the first research question, my interest is primarily in demonstrating how collocation search can be used to improve writing collocations.

Field of experience is defined in the next chapter

Research Question 2 can be considered as being a more specific case of Research Question 1, where the interest is more in showing that collocations can be corrected not only in terms of their grammaticality, but also in terms of their register appropriateness. This question is mainly addressed through the formalization of the register identification feature within Nutmeg, as introduced later in §8.2.1. This approach employs frequencies of collocations across different corpora to make inferences concerning their appropriateness for academic writing.

More specifically, responses to Research Question 2 require both the construction of a writing assistance system, based on the context models, and evaluations of the system in terms of both its contexts models and its features.

1.8 THESIS OUTLINE

Having sketched out some of the background necessary for subsequent chapters, this section concludes Part I of the thesis, which ultimately touches on several different academic fields such as linguistics, natural language processing and educational technology. Part II will introduce the concept of context in writing, which is analyzed in terms of register, topic, and readability. Part III will introduce two writing assistance systems and explain how the models introduced in Part II can be effectively used to improve the systems. Finally, Part IV concludes the thesis with evaluations of its contributions.

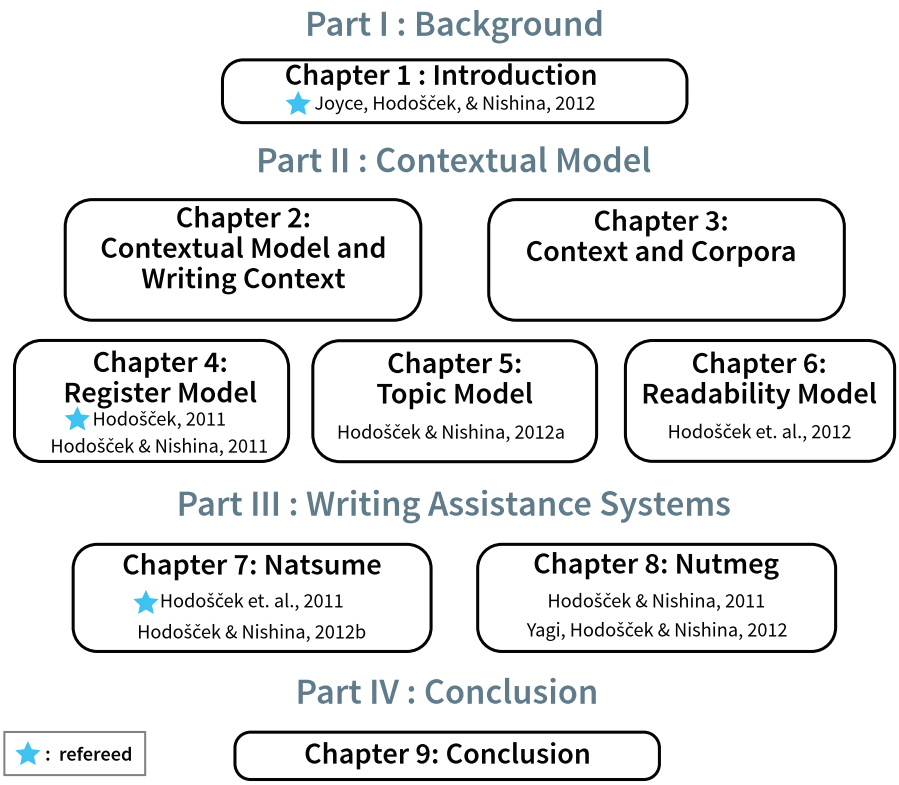


Figure 1: Schematic representation of the structure of the thesis.

Part II

CONTEXTUAL MODEL

2

CONTEXTUAL MODEL AND WRITING CONTEXT

2.1 WRITING AND COMMUNICATION

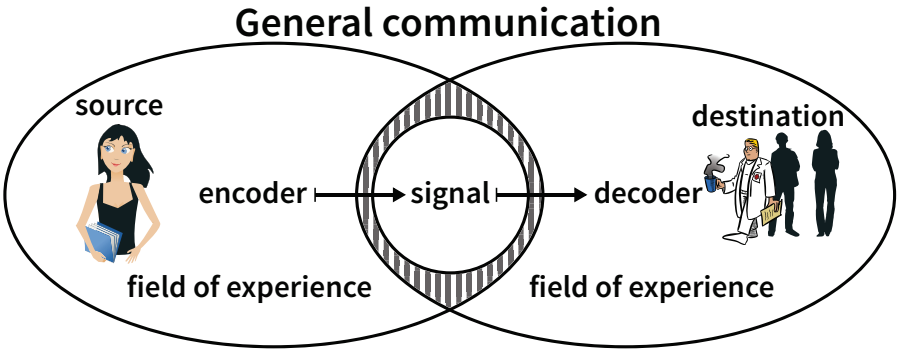


Figure 2: Model of general communication (adapted from Schramm (1997), p. 54).

In Wilbur Schramm's model of communication, the components of communication are the source, the signal, and the destination. In Figure 2, we see the source, a university student, encoding into the signal some concepts she wants to communicate to the destination, a university professor. The signal is then decoded by the destination. In this model, the intersection between the source and destination's respective fields of experience represents the common field of experience, which is the essential element for successful communication.

Given the definition of field of experience as the beliefs, values, experiences, and learned meanings of an individual (Schramm, 1955), clearly it is only possible for a signal to be successfully decoded when there is a significant degree of overlap between the respective fields of experiences that the source and destination bring to the communication.

As the form of communication examined in this thesis is written communication, the situation is somewhat more constrained than in the general model. → *Figure 3*
Specifically, it is one in which the writer is attempting to communicate with a reader, but cannot obtain any feedback—one-way communication. In order to increase the possibility of being correctly understood in such situations, writers must seek to accommodate the reader and attempt to narrow the gap between their respective fields of experience in the only way possible—by expanding their own field of experience.

The constrained version of the communication model has several direct implications for the writer and their field of experience:

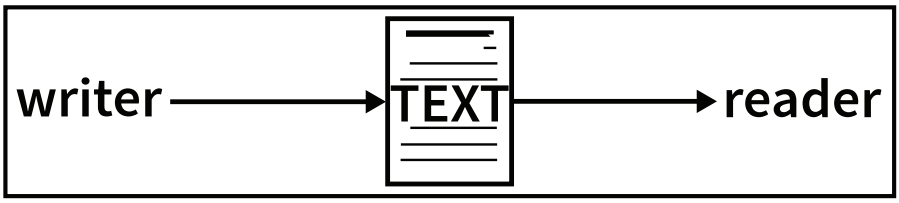


Figure 3: Constrained version of the communication model used within this thesis.

- The writer must consider the reader's perspective, which, within the model framework, refers to assuming, or targeting, to a certain level of knowledge on the part of the reader audience. Thus, the writer can only seek to convey information that falls within the field of experience that the readers are assumed to possess.
- Reflecting these assumptions, the writer must also employ appropriate scaffolding strategies, including their gradual removal, in order to facilitate the reader in reaching an appropriate interpretation of the writer's intended signal.

While the importance of the shared field of experience is clearly implied within Schramm's model of communication, this thesis specifically seeks to develop and elucidate the key notion of context that the shared field of experience represents, by focusing on the three contributing factors of register, topic, and readability. As illustrated in Figure 4, the model of context developed here consists of three distinct component models of register, topic, and readability, which are regarded as integral elements of the context model. Although these component models are examined in more detail in separate chapters later, that is more for clarity of exposition, as their modeling draws on different established NLP methods.

Within some approaches to register, such as that advanced by Biber and Conrad (2009), topic is situated as a separate but influencing factor within register. Similarly, in Halliday's Systemic Functional Grammar framework, register is defined in terms of field, mode, and tenor (Halliday & Matthiessen, 2004). Certainly, the interrelationship between register and topic is rather complex, and, indeed, while some NLP methods make no distinctions between them, others only focus on one dimension.

The example in Figure 5 contrasts two ways of writing about the same topic but using different registers. The first example is taken from a transcript of the Japanese Diet, and, given that it represents spoken language, it contains colloquial variations of adverbs, such as やはり /yahari/ "of course" (やっぱり /yappari/) and verb inflection contractions such as 〜ちゃ /~cha/ (では /dewa/). In contrast, the second example is an attempts to express the same sentence in a more formal, academic register. It, thus, contains more compact expressions,

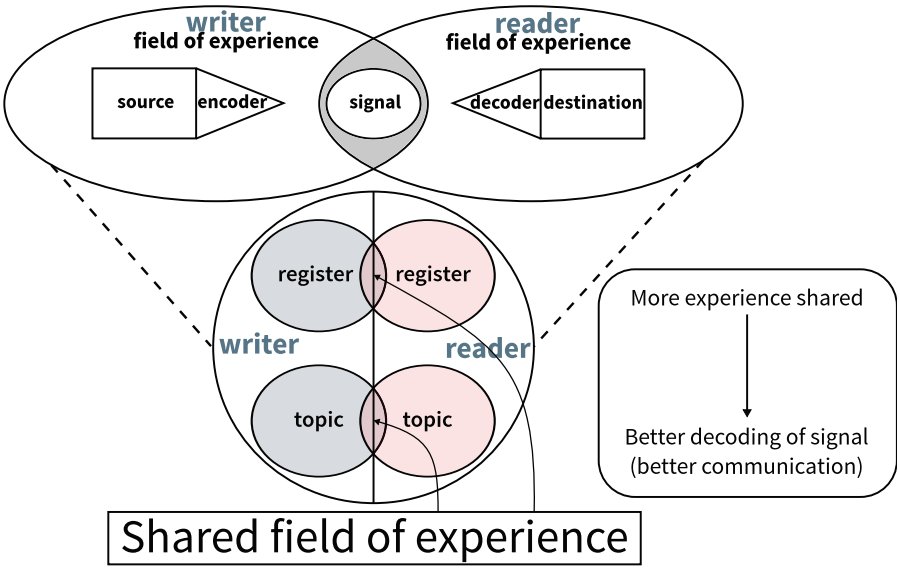


Figure 4: The context model developed in this thesis.

including the compound noun 共同研究 /kyōdōkenkyū/ “collaborative research”, and use of the である /dearu/ sentence final copula form.

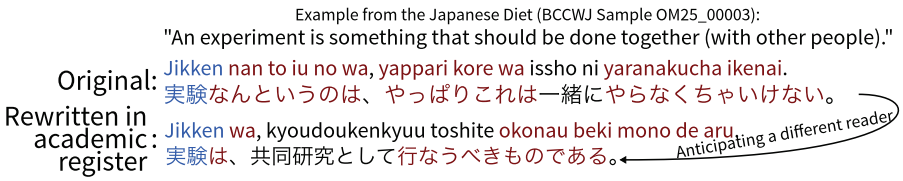


Figure 5: Contrast between two registers for essentially the same topic.

In this thesis, I adopt a practical approach towards the topic, in regarding it as a set of content words that can be expressed in many situations (i.e. different registers). Register, then, represents the writer’s repertoire of different expressions for a given topic, from which the writer selects the appropriate one according to the specific context/communication purpose/function of the communication. As components of field of experience, register and topic are discussed again in §4 and §5, which explains how they can be indirectly measured from the linguistic features extractable from written text (cf. Biber & Conrad, 2009).

A more detailed definition of topic as used here is provided in §5

2.2 REGISTER AND CONTEXT

Variations in language can be analyzed from several different viewpoints, depending on the distinction that one wishes to apply. The most general distinction is attributed to Saussure in his famous *Course in general linguistics*, where he

delineates language change into diachronic and synchronic change (1959, p. 22, 88, 110). When focusing on diachronic change, one attempts to trace the changes in language that occur with the passage of time. In contrast, when focusing on synchronic change, one is concerned with the changes that occur due to other factors, such as social, geographic, and functional changes.

Many studies of linguistic variation have tended to focus on language dialects, from either a social or geographical perspective. The social perspective is generally characterized by sociolinguistically-oriented studies that regard meaning-preserving phonetic variations in a language as an indicator of a social dialect (Labov, 1994). Studies from the social perspective usually assume that all dialects are communicatively equivalent and often limit the scope of their investigations to variations reflecting demographic differences such as gender, social position, ethnicity, and race. On the other hand, although studies from the geographic perspective also tend to focus on meaning-preserving phonetic variations, they usually limit investigations to variations associated with a particular location. One example of this orientation is the construction of dialect maps that document the variation in the pronunciation of /g/ between the velar plosive [g], nasal [ŋ], and others, in the word *kagami* (The National Language Research Institute, 1981). Recently, both perspectives have broadened the scopes of their investigations to cover various syntactic phenomena.

Following Biber and Conrad (2009), however, another approach to more recently emerge, which is essentially the one taken in this thesis, is to analyze language variation from a functional perspective. Thus, analyses of functional variation focus more on the contexts under which language exchanges occur, and variations in language are seen as being directly connected to the situation of language use. In this way, the approach requires an implicit mapping of the context to the actual selection of specific language patterns, which, in turn, correspond to both language-internal and language-external criteria. Language-internal criteria include linguistic features, such as the ratio of kanji in a sentence or the frequency of hedge words within a passage of text. Language-external criteria include various factors, such as media labels, author information, publication year, and other metadata, as well as consumer information. In terms of the constrained model of communication presented in §2.1, while language-internal criteria roughly equates to the shared field of experience, the language-external criteria encompasses the wider range of factors that constrain the context under which communication can occur.

The terms “register”, “genre”, “style”, “text type”, and “domain” have all been used within linguistic studies to refer to variations of one kind or another. Unfortunately, different scholars subscribe to different definitions of almost all of these terms and what one scholar might describe as register, another might describe as style or genre. For a comprehensive overview of the terms within linguistic studies, the reader is referred to Lee’s “Genres, Registers, Text Types, Domains,

Table 3: Tentative interpretations of register, genre, style, text type, and domain (Modified from Hodošček, 2010).

Term	Interpretation
Register	A variety of language associated with the specific situation of use. Example: register of written academic Japanese; classroom conversation
Genre	A category of language defined by a community, or associated with expected rhetorical structure and themes. Example: genre of Japanese research articles; crime novels
Style	Variations in language associated with an individual's "unique" uses of language. Example: sensationalist style; vague written style
Text type	A grouping of texts based purely on linguistic features. Example: informational text type
Domain	Text devoted to a single topic or a small set of related topics, often inside one genre. Example: domain of computational linguistics

and Styles" (2001). However, it is useful to briefly note how these terms can be interpreted with respect to (1) the internal-/external criteria contrast, (2) the incorporation/absence of topic, and (3) their characteristic linguistic features. A preliminary summary is offered in Table 3.

2.3 SYSTEMIC FUNCTIONAL LINGUISTICS (SFL)

Within Systemic Functional Linguistics (SFL), the functional bases of grammatical phenomena are described in terms of the INTERPERSONAL, TEXTUAL and IDEATIONAL metafunctions. They, in turn, relate to the three contextual vectors: *tenor*, *field*, and *mode*, that together comprise the context within the SFL framework. Increasingly, within the SFL framework, register has come to refer to the three contextual variables of field, tenor, and mode themselves (Hasan, 2009, p. 57). A conceptual image of the relationship between metafunction, context, and text is presented in Figure 6.

The textual metafunction relates to mode; the internal organization and communicative nature of a text. This comprises textual interactivity (disfluencies such as hesitators, pauses and repetitions), spontaneity (covering lexical density, grammatical complexity, coordination (how clauses are linked together) and the use of nominal groups) and communicative distance (cohesion).

The interpersonal metafunction relates to a text's aspects of tenor or interactivity. Like field, tenor comprises three component areas: the speaker/writer persona (stance, personalization and standing of the speaker or writer), social

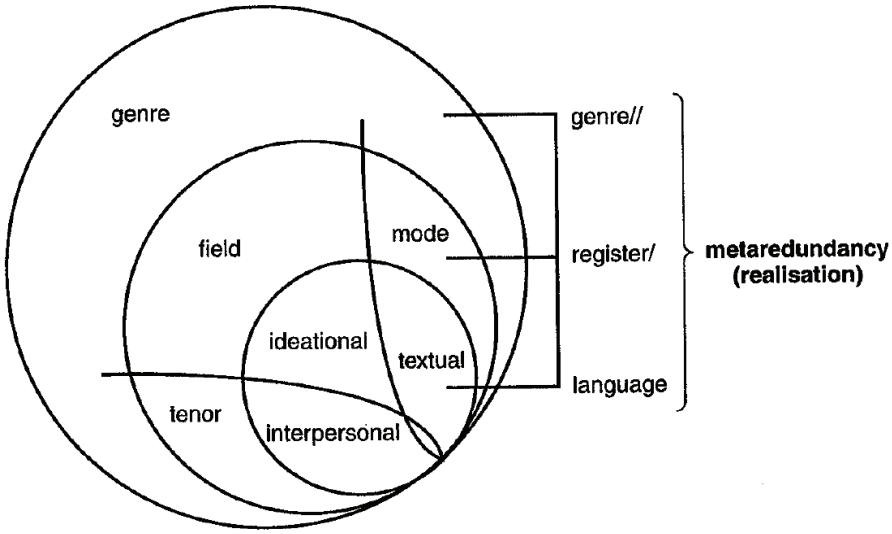


Figure 6: The relationships between metafunction, context, and text (Martin, 2009, p. 162).

distance (how close the speakers are to each other) and their relative social status (whether they are equal in terms of power and their knowledge of a subject).

The ideational metafunction is divided into two parts: the experiential and logical metafunctions. The ideational metafunction relates to the field aspects of a text, or its subject matter and the context of use. Field is divided into three areas:

- Semantic domain: the subject matter of a text through organizing its nominal groups (nouns / noun phrases) and its lexical verbs, adjectives and adverbs.
- Specialization: jargon or other technical vocabulary items.
- Angle of representation: types of processes, participants and circumstances.

Table 4: Metafunctions in SFL (Halliday & Matthiessen, 2004)

Metafunction	**Aspects of context**
Ideational (experiential, logical)	Field
Interpersonal	Tenor
Textual	Mode

The editors of the *Continuum Companion to Systemic Functional Linguistics* (2009, p. 246) define register as "a text type seen from the 'system' end, as a

functionally motivated subsystem within a language that is characterized by a 'general['] (or 'generalized') structure potential and by distinctive (usually quantitative) patterns of selection within the lexicogrammar and semantics." This definition is especially relevant for studies looking for ways to empirically quantify variation within the language in terms of some specific and concrete linguistic features. However, while the SFL framework offers a comprehensive theoretical device for the study of register, there have been far fewer serious attempts to apply the framework to actual language data compared to other approaches. Studies rooted within corpus linguistics, however, have tried to incorporate some of its elements and, over the course of the past two decades or so, there have been a few developments of empirical methods of capturing register variation based on exploratory multivariate statistics.

2.4 THE MULTI-DIMENSIONAL (MD) APPROACH

In contrast, the Multi-Dimensional (MD) approach to register analysis, pioneered by Douglas Biber in his book *Dimensions of register variation* (1995; also see Biber, 1993), adopts a multi-pronged exploratory approach to register variation within language. Specifically, the approach attempts to describe register variation in terms of the interactions between various linguistic features and their situations of use. Thus, the approach that has been used to investigate register variations within English can be summarized through the following six steps (reproduced from Biber (1995), Biber and Conrad (2009)):

1. An appropriate corpus is designed and collected based on previous research and analysis. The situational characteristics of each spoken and written register included in the study are documented.
2. Research is conducted to identify the set of linguistic features to be included in the analysis.
3. Computer programs are developed for automated grammatical analysis; the entire corpus of texts is analyzed to compute the frequency counts of each linguistic feature in each text.
4. The co-occurrence patterns among linguistic features are analyzed, using a factor analysis of the frequency counts.
5. Dimension scores for each text with respect to each dimension are computed; the mean dimension scores for each register are then compared to analyze the linguistic similarities and differences among registers.
6. The "factors" from the factor analysis are interpreted functionally as underlying dimensions of variation.

Clearly, however, the design and compilation of the corpus and associated NLP tools used in the feature extraction processes from steps 1.-3. require a wide variety of linguistic knowledge and research in order to yield significant results at the subsequent factor analysis steps. While the availability of large-scale corpora has ceased to be a major obstacle for conducting MD analyses, linguistic feature selection, extraction, and validation all still remain open problems, especially for languages other than English where MD analyses have yet to be undertaken. In other words, the amount of register variation that can be explained relies heavily on the careful selection of linguistic features, which, in turn, depends on the existence of detailed investigations of the linguistic features under consideration. Another problem that has been pointed out by Biber in *Dimensions of register variation*, especially for large comparative studies, is that quantitative changes in the distributions of lexicogrammatical features across registers, when considered individually, can quickly become far too complicated to interpret reliably (qualitatively) in terms of the specific situations of use (Biber & Conrad, 2009). However, since a register is characterized by a specific range, or pattern, of selections within the lexicogrammar, cooccurring lexicogrammatical features can be reinterpreted (qualitatively) as a small number of inherent dimensions of variation within lexicogrammar. The methodology employed to achieve this, namely, MD analysis, was developed by Biber (1995) and consists of the statistical factor analysis of the frequencies of lexicogrammatical features. The factors obtained from such analyses are then interpreted functionally as the underlying dimensions of register variation.

In a factor analysis, a large number of original variables are reduced to a small set of derived, underlying variables. For each factor (or dimension), there are both positive and negative features. When positive features occur with a high frequency within a text, negative features will occur with lower frequencies and vice versa. When all the dimensions are identified, a dimension score can be computed for each text or corpus, which can further be used to graphically represent their positions on various dimensions. Thus, linguistic variation can be described not in binary terms, but rather in terms of several continuous scales of variation that correspond to the identified dimensions.

As argued above, it has become possible, in principle, to utilize this approach for the Japanese language with the increased availability of large-scale Japanese language corpora. However, the main obstacle to such analyses is at the step of extracting the linguistic features, as it is still necessary to make suitable modifications in order to be able to extract the features that are salient for register variation within the Japanese language. As the automatic extraction of features from text remains a difficult task the higher one ascends up the lexicogrammar stratum, the SFL framework becomes even more attractive as an important source of inspiration for new directions to pursue towards realizing the corpus-based, exploratory approach.

2.5 SUMMARY

To recap, within the SFL framework, register can be represented through the three contextual vectors of field, tenor, and mode. In a similar vein, Biber has posited seven broad categories of situation type that describe the external characteristics of functional variation. A comparison of these two methods is provided in Table 5.

Table 5: Situational characteristics and the three contextual vectors

Biber and Conrad (2009, pp. 111–112)	Halliday and Matthiessen (2004)
Participants	Field
Relationships among participants	Field
Channel	Mode
Production and comprehension circumstances	Tenor
Setting	Tenor / field
Communicative purposes	Tenor
Topic	Mode / field / topic

As one concept closely related to register, it is vital to understand how the status of genre within both frameworks is intrinsically connected to both register and text.

As Martin (2009)'s depiction of the cline of realization in Figure 6 suggests, language is construed through the interpersonal, ideational, and textual metafunctions which, in turn, construe register through the contextual vectors of tenor, field, and mode and, thus, it is these contextual vectors that ultimately construe genre. Accordingly, within the SFL framework, the difference between genre and register is one of realization—they are on different *semiotic planes*.

Biber and Conrad (2009) provides a contrastive view of the three perspectives associated with language variation, as is shown in Table 6. Moreover, the definition of genre provided by the editors of the *Continuum Companion to Systemic Functional Linguistics* 2009 shares striking similarities with the one presented in Biber and Conrad (2009).

> [A genre is] a higher-level grouping of texts having the same compositional structure ("generic structure"), corresponding to rhetorical categories of procedural, expository, narrative and so on. The structure is specified in terms of a sequence of elements each having a distinct function with respect to the whole and each characterized by particular lexicogrammatical features.

Indeed, remarking on the similarities between these two frameworks, Biber and Conrad has stated that the "distinction between *register* and *genre* [...] clearly

Table 6: Relations between register, genre and style (reproduced from Biber and Conrad (2009, p. 16))

	Register	Genre	Style
Textual focus	sample of text excerpts	complete texts	sample of text excerpts
Linguistic characteristics	any lexicogrammatical feature	specialized expressions, rhetorical organization, formatting	any lexicogrammatical feature
Distribution of linguistic characteristics	frequent and pervasive in texts from the variety	usually once-occurring in the text, in a particular place in the text	frequent and pervasive in texts from the variety
Interpretation	features serve important communicative functions in the register	features are conventionally associated with the genre: the expected format, but often not functional	features are not directly functional; they are preferred because they are aesthetically valued

shares some characteristics with the use of the concepts in Systemic Functional Linguistics, especially with respect to the genre perspective emphasizing the conventional features of whole texts, while the register perspective emphasizes variation in the use of linguistic features" (Biber & Conrad, 2009, p. 22).

3 CONTEXT AND CORPORA

As the approach to writing assistance taken in this thesis is essentially data-driven, where corpora are used to model context, an important question that must be addressed is how can such data be used to effectively realize a writing assistance system. Clearly, in order to properly verify this approach, it is essential to utilize multiple corpora varying in terms of register, topics, and readability, which are important components of context.

This chapter introduces the linguistic resources used throughout the thesis. Native corpora are used to realize the data-driven learning (DDL) component of the developed writing assistance systems. Learner corpora are also introduced as they are used later in §8, when evaluating the register model. Finally, the chapter also introduced some dictionaries and natural language processing tools, with special focus on the representation and extraction of words and collocations in Japanese.

3.1 SCIENTIFIC AND TECHNICAL JAPANESE CORPUS (STJC)

A basic requirement for building a model of scientific and technical Japanese is the existence of a collection of authentic materials that adequately covers the target range of domains. Unfortunately, the data requirements for a writing assistance system for academic contexts are not fully satisfied by either the BCCWJ or the Japanese version of Wikipedia, which are both introduced in the succeeding sections. Although these corpora contain Japanese text covering many topics and registers, they do not offer a substantive and representative sample of scientific and technical Japanese, and no such corpus existed at the conception of this research. Therefore, a new corpus, called the Scientific and Technical Japanese Corpus (STJC), was constructed in order to meet these requirements.

The following criteria were used when choosing journals for inclusion into the corpus:

1. The journal specializes in some scientific or engineering field.
2. The journal is published by a society with at least a thousand members.
3. The journal has established review procedures.
4. The journal gave permission to use article texts within the developed systems.

Table 7: Journal citation index scores as of July 2013. Higher index scores generally indicate higher journal standing. Source: Google (2013)

Journal	h5-index	h5-median
J. of Natural Lang. Proc.	7	12
J. Inst. of El. Eng. of Japan B	8	9
J. Japan Soc. of Civil Eng. B	5	7
J. Japan Soc. of Civil Eng. A	5	6
J. Japan Soc. of Civil Eng. D	5	6
J. Japan Soc. of Civil Eng. C	4	5
J. Chem. Soc. of Japan	N/A	N/A
J. Inst. of El. Eng. of Japan A	N/A	N/A
J. Nippon Med. Sch.	N/A	N/A

Table 8: Composition of the Scientific and Technical Japanese Corpus.

Journal	Tokens	Chunks	Sentences	Sources
J. of Natural Lang. Proc.	1,655,975	541,277	45,885	201
J. Chem. Soc. of Japan	708,269	222,453	18,698	184
J. Inst. of El. Eng. of Japan A	640,259	206,464	17,057	163
J. Japan Soc. of Civil Eng. D	389,182	123,598	9,797	41
J. Inst. of El. Eng. of Japan B	241,303	78,669	6,521	50
J. Japan Soc. of Civil Eng. C	159,338	53,669	5,535	21
J. Japan Soc. of Civil Eng. A	156,644	50,679	3,873	34
J. Japan Soc. of Civil Eng. B	84,234	28,066	2,216	21
J. Nippon Med. Sch.	66,885	21,204	1,664	28
TOTAL	4,102,089	1,326,079	111,246	743

Currently, the corpus includes papers from the *Journal of Natural Language Processing*, the *Journal of the Japan Society of Civil Engineers*, the *Journal of Nippon Medical School*, the *Journal of the Chemical Society of Japan*, the *Journal of Environment and Natural Resources Engineering*, as well as the *Journal of the Institute of Electrical Engineers of Japan*. As illustrated in Table 7, six of the journals are listed within the top 100 of Google Scholar's Japanese index, which provides a measure of the relative impact of each journal. For comparative purposes, one may note that the top-scoring journal in the Google Scholar's Japanese index is the *Journal of Information Processing* (情報処理学会論文誌) with h5-index of 9 and h5-median of 14.

→ *Table 8*

Table 9: Composition of the BCCWJ sub-corpora.

Sub-Corpora	Approx. size	Year	Content
Publication	35 million words	2001-2005	Published books, magazines, and newspapers
Library	30 million words	1985-2005	Books cataloged at more than 13 public libraries in the Tokyo area
Special-purpose	35 million words	1975-2008	White paper text, Internet text, Diet minutes, best-selling books, etc.

3.2 BALANCED CORPUS OF CONTEMPORARY WRITTEN JAPANESE (BCCWJ)

The Balanced Corpus of Contemporary Written Japanese (BCCWJ) was developed as a five year project led by the National Institute for Japanese Language and Linguistics (NINJAL). Its aims were to construct a representative sample of written Modern Japanese from 1975 to 2005 (Maekawa, 2007b; Maekawa, 2007a; Maekawa, 2011). Additionally, the corpus was to be compiled as a tagged corpus and have sufficient scale and coverage of the sub-varieties of written language to offer a representative sample of the Japanese written language. The first official release of the corpus in 2012 (version 1.0) includes three sub-corpora of roughly equal size.

→ *Table 9 The composition of media in the BCCWJ corpus is presented in Table 10 in terms of token, chunk, sentence, paragraph and source counts. While the term token is commonly used to denote a word, a token here is defined as a Short Unit Word. Refer to §3.5 for more information.*

Reflecting their different purposes, different sampling methods were used in the compilation of each sub-corpus. The publication sub-corpus was sampled during a shorter time span than the library sub-corpus in order to be a representative sample of books, magazines, and newspapers that were sold and presumably consumed by the public. The library sub-corpus is more representative of the books for which public demand was deemed sufficient for copies to be available across many libraries within the Tokyo region. In contrast to the first two sub-corpora, the special-purpose sub-corpus serves more for comparative purposes, for, although not necessarily totally representative of each component register, the overall variety of registers is much greater than for the other two.

The BCCWJ corpus is encoded in an XML format that includes sentence and paragraph-level metadata annotations, such as tags for speech, titles, and quotations. As can be seen in the XML sample shown in Figure 7, the inclusion of quotations and other tags that relate to the register of a sample make this data particularly useful for investigating or controlling for register variation within the samples, such as, for example, controlling for the presence of dialog within works of fiction.

```
<?xml version=”1.0” encoding=”UTF-8”?>
<sample sampleID=”LBa1_00004” version=”1.0” type=”variableLength”>
<article articleID=”LBa1_00004_V001” isWholeArticle=”false”>
<info arg=”article/@isWholeArticle” value=” 完結-不完全” />
<titleBlock>
<title>
<sentence type=”quasi”>第十　復興の時代</sentence>
<br type=”automatic_original” />
</title>
</titleBlock>
<cluster>
<titleBlock>
<title>
<sentence type=”quasi”>一　再婚</sentence>
<br type=”automatic_original” />
</title>
</titleBlock>
<paragraph>
<sentence>　大正六年（一九一七）二月十一日には『ときのこゑ』禁酒号を発
行した。</sentence>
<sentence type=”quasi”>山室は、</sentence>
<br type=”automatic_original” />
</paragraph>
<quotation>
<citation>
<paragraph>
<sentence>　紀元節の目出度い祭日に世間では祝盃をあげて、これを祝う
といふ際、私共は却つて禁酒の主義を拡張するために此の日を用いようと
する。</sentence>
<br type=”automatic_original” />
</paragraph>
</citation>
</quotation>
...
```

Figure 7: Sample from BCCWJ XML format file.

Table 10: Breakdown of major BCCWJ media labels.

Media label	Tokens	Chunks	Sentences	Paragraphs	Sources
Books	70,472,742	24,654,541	3,155,084	1,552,490	22,058
Yahoo! Blogs	13,212,757	4,275,507	943,646	783,871	52,676
Yahoo! Q&A	12,088,127	3,947,304	780,510	624,616	91,445
Minutes of the Diet	5,600,649	1,858,486	139,802	45,810	159
White papers	5,495,254	1,705,664	139,373	101,587	1,500
Magazines	5,114,752	1,718,895	281,765	155,260	1,996
Gov. pamphlets	4,739,306	1,295,337	255,841	209,679	354
Legal documents	1,206,481	401,505	33,289	25,364	346
Textbooks	1,126,214	392,651	63,370	45,952	412
Newspaper	1,036,285	343,304	50,960	26,546	1,473
Verse	237,685	106,265	18,974	18,974	252
TOTAL	120,330,252	40,699,459	5,862,614	3,590,149	172,671

3.3 JAPANESE WIKIPEDIA

Currently, the largest freely available corpus of Japanese text is the Japanese version of Wikipedia[1]. The decision to include it, in addition to the STJC and BCCWJ, was made for two reasons.

The first reason is that for many tasks, the quantities of data provided by the STJC and BCCWJ are insufficient. The amount of data required to accurately capture some language pattern is a function of both the number and the frequencies of its components, i.e. their complexity—also commonly referred to as the "sparseness problem"[2]. Due to the nature of the data used within Natsume, which includes triplet combinations of nouns, case particles, and verbs, the amount of extractable data is only really sufficient for the most common expressions. Moreover, other NLP technologies deployed within the Natsume system, such as `getassoc`[3], are more precise at the scales of data available with Wikipedia. The total size of the Wikipedia corpus is around four times larger than the BCCWJ. → *Table 11* However, one unfortunate side effect of including Wikipedia is that for many less frequent collocations, the only information available is from Wikipedia, making it impossible to compare between corpora.

Another requirement of the project is that there are no legal restrictions relating to displaying online examples obtained from the corpora texts. The permissive license of Wikipedia allows for the display of all text sentences as example

1 Accessible from https://ja.wikipedia.org/.

2 An intuitive way of thinking of this is of conducting a search query on the Internet, where each consecutive word entered into the search box returns fewer results.

3 Available from http://getassoc.cs.nii.ac.jp/.

Table 11: Various unit sizes for the Wikipedia corpus (as of 7/3/2013).

Corpus	Tokens	Chunks	Sentences	Paragraphs	Sources
Wikipedia	416,472,135	131,701,931	15,915,085	6,525,344	853,975

sentences in the systems. As a community effort, the quality of the Wikipedia data can vary considerably, and in some cases, texts contain ungrammatical elements or overly complex sentences.

3.4 NATANE LEARNER CORPUS

Natane is a Japanese language learner corpus annotated for learner errors. Compared to solely utilizing native corpora, the principle benefit of also incorporating a learner corpus from the perspective of developing a writing assistance system is that the learner corpus can highlight the kinds of errors that learners tend to make. For example, in the case of Natane, by comparing learner error tendencies based on their first language, it is possible to provide customized guidance with lesson plans that are more appropriate to the learner's first language.

Compared to native corpora, learner corpora tend to be much smaller in size and variety. This is due to the difficulties of obtaining learner writing, which in most cases is specifically elicited for the construction of a corpus rather than collected from readily-available sources, as was the case in the construction of the BCCWJ or STJC. Another important differentiator is the inclusion of error annotations and background information about the learners who generated the original texts.

The end goal of the on-going construction of the Natane corpus is to compile a database that is both well-formed and sufficient in size for the automatic identification and correction of writing errors. It should be noted that, even though it is possible to construct relatively simpler systems such as the spellcheckers, co-occurrence checkers, and writing style checkers, it is far more difficult, even for state-of-the-art NLP systems, to engineer features that require more sophisticated understandings of semantics and discourse. As more features are annotated within corpora, it is likely that the number of automatic correction targets will increase.

As an on-going joint project involving several Japanese language teachers, the collection and annotation of the corpus initially progressed through the following stages:

1. Collection of learner essays and their transcription.
2. Initial annotation of learner errors using Excel (Cao & Nishina, 2010; Cao, Kuroda, Yagi, Suzuki, & Nishina, 2010).

3. Analysis of the initial annotations leading to specification of framework for classifying learner errors (Cao, Kuroda, Yagi, & Nishina, 2011; Cao, Yagi, Kuroda, & Nishina, 2012).
4. Use of the multipurpose annotation tool Slate[4] for error tagging (Kaplan, Iida, Nishina, & Tokunaga, 2012).

For more detailed information about each stage in the construction process, the reader is referred to the papers cited above.

3.4.1 *Collection*

The essays were collected from undergraduate and graduate students as well as students attending Japanese language schools. All essays were written on a specific topic, though not all topics were the same. Data concerning learner age, nationality, university level, first language, major, and Japanese language learning experience, as well as other background information, was recorded with the essay. Additionally, the learners signed a waiver authorizing the anonymous use of their essay within Natane, which includes a web interface with full-text search features[5]. Although more than 5,000 sentences have been collected, currently, only around 3,500 have been annotated. In its present state, Natane consists of 285 essays obtained from 192 learners, totaling 205,520 characters. From a total of 9,041 annotations, there are 6,789 learner errors. The distribution of learners by their first language, as shown in Table 12, is biased towards Mandarin Chinese speakers, who account for more than half of the learners and the essays. The remaining languages are predominantly from Asia.[6]

3.4.2 *Initial Annotation*

While error classification frameworks for languages such as English and French already exist (Dıaz-Negrillo & Fernández Domınguez, 2006; Granger, 2003; L'Haire & Faltin, 2003), there was no preexisting comprehensive error annotation scheme or descriptive framework for Japanese language learner errors. Given the absence of a suitable framework, the project decided to proceed with constructing one by drawing on both previous research and the teaching experience of the annotators (Cao & Nishina, 2010). During the initial annotation process, it became clear that there were two major kinds of errors. The first were unambiguous errors (*ordinary errors*), while the second were errors that the annotators regarded as being unnatural language uses (*unnatural production*

4 More information on Slate is available on its homepage: http://www.cl.cs.titech.ac.jp/slate/.
5 Available at http://hinoki.ryu.titech.ac.jp/natane.
6 Current statistics shown in Table 12 are available on the Nutmeg website (http://hinoki.ryu.titech.ac.jp/natane/stats).

Table 12: Distribution of Natane essays by first language and gender.

First language	Male	Female	Unknown sex	Total
Mandarin Chinese	62	64	26	152
Marathi	6	23	7	36
Vietnamese	18	9	0	27
Korean	24	3	7	34
Spanish	2	0	0	2
Malay	8	0	0	8
Slovenian	7	0	0	7
Hungarian	1	0	0	1
Thai	1	0	0	1
Unknown	5	0	12	17
TOTAL	133	90	62	285

errors), but not ungrammatical. The ordinary errors include deviations from standard orthography, syntactic function (voice, tense, aspect, and modality), conjugation, and subject-predicate incongruity. They are typically easy to annotate and occur frequently. In contrast, the unnatural production errors include word choice, addition, or omission of text units (phrase, paragraph, etc.). They are typically less frequent, and being harder and more subjective to annotate, there tends to be less agreement between annotators.

3.4.3 *Framework of Error Classifications*

The feedback gained from the initial annotation process was crucial for refining the framework of error classifications (Cao et al., 2011). The resultant error annotation framework is essentially hierarchical in nature, and is able to take into account different viewpoints regarding the learner errors, as well as allowing for the systematic annotation of such errors (Yagi & Suzuki, 2012). The hierarchy consists of up to four levels, with the higher levels corresponding to courser, more abstract categories, and branching out into three principal dimensions:

1. Error domain: linguistic level of the error (i.e. phoneme, word, phrase, …, discourse; the word tag is further classified into word classes like noun, verb, etc.). → *Figure 8*
2. Error category: type and form of the error. → *Figure 9*
    - Type: addition, omission, word order, deviation from standard orthography, etc.
    - Form: conjunction, conjugation, collocation, and (Japanese letter) script.

3. Error source: reason or background for the error (i.e. annotator's subjective opinion about the source of the error—register and style mismatch, coherence, first language interference, etc.). → *Figure 10*

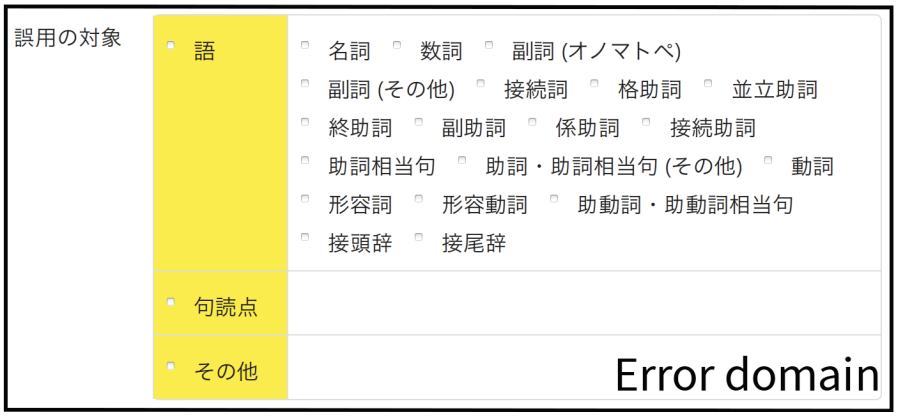


Figure 8: Error classification hierarchy: error domain.

3.4.4 *Other Japanese Language Learner Resources*

Although there are still relatively few other Japanese language learner corpora, the situation has improved somewhat recently, with the release of the following resources: the Learner's Language Corpus of Japanese[7], Teramura corpus[8], NINJAL's learners corpus[9], and the JC Corpus[10]. The major difference between Natane and these learner corpora is that, because they are more focused on the annotation of grammatical errors, they have less comprehensive error classification frameworks than the one used in Natane. That is especially important for the present research, as errors related to register misuse included in Natane can be used to evaluate register identification systems. → *Section 8.2*

3.4.5 *Conclusion*

Natane is a learner corpus that has many potential uses, although Japanese language educators and NLP researchers are naturally its primary target users. Using the search interface for the Natane corpus, Japanese language educators can utilize Natane to find examples of learner errors. Moreover, the data provided is particularly useful for analyzing the error tendencies due to first language

7 More information available from http://cblle.tufs.ac.jp/llc/ja/.
8 More information available from http://teramuradb.ninjal.ac.jp/.
9 More information available from http://jpforlife.jp/taiyakudb.html.
10 More information available from http://www34.atwiki.jp/jccorpus/pages/21.html.

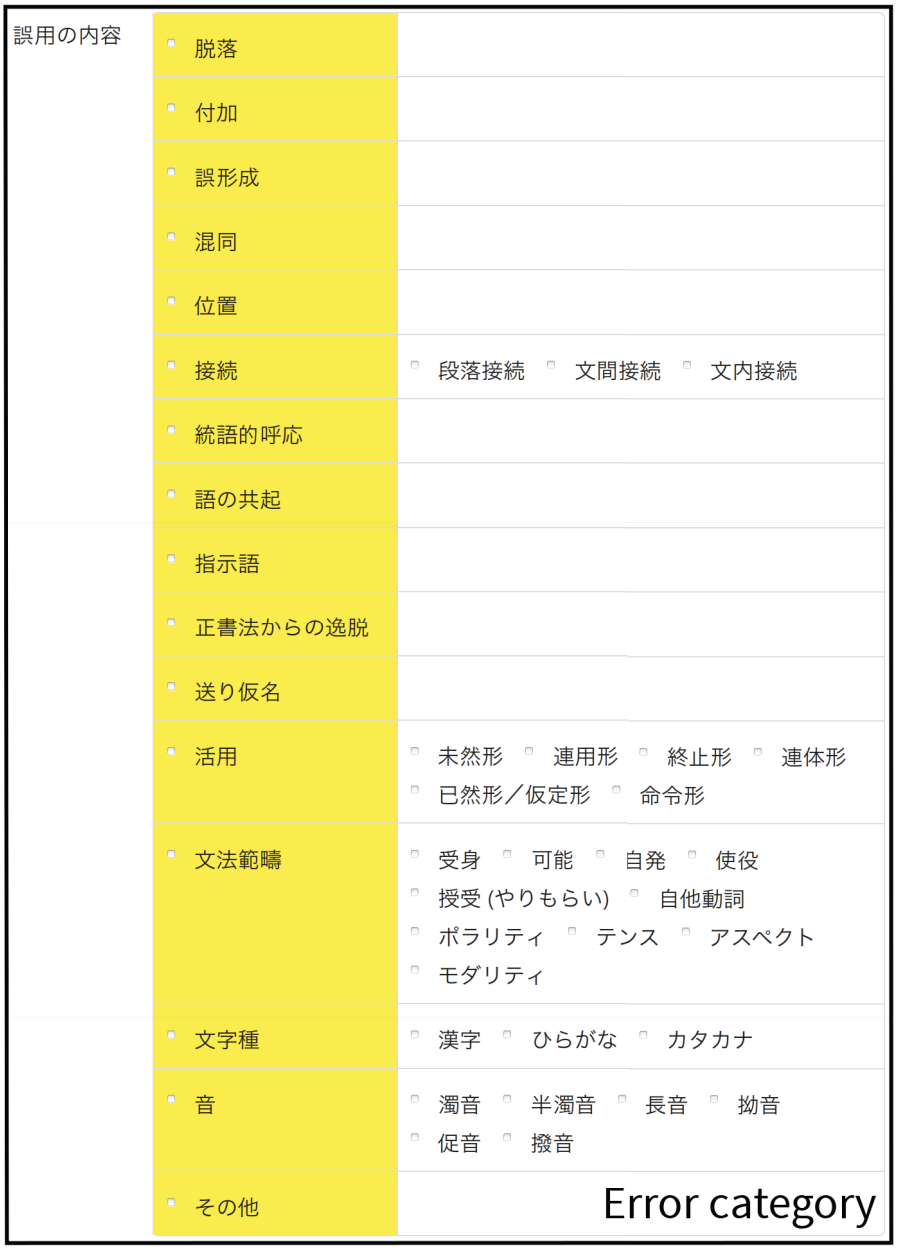


Figure 9: Error classification hierarchy: error category.

Figure 10: Error classification hierarchy: error source.

interference, as well as for helping to observe the overall language acquisition process. The latter usage is primarily aimed at NLP and machine learning applications, where Natane can be applied to the construction of novel error correction systems. An example of one such system is the Nutmeg writing assistance system introduced in §8.

While the availability of more Japanese language learner corpora is undoubtedly a welcome situation, given that they all use different frameworks of error classification, one pressing issue is for some common standard to emerge. It is hoped that the present research, in particular, can contribute to identifying some of the issues surrounding the annotation of register-related errors.

Refer to §8 for a discussion on some issues relating to register annotations

3.5 UNITS OF LANGUAGE

The concept of the word unit in Japanese is not as straightforward as in many Western languages, in large part because the Japanese writing system does not employ spaces to delimit word boundaries. One of the goals of the BCCWJ project was to provide a standardized unit of language that could simplify research on language (Maekawa, 2007b). The basic word unit taken up by the project was the Short Unit Word (SUW), which represents a relatively short unit corresponding to one morpheme. The second unit, called the Long Unit Word (LUW), is closer to the 文節 /bunsetsu/ "phrase" that consists of an independent element (noun or verb) and a dependent element (particle or inflection suffix). The potential benefit of adopting two language units is in allowing different levels of linguistic analysis. For lexical and morphological research involving concordance searches of the corpus, the SUW provides the greatest flexibility for identifying and selecting appropriate examples, while the LUW may be a more appropriate unit for investigating some issues such as lexical comparisons within genre-related studies.

→ *Table 13*

Table 13: Relationship between SUWs, LUWs, and chunks (Source: Yomiuri shimbun (evening edition), 4/28/2004; BCCWJ sample ID: PN4c_00026).

Chunk	今回、		この	ホテルを		使って		大型夜景鑑賞イベントを					企画した。			
LUW	今回	、	この	ホテル	を	使っ	て	大型夜景鑑賞イベント				を	企画し		た	。
SUW	今回	、	この	ホテル	を	使っ	て	大型	夜景	鑑賞	イベント	を	企画	し	た	。
Reading	Konkai	,	kono	hoteru	o	tuka	tte	ōgata	yakei	kansyo	ibento	o	kikaku	si	ta	.

Table 13 shows how a sentence can be broken down into chunks, LUWs, and SUWs. Although based on the bunsetsu, the process of identifying LUWs is rather complex because it involves both top-down and bottom-up processing. In the initial top-down parsing step, a sentence is chunked into bunsetsu that are analyzed in terms of SUWs and in the subsequent bottom-up parsing step, LUWs are constructed from component SUWs (independent and dependent elements). Accordingly, LUWs include combinations formed from SUWs plus common conjugations, inflections and affixes, and combinations formed by joining noun and verb elements into compound units.

Specifically constructed for morphological parsing of the BCCWJ corpus, the morphological parser dictionary UniDic is a hierarchical dictionary that essentially adopts SUWs for its dictionary entries ("UniDic Project Top Page," 2013). The top-level word classes of the UniDic hierarchy are nouns (名詞 /meishi/), pronouns (代名詞 /daimeishi/), verbs (動詞 /dōshi/), i-adjectives (形容詞 /keiyōshi/), na-adjectives (形状詞 /keijōshi/), adverbs (副詞 /fukushi/), prefixes (接頭辞 /settōji/), suffixes (接尾辞 /setsubiji/), interjections (感動詞 /kandōshi/), particles (助詞 /joshi/), auxiliary verbs (助動詞 /jodōshi/), pre-nominals (連体詞 /rentaishi/), conjunctions (接続詞 /setsuzokushi/), symbols (記号 /kigō/), punctuation (補助記号 /hojokigō/), and white-space characters (空白 /kūhaku/) ("UniDic Project Top Page," 2013). As LUWs are constructed from SUWs, the word class hierarchy is identical for the most part. One important difference, particularly relevant for the present study, is that ambiguous SUWs that, depending on context, can be either a noun or adverb, or a noun or verb, are disambiguated in the process of becoming LUWs.

The relationship between SUWs and LUWs may be further elucidated with reference to the notion of chunks within the Japanese dependency grammar analyzer CaboCha[11] (Kudo & Matsumoto, 2013). CaboCha's dependency grammar consists of chunks linked by grammatical dependency. The chunks consist of either one or more morphemes (SUWs) that can be separated into content and grammatical or functional portions. Content portions roughly correspond to LUWs, while there is no direct analog to the grammatical and functional portions, which can comprise several SUWs or even several LUWs (commonly consisting of particles), as illustrated in Figure 11.

11 Available from http://code.google.com/p/cabocha/.

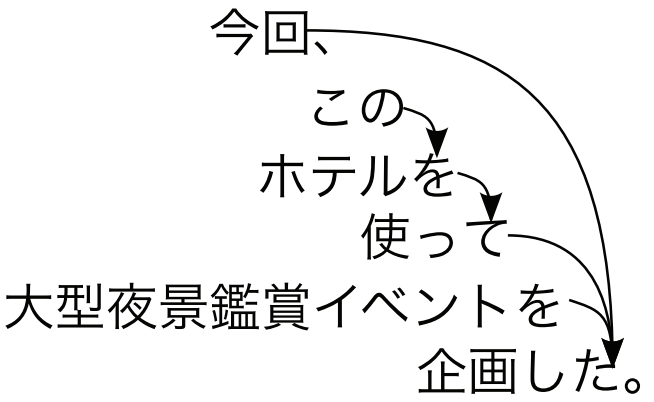


Figure 11: Chunks from Table 13 arranged in the dependency grammar of CaboCha.

4

REGISTER MODEL

This chapter provides an outline of the research project's attempt to model the register component using word class-based linguistic features.

4.1 MODIFIER-VERB RATIO (MVR)

The modifier-verb ratio, commonly abbreviated as MVR, was proposed in 1965 by Kabashima and Jugaku as part of a study on Japanese stylistics. More recently, as more sophisticated language processing tools and larger, more varied, corpora have become available, word class ratios have been reexamined in studies such as that of Fujiike, Konishi, Ogura, Ogiso, and Koiso (2011). This section attempts to explore the variability in register between and within the media in the BCCWJ, by observing the distribution of this linguistic-internal measure.

According to Kabashima and Jugaku, texts can be classified on a scale ranging from summative (要約的 /yōyakuteki/) to descriptive (描写的 /byōshateki/) (1965). Summative texts convey only the bare minimum—the skeleton or frame—of what they are describing. In contrast, descriptive texts specify in detail what they are describing, making the reader feel as if he is part of the situation described. Kabashima and Jugaku (1965) further characterizes descriptive texts as either active (動き描写的 /ugokibyōshateki/) or static (ありさま描写的 /arisamabyōshateki/).

The modifier-verb ratio was first introduced as a quantitative method of classifying texts into these categories using only the ratio of modifier to verb counts. Kabashima and Jugaku define modifiers as consisting of adjectives, adverbs, and pre-nominals (1965, p. 122). Once words are classified according to their word classes and ratios for each word class are calculated, one can then calculate the modifier-verb ratio by dividing the ratio of modifiers with the ratio of verbs in a text:

$$MVR = 100\frac{\text{modifiers}}{\text{verbs}} \tag{1}$$

In summary, texts with a high noun ratio and low modifier-verb ratio tend to be summative, while those with low noun ratios tend to be descriptive. Furthermore, descriptive texts with low modifier-verb ratios tend to be active, while those with high modifier-verb ratios tend to be static.

→ *Figure 12*

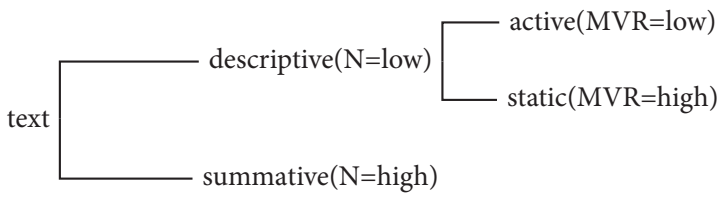


Figure 12: Categorization of texts using noun and modifier-verb ratios (adapted from Kabashima and Jugaku, 1965, p. 25; N and MVR information added by author).

4.2 MVR EXTRACTION

Modifier-verb ratios were computed for a pre-release version of the BCCWJ, as follows. First, using plain text samples from the BCCWJ, sentences were split based on a set of common sentence delimiters (from the set of both half- and full-width characters ".!?。 ！ ？"), except for when a delimiter was encountered inside a quotation. Next, the sentences were morphologically analyzed into SUWs using the morphological parser MeCab (version 0.98) and UniDic dictionary (version 2.1.0) (Kudo, 2013; "UniDic Project Top Page," 2013; NINJAL [National Institute for Japanese Language and Literature], 2011). In the next step, the SUW data was fed into the LUW analyzer Comainu (version 0.53a), which produced morphologically parsed Japanese text with LUWs as the base unit.

Finally, following Kabashima and Jugaku (1965), all word classes were coded and counts were made for the five classes of modifiers (M), nouns (N), verbs (V), interjections (I), and other (O). Using these frequency counts, it was straightforward to calculate the modifier-verb ratio as $MVR = 100\frac{M}{V}$. Special care needed to be taken for samples that contained no verbs, where it was not possible to calculate a modifier-verb ratio, and such samples were treated as outliers (572 and 492 samples from Yahoo! Q&A and Yahoo! Blogs, respectively, were treated as such).

Kabashima and Jugaku (1965)'s method of sampling of materials differs slightly from the one used for the BCCWJ, because in their sampling method, they took random sentences from each book, to yield an average word class ratio of each book. In contrast, for at least the library and publication sub-corpora in the BCCWJ, the sampling was done on the whole body of material, with samples taken in fixed- and variable-length chunks from random books and other media (Maruyama et al., 2010). Thus, while the sample may be less than representative of the sampled work, when taken together, the samples do constitute a representative sample of the complete body sampled.

4.3 RESULTS

Compared to Fujiike et al. (2011), whose study was based on smaller sample sizes and used a scatterplot to visualize noun and modifier-verb ratios, the relationship between noun and modifier-verb ratios is plotted for each media using a bivariate visualization method called a bagplot. The bagplot, first proposed by Rousseeuw, Ruts, and Tukey (1999), is a 2D generalization of a boxplot used to analyze the relationship between two variables. According to Rousseeuw et al. (1999), "the bagplot visualizes the location, spread, correlation, skewness, and tails of the data." More specifically, it consists of a bag containing 50% of the data points, a fence (computed by magnifying the bag by a factor of 3) that separates inliers from outliers, and a loop containing the points outside of the bag but inside the fence. In addition, the central point for each bagplot is defined as the point with the highest halfspace depth and is denoted by a star-shaped point, while any point outside the fence is considered an outlier. → *Figure 13*

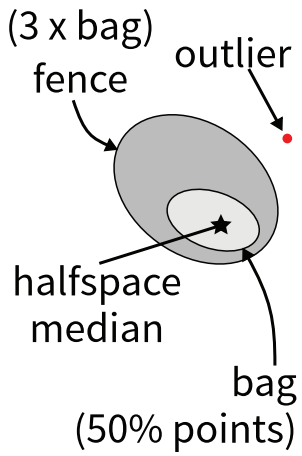


Figure 13: Visual explanation of bagplots.

The ratios computed by the method outlined in §4.2 were used to plot the bagplots using the `aplpack` package for the statistical programming environment R (Wolf & Bielefeld, 2010; R Core Team, 2013). Figure 14 shows the bagplots plotting the relationship between the noun and modifier-verb ratios for each media, here limited to noun ratios of 0-60% and modifier-verb ratios of 0-200 for clarity.

Furthermore, for easier comparisons and quantitative assessments of each media, Table 14 presents the halfspace median of N and MVR, the percentages of samples inside a bag, the percentages of samples outside a bag but inside a fence, as well as the percentage of outliers for each media.

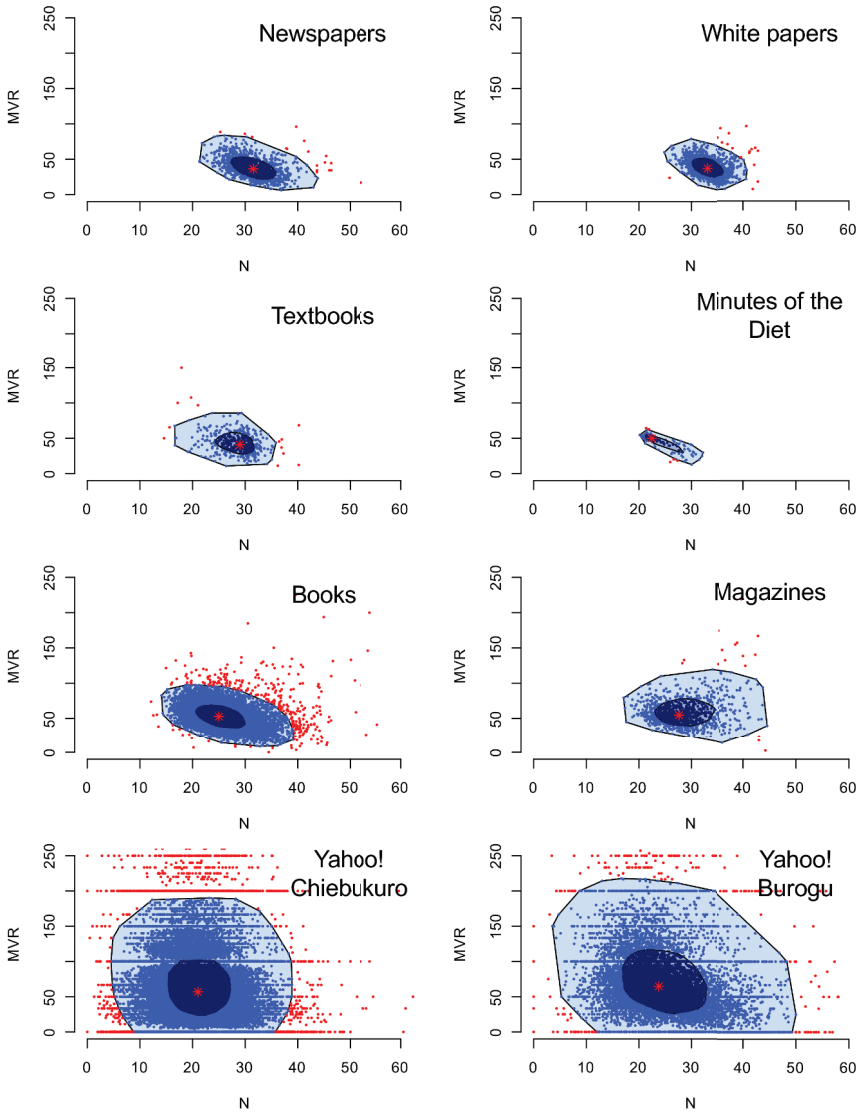


Figure 14: Bagplots for the noun ratio (N) to the modifier-verb ratio (MVR) for each media.

Table 14: Summary of the bagplot statistics for each media, sorted by MVR median.

Media	Bag(%)	Fence(%)	Outliers(%)	N median	MVR median
Newspapers	51.17	47.45	1.37	31.50	36.35
White papers	49.47	49.40	1.13	33.07	36.93
Textbooks	49.69	47.62	2.69	29.00	40.82
Minutes of the Diet	49.06	48.43	2.52	22.46	49.86
Books	52.36	46.22	1.41	25.00	52.86
Magazines	50.33	48.14	1.53	27.63	54.96
Yahoo! Q&A	53.91	41.35	3.66	21.01	56.82
Yahoo! Blogs	48.51	43.28	3.71	23.76	64.77

4.4 DISCUSSION

In general, the negative correlation of noun ratios with modifier-verb ratios observed by Kabashima and Jugaku (1965) was re-confirmed for the BCCWJ, which is indicated by the consistently downward orientations of the bags. The media of Magazines was the only exception to this tendency and merits further investigation. Interestingly, Magazines have both a relatively high noun ratio as well as a high modifier-verb ratio, a combination not adequately treated in Kabashima and Jugaku (1965). Although both Internet corpora, Yahoo! Blogs and Yahoo! Q&A, showed the largest bag and fold areas, as well as the highest outlier percentages, Yahoo! Q&A, in particular, has the highest concentration of samples in the bag and the least between the fence and the bag, suggesting a different distribution of variations in the word classes compared to the other media. Not surprisingly, White papers, Newspapers, and Textbooks have the highest noun ratios, as well as the lowest modifier-verb ratios, classifying them as summative texts, while the Books media has an average noun ratio, but a higher than average modifier-verb ratio.

4.5 CONCLUSION

This section has attempted to shed some light on the ways in which the modifier-verb ratio can be applied to the modeling of inter- and intra-corpus register differences. More specifically, it briefly touched on the issues of extracting word class ratios pertaining to word classes and word units in the morphological parser dictionary UniDic and Comainu LUW parser. From using bagplots to visualize the distributions of word classes within the BCCWJ, it was possible to essentially replicate the negative correlations between noun ratios and modifier-verb ratios observed by Kabashima and Jugaku (1965). Finally, the combination of the bagplot and modifier-verb ratio revealed clear differences on both the

summative-descriptive and static-active scales between media, while also providing insight into the amount of variation inherent inside a particular media.

4.6 FUTURE WORK

Modifier-verb ratios are a relatively simple index of the variety observable both inside and between corpora, and, clearly, more needs to be done to extend Kabashima and Jugaku's analysis to new varieties of writing, such as those found in Internet corpora, and to possible investigations of the positive correlation between nouns and MVR observed for Magazines. Moreover, comparisons with other measures, such as lexical density, for example, in Halliday (2009, p. 75-77), or the multidimensional feature approach outlined in Biber (1995), should be attempted.

As a word class based measure, one may expect the modifier-verb ratio to be relatively robust for topic changes in an otherwise situationaly homogeneous register. However, this could be further investigated by, for example, using the Nippon Decimal Classification (NDC) library classification system supplied for the Books media, the topical information available for the Internet corpora, as well as a model of topic as introduced in the next chapter, for a deeper understanding of the intra-corpus variation phenomena. Finally, although the BCCWJ contains material sampled from a relatively short timespan of up to 30 years, there is no guarantee that any inferences derivable from the empirical data can be attributed to diachronic differences.

5 TOPIC MODEL

Within this thesis, topic is used to refer to sets of words corresponding to the informational content in a text. This, in contrast, to the more commonly encountered definition as the grammatical subject of a sentence. Thus, the term topic refers to more than just one word, but to a set of words connected by an underlying concept.

5.1 LATENT DIRICHLET ALLOCATION TOPIC MODEL

Latent Dirichlet Allocation (LDA) is a type of probabilistic topic model. From the perspective of document classification, it is an unsupervised approach to annotating document collections with thematic information, while from the perspective of discovery, it can help the process of hypothesis formation concerning some aspects of documents (Blei, 2012). Blei (2012) offers the following comments of explanation concerning the intuitions behind latent Dirichlet allocation. From the assumption that some number of "topics," as distributions over the words, exist for a whole collection, it is possible to select a distribution of the topics, then, select a topic assignment from the distribution, and, finally, select a word from the selection of corresponding topics. Thus, the overall relationships between documents, topics, and words within the LDA topic model can be represented as the graphical model in Figure 15. Taking an example of a posting to Yahoo! Blogs, Figures 16 and 17 also illustrate the relationships between documents and topics and between topics and words, respectively.

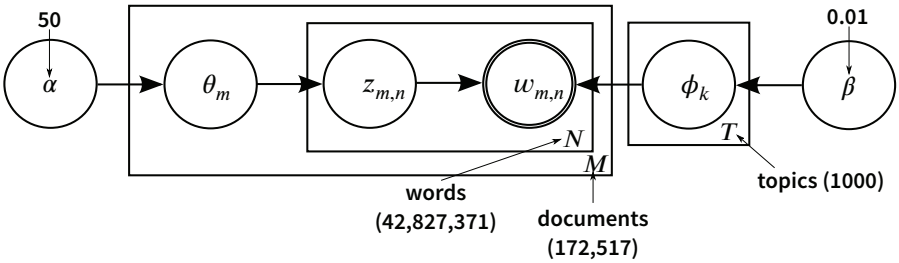


Figure 15: Graphical model of Latent Dirichlet Allocation topic model used in Hodošček and Nishina (2012a).

As LDA uses words as features, it is natural to ask questions about what kinds of words should or should not be included. From the perspective of attempting to maintain a practical distinction between topic and register, it is important

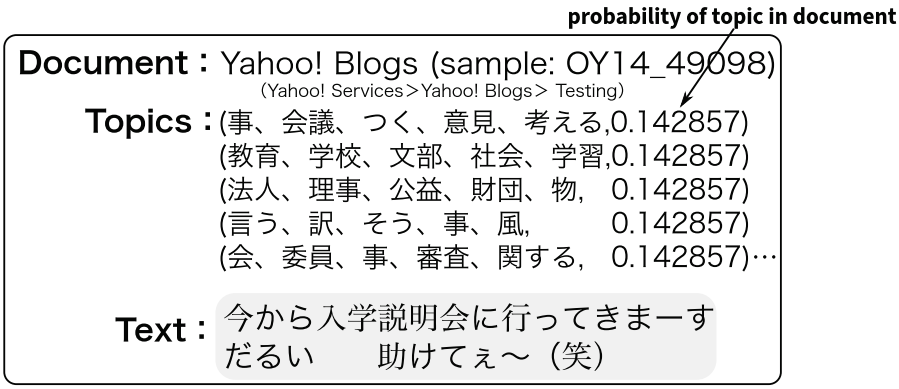


Figure 16: Relationship between documents and topics.

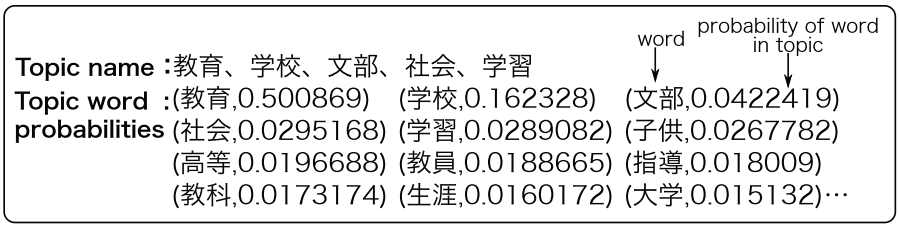


Figure 17: Relationship between topics and words.

to avoid the inclusion of grammatical particles and functional expressions that reflect features relating to register. On the other hand, content words that reflect the theme and the topic of a text should be included. From this perspective, the distribution of content words within a passage of text corresponds to the topic of that text. Correspondingly, the topic features used within the model are nouns (excluding numbers), verbs (excluding dependent verbs), adjectives (excluding dependent adjectives), and adverbs based on the SUWs provided by UniDic. Finally, an LDA model of 1,000 topics (with parameters: $\alpha = 50, \beta = .01$) was built using Yahoo! LDA (Smola & Narayanamurthy, 2010) for the BCCWJ using these features.

5.2 RESULTS

Similar to the methodology employed in the previous chapter, the topic proportions within a particular media were examined by using the topic and word proportions generated by the LDA model for every BCCWJ sample. Table 17 lists the top five topics obtained per BCCWJ media label. Overall, the topic model is able to differentiate between media that are more similar in style from those that are more similar in terms of their topics. For instance, conducted

analyses revealed just some of the following correlations from Table 15, arranged according to media:

- Books

  First, correlations between the three Books corpora reflect some aspects of the sampling methodology used in their construction. The Library (LB) and Publication (PB) media, though employing different sampling bodies, were highly correlated. However, the Bestseller (OB) media showed markedly weaker correlations with both, indicating that its sampling method has exerted some influence on the selection of topics. Accordingly, while topic correlations for both the Yahoo! Blogs (OY) and Q&A (OC) are weaker than for the Library and Publication media, Bestseller media has slightly stronger correlations.

- Magazines

  This media is more strongly correlated with the Internet media (OY and OC), the Books media (PB and LB), as well as the Newspapers media. Although individual magazines tend to focus on specific topics, when taken together as the Magazines media, they also cover a wide range of general interest topics and news, which results in the considerable degree of topic variation observed for this media.

- Yahoo! Blogs and Q&A

  For these media, the correlations are generally low, apart from stronger mutual correlations and their correlations with the Magazines (PM) media. While the variety of topics within the Yahoo! Blogs and Q&A media is considerable, as evidenced by the fact that together they cover 1,000 topics, their weak correlations with the Books media, which also contains similar numbers of topics, imply that the proportions of these topics are different.

- White papers

  The strongest topic correlations observed are for the Legal documents media, the Minutes of the Diet media, and with the Newspapers (PN) media. These correlations reflect the fact that White papers are also closely related to government and politics, which are also central domains for these other media. Interestingly, even though the Minutes of the Diet media is different in register, it is remarkably similar in topic. This would seem to suggest that the topic model is able to successfully discriminate between topic and register, at least in this particular instance.

- Minutes of the Diet

The strongest correlations were observed with the Legal documents (OL) and White papers (OW) media. In stark contrast to their close similarities as media forms in both containing colloquial communications, the Minutes of the Diet media displays a weak negative correlation with the Yahoo! Q&A and Blogs media. This also represents evidence that the topic model's ability to successfully differentiate between register and topic.

The overall patterns of topic correlations between the various media also relate directly to the results obtained from a hierarchical cluster analysis, as shown in Figure 18. Within the figure, there are two clear clusters: one on the left side that relates to government and public policy (OP, OM, OW, and OL), and one looser clustering on the right side consisting of one statistically significant cluster representing a balanced range of general topics (OC, PB, OY, LB, PM, OB, and PN), together with the Textbook (OT) and Verse (OV) media that capture yet another distinctive range of topics. Table 16 presents a summary of the topic coverage across BCCWJ media labels.

Performed with the `pvclust` *package for* R *(Shimodaira, 2004)*

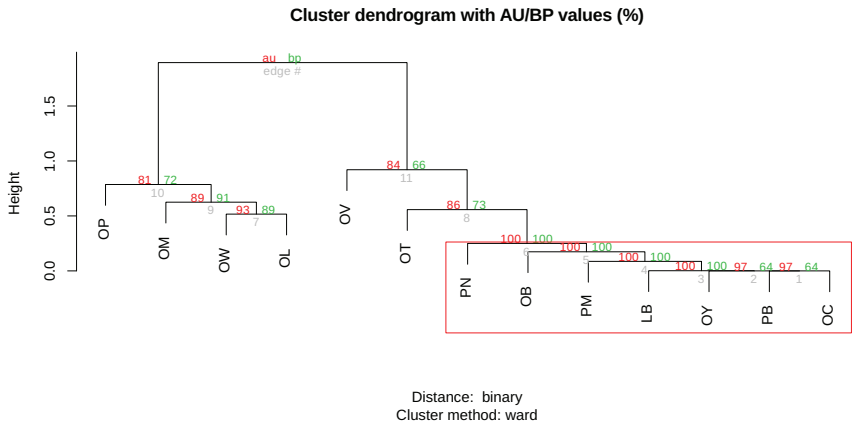


Figure 18: Hierarchical cluster analysis of BCCWJ media by topic using binary distance with Ward's method. The AU/BP values correspond to approximately unbiased and bootstrap probability values, respectively. The enclosed cluster grouping signifies a grouping with AU values greater than 95%.

5.3 DISCUSSION

Results from the correlation and hierarchical clustering analyses would seem to suggest a split between the media that cover many topics and the media that cover only a specific subset. An investigation of frequent topics from each of these groups might provide a clearer sense of these differences. Referring to Table 17, the top topics for the Books media, such as "face, eyes, so, hands" and "say, thing,

Table 15: Spearman correlations between topic probabilities assigned to all samples of BCCWJ media. Strong correlations are emphasized in bold ($> .50$).

		PB	LB	OB	PM	PN	OC	OY	OW	OV	OT	OP	OL
Books	PB												
Books	LB	**0.85**											
Books	OB	**0.50**	**0.66**										
Magazines	PM	**0.58**	**0.56**	0.43									
Newspapers	PN	**0.56**	**0.55**	0.29	**0.59**								
Yahoo! Q&A	OC	0.25	0.18	0.28	**0.54**	0.33							
Yahoo! Blogs	OY	0.18	0.22	0.34	**0.60**	0.37	**0.77**						
White papers	OW	0.33	0.17	-0.09	0.10	0.49	0.06	-0.11					
Verse	OV	0.28	0.42	0.42	0.27	0.16	0.16	0.33	-0.20				
Textbooks	OT	**0.52**	**0.54**	0.35	0.34	0.38	0.19	0.20	0.24	0.29			
Govt. pamphlets	OP	0.14	0.09	-0.02	0.17	0.37	0.10	0.13	0.29	-0.01	0.13		
Legal documents	OL	0.24	0.04	-0.15	-0.08	0.29	0.04	-0.17	**0.61**	-0.22	0.12	0.23	
Minutes of the Diet	OM	0.18	0.10	-0.04	-0.01	0.36	-0.03	-0.15	**0.53**	-0.18	0.08	0.21	**0.56**

Table 16: Topic coverage across BCCWJ media labels sorted by topics found.

Media		Topics found
Books	PB	1,000
Yahoo! Q&A	OC	1,000
Yahoo! Blogs	OY	1,000
Books	LB	999
Magazines	PM	946
Books	OB	892
Newspapers	PN	847
Textbooks	OT	646
White papers	OW	484
Verse	OV	409
Legal documents	OL	329
Govern. pamphlets	OP	264
Minutes of the Diet	OM	224

think, so, time", clearly reflect the general, descriptive style of fiction writing, which accounts for a large portion of this media. These topics are in fact similar to those identified for the Yahoo! Q&A and Blogs media, which also underscores the similarities in their ranges of topics. In contrast, the other media include more specific topics that do not feature as highly in the other media. For example, the Newspaper media includes a sports-related topic represented by "win, times, tournament, first, finals" and several topics relating to politics such as one of foreign diplomacy ("president, USA, relation, government, diplomacy") and another of domestic politics ("premier, LDP, Koizumi, politics, prime minister"). In a similar vein, the Verse media contains topics that are only typically found in verse, such as "night, day, flower, autumn, moon" and "wind, white, center, sky, red". Thus, the other media category clearly contains more selective sets of topics.

5.4 CONCLUSION AND FUTURE WORK

In summary, the LDA topic model is capable of generating topics that, arguably, reflect fairly faithfully the real topics and themes of their respective media. By using this model, it is possible to obtain three advantages: (1) identification of media that contain a wide variety of topics, (2) identification of media that contain similar or specific/original topics, and (3) a supplement to the classification of media according to register presented in §4, that is more capable of capturing context.

Table 17: Top five topics per BCCWJ media.

Rank	Published books (PB)	Library books (LB)	Bestselling books (OB)	Magazines (PM)	Newspapers (PN)
1	顔, 目, 声, そう, 手	言う, 事, 思う, そう, 時	言う, 事, 思う, そう, 時	人気, スタイル, 感, デザイン, 使う	優勝, 回, 大会, 初, 決勝
	face, eyes, so, hands	say, thing, think, so, time	say, thing, think, so, time	popular, style, feel, design, use	win, times, tournament, first, finals
2	言う, 事, 思う, そう, 時	顔, 目, 声, そう, 手	顔, 目, 声, そう, 手	バッグ, スカート, ニット, パンツ, スタイル	大統領, 米国, 関係, 政府, 外交
	say, thing, think, so, time	face, eyes, so, hands	face, eyes, so, hands	bag, skirt, knit, shorts, style	president, USA, relation, government, diplomacy
3	事, 於く, つく, 因る, 物	言う, 声, 顔, 事, そう	言う, 出る, 男, 電話, 入る	シャツ, ブランド, ジャケット, プリント, カラー	首相, 自民, コイズミ, 政治, 総理
	thing, at, about, due to, thing	say, thing, think, so, time	say, leave, man, telephone, come	shirt, brand, jacket, print, color	premier, LDP, Koizumi, politics, prime minister
4	言う, 於く, 因る, 関係, 意味	言う, 出る, 男, 電話, 入る	言う, 声, 顔, 事, そう	映画, 監督, 作品, 出演, 描く	イラク, テロ, イスラエル, 戦争, イラン
	say, at, due to, relation, meaning	say, leave, man, telephone, come	say, voice, face, thing, so	movie, director, work (film), production, depict	Iraq, terrorism, Israel, war, Iran
5	表示, クリック, 設定, ボタン, 選択	事, 言う, 時, 年, 思う	事, 者, 申す, 物, 今	モデル, ボディー, シート, リットル, エンジン	町, 年, 月, 会, 事
	display, click, setting, button, option	thing, say, time, year, think	thing, person, say, thing, now	model, body, seat, liter, engine	town, year, month, meeting, thing
	White papers (OW)	Verse (OV)	Textbooks (OT)	Government pamphlets (OP)	Legal documents (OL)
1	於く, 為, つく, 行う, 図る	夜, 日, 花, 秋, 月	実験, 調べる, 事, 考える, 分かる	月, 日, 市, 申し込み, センター	条, 項, 規定, 当該, 於く
	at, for, about, perform, plan/aim	night, day, flower, autumn, moon	experiment, examine, thing, think, know	month, day, city, apply, center	clause, paragraph, regulation, concern, at
2	年, パーセント, 増加, 図, 別	風, 白い, 中, 空, 赤い	計算, 数字, 数, 答え, 桁	課, ■■, 月, 平成, 市	条, 項, 規定, 因る, 業務
	year, percent, increase, plan, separate	wind, white, center, sky, red	calculation, numeral, number, answer, line	section, ■■, month, heisei, city	clause, paragraph, regulation, due to, business
3	事業, 整備, 年度, 施設, 実施	姿, 巨大, 物, 光, 今	運動, 力, 速度, 時, 物体	時, 日, 午後, 分, 午前	事業, 事, 指定, 定める, 大臣
	project, maintenance, fiscal year, enforce	figure, huge, thing, light, now	exercise, power, speed, time, body	time, day, years after, minute, morning	project, thing, regulation, decide, cabinet minister
4	年, 国際, 月, 於く, 協力	顔, 目, 声, そう, 手	反応, 液, 化学, 分子, 水溶	区, 月, ファックス, 日, 役所	法, 法律, 改正, 物, 年
	year, international, month, at, cooperation	face, eyes, voice, so, hand	reaction, liquid, chemistry, molecule, water-soluble	precinct, month, fax, day, govern. office	act, law, revision, thing, year
5	率, 上昇, 年, 因る, 事	目, 中, 声, 手, 体	式, 関数, 時, 次, 事	町, 年, 月, 会, 事	執行, 命令, 債権, 裁判, 事
	rate, rise, year, due to, thing	eyes, center, voice, hands, body	formula, function, time, next, thing	town, year, month, meeting, thing	enforcement, order, credit, trial, thing
	Yahoo! Q&A (OC)	Yahoo! Blogs (OY)	Minutes of the Diet (OM)		
1	方, 教える, どう, 分かる, 出る	所, もう, 後, 前, 気	委員, つく, 事, 大臣, 政府		
	manner of, teach, how, know, leave	place, already, after, before, mood	cmt. member, about, thing, cabinet minister, govern.		
2	言う, 今, 人, 時, 知る	言う, 事, 思う, そう, 時	言う, 訳, そう, 事, 風		
	say, now, person, time, know	say, thing, think, so, time	say, reason, so, thing, way		
3	言う, 事, 思う, そう, 時	今日, 明日, 笑い, 頑張る, まあ	案, 国会, 提出, つく, 法案		
	say, thing, think, so, time	today, tomorrow, laugh, persist, you might say	bill, Diet, submit, about, bill		
4	落札, 出品, 評価, 連絡, メール	食べる, 弁当, 美味しい, 御飯, 今日	言う, 事, 思う, 問題, 訳		
	prize winning, exhibit, valuation, contact, mail	eat, lunchbox, delicious, food, today	say, thing, think, problem, reason		
5	所, もう, 後, 前, 気	言う, 今, 人, 時, 知る	言う, 思う, 参考, 事, 風		
	place, already, after, before, mood	say, now, person, time, know	say, think, reference, thing, way		

There are three possible directions that could be profitably pursued in future work: (1) improving the specification of the topic model by incorporating other aspects of the media, such as media labels relating to language-external criteria, and by further investigating the employment of LUWs rather than SUWs, for the language-internal criteria, given that LUWs further constrain the situational contexts in which they occur; (2) improving the topic model by varying the number of topics utilized and by comparing separate topics generated from separate topic models for individual media, and (3) evaluating the utility of these topics within writing assistance systems.

6 READABILITY MODEL

Readability may be simply defined as the relative difficulty of reading or understanding a text. It is an important part of written communication as it is a function of both the writer's and reader's fields of experience.

Constructing a model of readability can potentially aid in realizing a number of important applications for writing assistance systems. For example, a writing assistance system that presents example sentences to the user would be greatly enhanced by being able to show examples that are highly appropriate to the user's proficiency level. Similarly, another interesting direction would be to develop the ability to consider the user's proficiency level and, based on that information alone, to weight different kinds of search queries and suggestions.

6.1 PREVIOUS WORK

Readability research in English has a long and varied history beginning as early as the late 19th century (Benjamin, 2012). In contrast, there is considerably less published research on readability formulas for Japanese, with the exception of the readability formula of Tateisi, Ono, and Yamada (1988), which is essentially a linear regression that takes the average number of characters, average roman, hiragana, katakana, and kanji characters per sentence, and the ratio of commas to periods as its predictors. More recently, readability assessment systems have been developed that incorporate a number of novel features or newer statistical methods: for example, Shibasaki and Hara (2010) uses a multivariable polynomial model, based on surface features, but with predicate counts, while Sato, Matsuyoshi, and Kondoh (2008) proposes another using a bigram language model.

6.2 DATA

It should be noted, however, that there are some issues with applying the Japanese readability formulas and classifiers, just described, to the context of L2 Japanese learners. As a working solution to the absence of an appropriate corpus graded for L2 Japanese learners, this research project utilizes the Textbook media of the BCCWJ version 1.0. The Textbook media consists of samples from nationally approved K-12 textbooks, with every sample containing information on its

See §3.2 for more information about the BCCWJ

Table 18: Textbook media composition by grades.

School	Grade	Sentences	Paragraphs	Samples
Elementary	1	363	121	9
	2	505	173	10
	3	1,492	460	19
	4	1,418	414	15
	5	2,228	552	20
	6	3,340	888	21
Middle	7	1,570	523	14
	8	5,172	1,149	29
	9	4,438	1,056	24
High	10	32,926	6,744	251
	TOTAL	53,513	9,750	412

grade level, which is used in the present research as an approximate proxy for readability levels.

As can be seen in Table 18, there were imbalances in the amounts of Textbook media texts across the grades. In order to redress this situation somewhat, grade levels were grouped under three school levels: elementary (grades 1-6; coded as E), middle (grades 7-9; coded as M), and high (grade 10-12; coded as H) school. Moreover, as Figure 19 shows, there were also imbalances in the distributions of subjects across the elementary, middle, and high school levels.

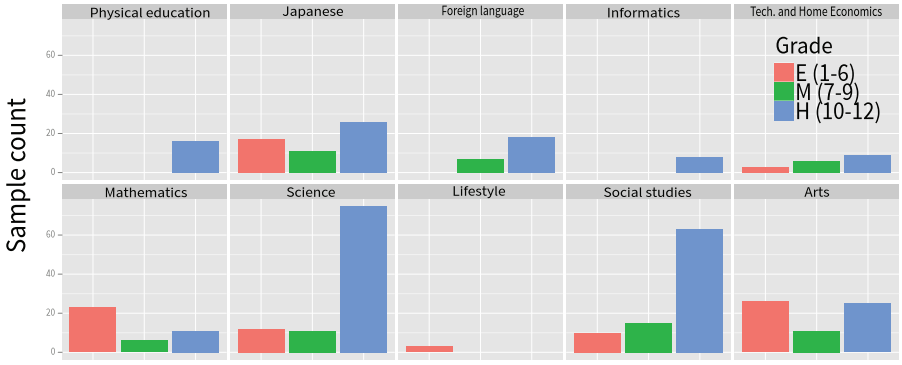


Figure 19: Subject distribution of the BCCWJ Textbook media by grade.

In order to tune and evaluate models of readability, the data was split into training and test sets according to a 3:1 ratio. Additionally, a stratified random split strategy based on school level was employed to maintain similar school level distributions for both the training and test sets.

→ *Table 19*

Table 19: Textbook sub-corpus training and test sets.

	Sentences	Paragraphs	Samples	% of total
Training set	40,136	7,314	311	75
Testing set	13,377	2,436	101	25
TOTAL	53,513	9,750	412	100

6.3 PREDICTORS OF READABILITY

In contrast with previous studies that have used more traditional surface features, the present research includes several vocabulary and syntactic-level features that may be more applicable to L2 learners (Heilman, Collins-Thompson, Callan, & Eskenazi, 2007). Overall, 14 different linguistic features were selected for their potential impact on readability, which can be grouped according to their complexity and linguistic level:

- Surface features: → *Figure 20*

  In addition to the number of characters in a sentence, tokens (SUWs) and chunks are counted from the morphological and dependency structures of the sentences.

- Syntactic features: → *Figure 21*

  Based on CaboCha, calculate average chunk depths and average distances between chunks. While not strictly speaking a syntactic feature, numbers of predicates per sentence were also calculated following the definition in Shibasaki and Hara (2010).

- Writing system features:

  As described in §1.1, the Japanese writing system employs several different scripts in writing: rōmaji, katakana, hiragana, kanji, and symbols. Accordingly, ratios of each word class were recorded at the sentence, paragraph, and source levels.

- Vocabulary features:

  In contrast to the other features that focus on surface or structural features of language, two vocabulary level features were also included. The first is the average (pre-2010) Japanese Language Proficiency Test (JLPT) word level[1], which is consistent with the goal of assessing L2 readability. The second is the average log-scaled probability of tokens (matched on word

1 A JLPT vocabulary list is available from http://www7a.biglobe.ne.jp/nifongo/data/noryoku.html.

class and lemma) found in the BCCWJ Library Books (LB) media, which is used as a measure of the general vocabulary usage in the context of printed materials for adult native speakers—contrasting with the vocabulary usages encountered within the Textbook media, which is related to adolescent readers.

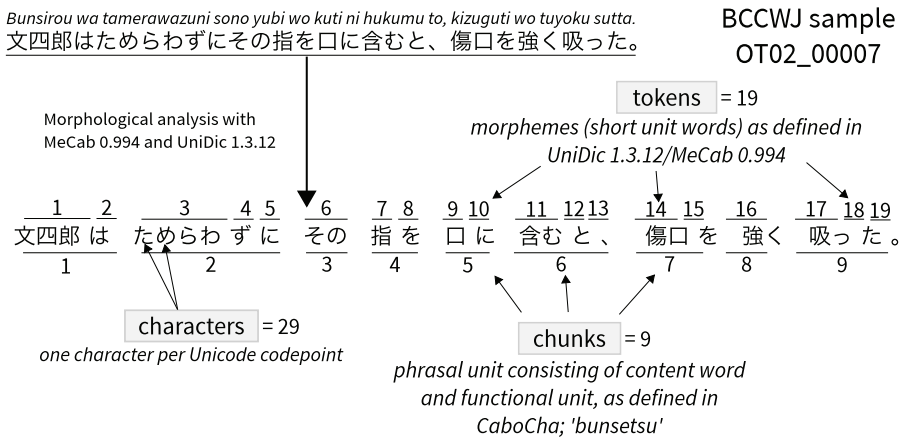


Figure 20: Extraction of surface readability features.

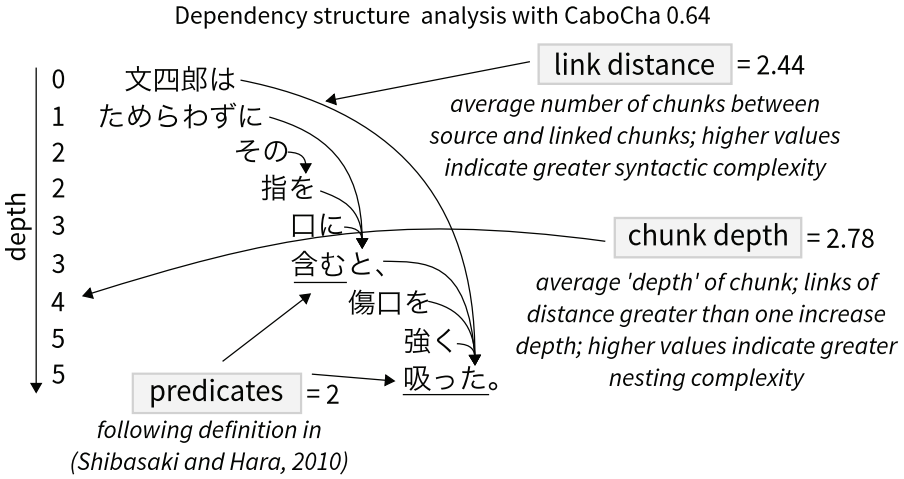


Figure 21: Extraction of syntactic readability features.

In order to identify the correlates of readability at the sentence, paragraph, and sample levels, Pearson correlation coefficients were calculated for each feature with grade level. Hiragana and kanji consistently showed the highest correlations to grade level at all levels of analysis. That is to be expected, as the nationally accredited textbooks that form the Textbook corpus conform to stringent guidelines on which kanji are to be used in which grades. When

→ *Table 20*

Table 20: Pearson correlations for the linguistic features with Textbook grade level for sentences, paragraphs, and samples (top 3 in bold).

Level	Factor	Sentence	Paragraph	Sample
Writing system	Hiragana	**-0.334**	**-0.470**	**-0.720**
	Katakana	0.091	0.142	0.168
	Kanji	**0.331**	**0.443**	**0.730**
	Rōmaji	0.067	0.145	0.153
	Symbols	-0.071	-0.056	-0.373
	Commas	-0.047	-0.114	-0.028
Surface	Characters	0.206	0.147	0.334
	Tokens	**0.209**	0.303	0.589
	Chunks	0.201	0.290	0.597
Syntactic	Link distance	0.174	0.302	**0.631**
	Chunk depth	0.182	**0.311**	0.606
	Predicates	0.126	0.190	0.544
Vocabulary	JLPT word level	-0.206	-0.169	-0.283
	BCCWJ-LB word level	-0.154	-0.185	-0.301

comparing between levels and groups, hiragana and kanji from the writing system group, and the surface and syntactic groups in general, correlate higher as the level of analysis increased from the sentence to the sample, while the vocabulary group features remain mostly unaffected. Finally, the group of surface features (characters, tokens, and chunks), as well as link distance and chunk depth from the syntactic group were all found to be highly correlated ($r > 0.9$, $p < 0.001$) at all levels of analysis (sentence, paragraph, and sample).

6.4 MODELS

After examining the correlations for individual predictors with grade level, the research project next turns to investigate which combinations of predictors would be the most effective in predicting overall readability classification performance. Expanding on the modeling conducted in Hodošček et al. (2012), several supervised learning and statistical classification algorithms were selected for variety and performance across a range of data sets to predict the three readability classes (E, M, and H) using the 14 predictors:

- SVM

  A Support Vector Machine using a Gaussian radial basis kernel function. Contained in the `kernlab` package for `R` (Karatzoglou, Smola, Hornik, & Zeileis, 2004).

- C5.0

  According to the Kuhn, Weston, Coulter, and Quinlan (2013), "C5.0 is an evolution of the C4.5 model described in Quinlan (1993) containing several new features, such as winnowing for feature selection, boosting as well as using a cost function for building the model. C5.0 can create both tree- and rule-based models." Contained in the `C50` package for `R`.

- Random forest

  Classification and regression based on a forest of trees using random inputs, as described in Breiman (2001). Contained in the `randomForest` package for `R` (Liaw & Wiener, 2002).

- Neural net

  Single hidden-layer neural networks with possible skip-layer connections. Contained in the `nnet` package for `R` (Venables & Ripley, 2002).

- Lasso and elastic-net regularized generalized linear model

  A multinomial regression model using the `glmnet` package for `R` (Friedman, Hastie, & Tibshirani, 2010).

Depending on the machine learning model, feature selection can greatly impact model performance, such as when there is colinearity between features. For this reason, most of the selected models incorporate some form of automatic feature selection[2].

Training and validation of the models was carried out using the `caret` package for `R`, which provides functions for data splitting, pre-processing, feature selection, model tuning using resampling, and variable importance estimation, among others (Kuhn, 2008). The algorithm used for the training and validation of the models, presented in Algorithm 1, is essentially that of Kuhn (2008), and is reproduced here for the purpose of exposition.

2 The variety of neural networks used in this thesis being the exception.

```
1  Define sets of model parameter values to evaluate
2  for each parameter set do
3  |  for each resampling iteration do
4  |  |  Hold—out specific samples
5  |  |  Fit the model on the remainder
6  |  |  Predict the hold—out samples
7  |  end
8  |  Calculate the average performance across hold—out predictions
9  end
10 Determine the optimal parameter set
11 Fit the final model to all the training data using the optimal parameter set
```

Algorithm 1: Model training and validation.

Parameter tuning was carried out using 10-fold cross-validation with 3 repetitions (lines 3-7). The accuracy performance measure was used (line 10) to determine the best model, which was then fit on the held-out test data (line 11). This procedure was repeated for every type of model on the three granularity levels: sources, paragraphs, and sentences. Thus, the total number of final models produced was 15.

The tuning search grid specification for each model corresponding to lines 1-9 is included in Appendix §10.1.1

6.5 RESULTS

The results of the models are summarized in Table 21. As the classification is three-way in nature, the results are calculated by comparing each class factor level to the other two levels. Accuracy rates for the models are computed along with 95% confidence intervals. The Kappa statistic is used to measure the amount of agreement between the predicted and reference classes. p-values from McNemar's test are used to test for the significance of this agreement. One-sided tests were also conducted to test whether the accuracy was better than for the "no information rate," which is set to the largest class percentage in the data (high school level).

Analysis of model differences based on accuracy alone revealed no significant differences between models at the sources and paragraphs levels, while significant differences were found for almost all models at the sentences level.

Pairwise differences between models are presented in Appendix §10.1.2

A more detailed view of classifications is offered in Table 22 using confusion matrices that tabulate the predicted and reference classes for each model. When comparing the grade levels, it is clear that the middle school level (M) consistently had more misclassifications compared to the other school levels, where, for example, the sentence model only yielded a two-way classification, as either belonging to the elementary school or high school levels.

The analysis of confusion matrices across the five models and three levels revealed that the glmnet model, while exhibiting similar classification to the other models, also exhibited the highest count of undesirable classifications, which can be attributable to the model's low performance on the middle school class.

In summary of the prediction performance, the confusion matrices (Table 22) and performance statistics (Table 21) for each model show a general increase in the number of misclassifications from samples to sentences, and, in particular, at the middle school level.

Detailed results of model statistics by class is presented in Appendix Table 35

Finally, as a sub-goal of the research project is to identify the relative importance of the selected predictors, Table 23 shows the variable importance data for the SVM model, while Table 24 shows another for the C5.0 model. The analysis here is limited to the sentences level, as the performance of predictors at that level has the most applications for an example sentence display feature.

6.6 DISCUSSION

Examinations for the importance of the predictor variables for the SVM and C5.0 models reveals that features that do not necessarily appear in most sentences, such as rōmaji, katakana, symbols, and commas, are less important predictors for the sentence model. The two most important features were identified as hiragana and kanji. Other important features were chunk depth and link distance from the syntactic feature group, and tokens, length, and chunks from the surface feature group. On the other hand, the results for the BCCWJ and JLPT level vocabulary-based features indicate that they are less relevant to classification performance compared to the other feature type groups. This also highlights the inappropriateness of the JLPT level feature within an L1 data context. Finally, these data suggest that for sentence-level predictors to be useful, they should be sufficiently general to be of relevance for most sentences or, at the very least, are useful for discrimination where other more general features fail.

In cases where some predictors are highly correlated with some other predictors (such as the general inverse relationship between hiragana and kanji in a sentence), it may be detrimental to classifier performance to include all such predictors within the model. Given, however, that an important sub-goal of the present research project is to try and identify which predictors are the most predictive of readability within actually-implemented classification models, highly-correlated predictors were not excluded from the conducted analyses. On the other hand, it is also relevant to note that some of the classification models used here are not greatly influenced by the presence of such highly-correlated predictors, so decisions about whether or not to exclude can be rather moot.

Table 21: Model statistics for each level.

Source					
	SVM	C5.0	Random Forest	Neural Net	glmnet
Accuracy	0.891	0.901	0.921	0.851	0.881
95% CI	(0.813, 0.944)	(0.825, 0.951)	(0.85, 0.965)	(0.767, 0.914)	(0.802, 0.937)
No Information Rate	0.614	0.614	0.614	0.614	0.614
p [Acc > NIR]	4.7e-10	8.83e-11	2.25e-12	1.53e-07	2.27e-09
Kappa	0.802	0.813	0.852	0.722	0.766
Mcnemar's Test p	0.543	0.062	0.112	0.228	0.0186
Paragraph					
	SVM	C5.0	Random Forest	Neural Net	glmnet
Accuracy	0.736	0.724	0.735	0.703	0.706
95% CI	(0.72, 0.752)	(0.708, 0.74)	(0.719, 0.75)	(0.686, 0.719)	(0.69, 0.722)
No Information Rate	0.558	0.558	0.558	0.558	0.558
p [Acc > NIR]	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16
Kappa	0.525	0.509	0.528	0.471	0.456
Mcnemar's Test p	<2e-16	<2e-16	<2e-16	<2e-16	<2e-16
Sentence					
	SVM	C5.0	Random Forest	Neural Net	glmnet
Accuracy	0.678	0.682	0.692	0.67	0.657
95% CI	(0.67, 0.686)	(0.674, 0.69)	(0.684, 0.699)	(0.662, 0.678)	(0.649, 0.665)
No Information Rate	0.616	0.616	0.616	0.616	0.616
p [Acc > NIR]	<2e-16	<2e-16	<2e-16	<2e-16	< 2.2e-16
Kappa	0.291	0.343	0.354	0.296	0.229
Mcnemar's Test p	<2e-16	<2e-16	<2e-16	<2e-16	< 2.2e-16

Table 22: Model classification confusion matrices for E, M, and H on the sources, paragraphs, and sentences held-out test sets. Rows represent predicted classes while columns represent the true classes. Best prediction numbers across models are marked with bold weight, while the highest undesirable prediction (misclassifying E as H or H as E) numbers are underlined.

		Reference														
		Source														
		SVM			C5.0			Random Forest			Neural Net			glmnet		
		E	M	H	E	M	H	E	M	H	E	M	H	E	M	H
Prediction	E	**22**	1	2	**22**	2	2	**22**	1	2	**22**	3	2	**22**	2	1
	M	1	**11**	3	1	9	0	1	**11**	0	0	7	3	0	6	0
	H	0	4	57	0	5	60	0	4	60	1	6	57	1	8	**61**
		Paragraph														
		SVM			C5.0			Random Forest			Neural Net			glmnet		
		E	M	H	E	M	H	E	M	H	E	M	H	E	M	H
	E	471	117	60	**481**	135	96	453	111	70	467	129	91	466	122	85
	M	60	218	92	68	207	92	89	**248**	98	76	183	122	32	115	49
	H	121	347	1534	103	340	1498	110	323	1518	109	370	1473	154	445	**1552**
		Sentence														
		SVM			C5.0			Random Forest			Neural Net			glmnet		
		E	M	H	E	M	H	E	M	H	E	M	H	E	M	H
	E	1114	330	368	1216	416	460	1228	353	419	**1246**	479	538	1009	367	454
	M	69	176	91	203	474	349	163	**475**	275	79	136	126	1	0	3
	H	1153	2289	7772	917	1905	7422	945	1967	7537	1011	2180	7567	1326	2428	**7774**

Table 23: The results of variable importance in the tuned SVM model based on ROC curve analysis. Variables are sorted by average importance across the classes.

Variable	E	M	H
Hiragana	100.0	72.4	100.0
Kanji	97.1	70.1	97.1
Tokens	61.6	40.8	61.6
Length	59.9	40.6	59.9
Chunks	57.0	40.1	57.0
Chunk depth	52.0	39.8	52.0
Link dist	51.0	39.3	51.0
JLPT level	38.9	27.5	38.9
Katakana	36.8	25.9	36.8
Rōmaji	35.5	24.8	35.5
Predicates	35.1	31.4	35.1
BCCWJ level	24.6	24.6	19.2
Symbols	24.5	20.5	24.5
Commas	0.0	11.0	11.0

Table 24: The results of variable importance in the tuned C5.0 model based on the percentage of training set data that fall into all the terminal nodes after the split.

Variable	Overall
Predicates	100.0
Commas	100.0
Chunk depth	100.0
Chunks	100.0
JLPT level	100.0
Length	100.0
Link dist	100.0
Tokens	100.0
Hiragana	100.0
Kanji	100.0
Rōmaji	98.8
Symbols	96.3
BCCWJ level	89.0
Katakana	0.0

As differences in terms of model accuracy were not significant for the source and the paragraph levels, in considering which model might be more preferable over the other, it is important to take into account criteria such as the ease of interpretability. In terms of that criteria, the glmnet model would seem preferable. However, closer inspection of the model's classification profile across classes reveals a high rate of misclassifications for the elementary school level with the high school level and the high school level with the elementary school level. These are naturally undesirable properties, as applying such a model to the selection of example texts to users of writing assistance systems would lead to the systems offering sentences that are less readable to beginner-level L2 learners. Based on this criteria, then, any of the following three models would be more preferable: random forest, SVM, or C5.0.

6.7 CONCLUSIONS AND FUTURE WORK

Having investigated which linguistic features correlate best with Japanese K-12 Textbooks' grade levels, the research project next constructed classification models using a combination of the features as predictors, and measured the effectiveness of each predictor within the model at each level of analysis (sentence, paragraph, and sample). The preliminary findings suggest that for certain predictors to be effective, they should be applied to the paragraph or higher levels. However, the results also indicate that the relative importance of some other predictors, such as JLPT word level, is less influenced by the level of analysis.

While the present models and predictors have been shown to be effective at the sample level, the findings also point to the value of future work that would seek to develop new predictors that are also effective at the sentence level. Moreover, a clear prerequisite for the development of an L2 readability measure is to construct a Japanese language corpus that is graded in terms of its readability for L2 learners, as that would be fundamental to the construction of models that are capable of predicting the readability of texts for L2 learners. In such models, the L1 of the target learner might also play an important role, because, for example, learners from China may experience different problems in understanding sentences that have a high kanji ratio compared to learners from languages that do not use Chinese characters.

Finally, other future work could attempt to vary ratios between the training and test data sets to investigate the impact on performance, as well as investigating different remedies for the class imbalances between the school levels within the data, perhaps by using down-sampling techniques or setting prior probabilities. As the primary goal of building the readability model is to utilize it in classifying sentences from the BCCWJ, STJC, and Wikipedia, which differ

from the Textbook media in both topic and register, particular care should be taken to not over-fit the model.

Part III

WRITING ASSISTANCE SYSTEMS

7

NATSUME

Natsume is an online writing assistance system that began operating in 2009 (Hodošček, 2013; Abekawa et al., 2011a). The focus during the initial development of Natsume was to allow users not only to search for collocations, but also to provide several ways for the learner to obtain some degree of confidence concerning the correct usages of the collocation. For example, search results would provide not only raw collocation information, but would also allow the user to look for a selection of similar collocations from which to make a final selection. Similarly, by comparing collocational tendencies across different genres, the user would be able to make more informed decisions concerning the suitability of selected collocation according to their specific writing goals.

Thus, Natsume focuses on assisting L2 learners of Japanese in writing technical Japanese. For example, writing reports or papers at universities can be hard if the students cannot differentiate between candidate words or expressions in terms of whether they correspond to spoken or written Japanese. As a study and writing aid, the use of conventional (non-corpus-based) electronic dictionaries is prevalent among L2 Japanese learners. However, these dictionaries seldom contain information concerning a word's usage with respect to written and spoken language. Natsume, by virtue of having access to corpora representing various registers, contains information that can be used to determine if a word is more appropriate for spoken Japanese or for written Japanese.

When writing in a second language, it is often the case that one knows the meaning of a noun or verb, but does not know what verb goes together with what noun. Conventional dictionaries often contain only a limited amount of information on the frequently co-occurring patterns of words. These frequently co-occurring patterns of words—*collocations*—are important because they offer more contextual information about a word than available from conventional dictionaries. Moreover, knowledge of the collocations has been identified as an important factor in achieving high second language proficiencies (Pawley & Syder, 1983).

See §1.2
See §1.4
See §1.3

With Natsume, users can use the system to find the collocations of a word, compare the observed frequencies for various genres, and check correct uses by looking at example sentences. These tasks are consistent with the philosophy of data-driven language learning (cf. §1.4), where providing users with access to authentic information is essential for them to be able to make decisions about word choices and regarding their writing. Currently, Natsume is targeted

→ *Figure 22*

primarily at intermediate to advanced learners of Japanese, although it can be used by Japanese native speakers as well.

The opening interface to Natsume includes a search box and a number of selectable options relating to collocational patterns and sorting. After choosing an appropriate word to search for and clicking the search button, the user is presented with the collocation view, from which the user can access additional information from both genre comparison and example sentence views:

1. Collocation view: This displays collocates of the searched keyword. → *Figure 23*
2. Genre comparison view: This displays the distribution of genre frequencies for a given collocation, in revealing use trends for the collocation across different genres. → *Figure 24*
3. Example sentence view: This displays authentic examples so the user can see how the collocation is actually used in context. → *Figure 25*

7.1 COLLOCATION SEARCH INTERFACE

Upon loading the Natsume webpage in a browser, users must select the particular collocation pattern they wish to search for and a matching noun, verb or adjective into the search box to commence a search. The search will present a table containing 8 columns representing different case particles. Inside every column are relevant collocates of the search keyword, which, by default, are sorted by frequency. It is possible to obtain different sets of collocations by changing the sorting of results according to various criteria (frequency, Dice's coefficient, t score, Jaccard similarity coefficient, Log-likelihood ratio, χ^2 coefficient, and Mutual Information score). Color bars to the right of each collocate indicate either the relative frequency or the association measure score of the collocation in all corpora. Additionally, users can search for and compare two or more similar patterns at the same time to help decide which one is more suited for them. By using this feature, users can also resort on any input word, which, in turn, makes it easy to see at a glance which words collocate with which input words. The example in Figure 23 includes such a three-way comparison between the semantically similar verbs やる /yaru/ "do (colloquial)" , する /suru/ "do", and 行う /okonau/ "perform, do".

→ *Figure 23*

The formulas for these measures are given in §10.4

If the user is interested in seeing more information on a particular collocation triplet, they can click on a collocate within the table in order to load the genre comparison view below the collocation view. It is also possible to determine the behavior associated with clicking, from the following options:

- Particle/conjugation expansion: Can be used to compare among different grammatical uses of collocates.

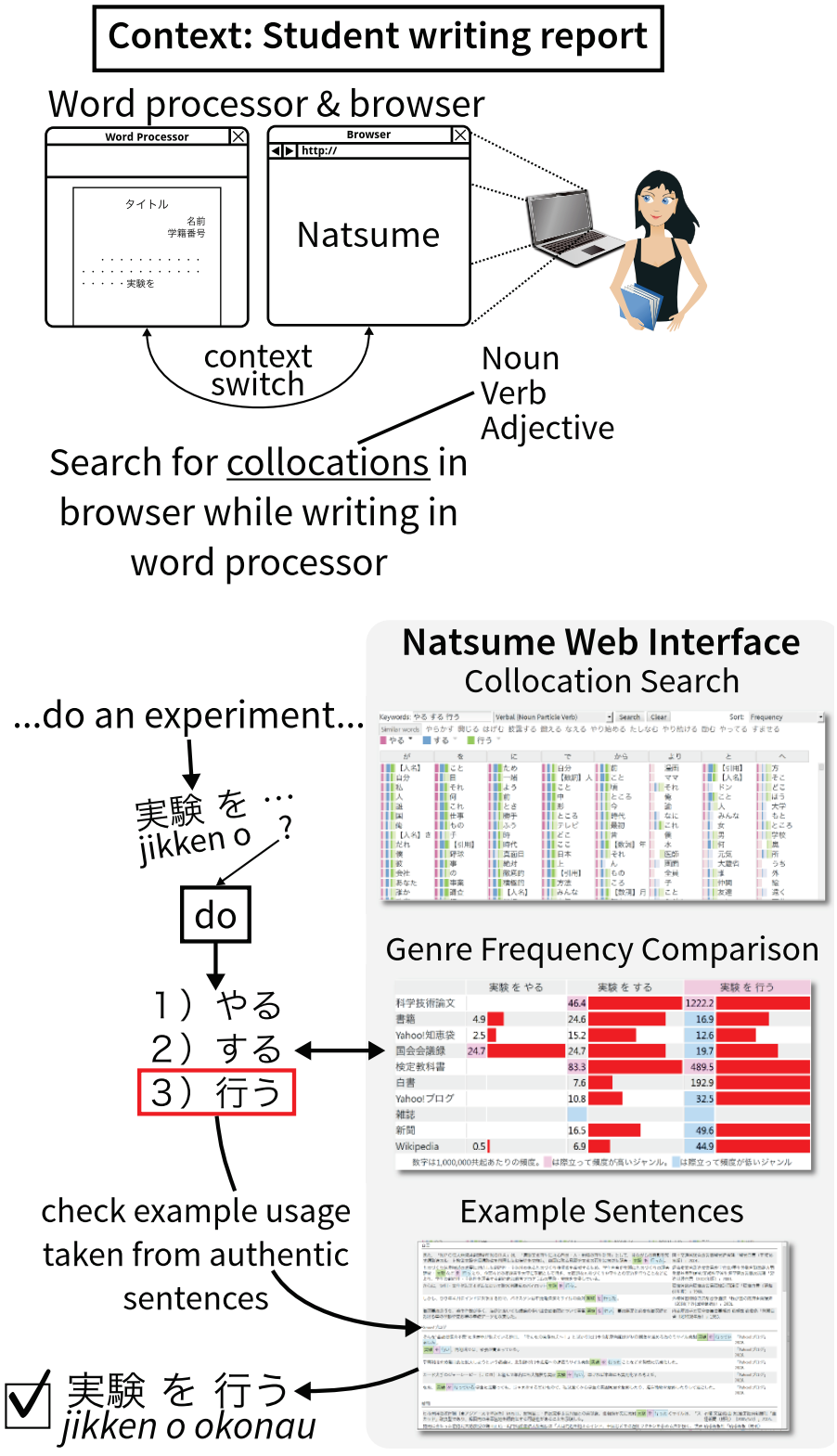


Figure 22: Usage example where student selects the appropriate way of writing “to do an experiment” in a report.

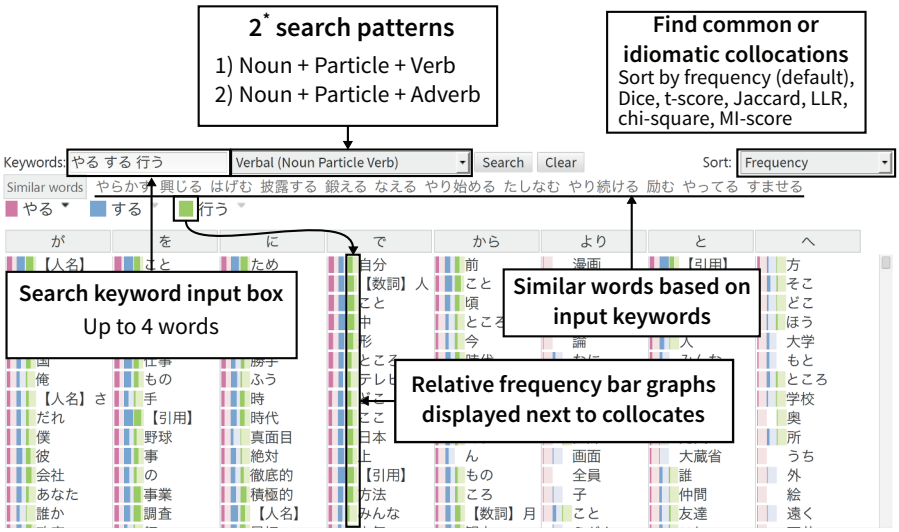


Figure 23: Natsume collocation search interface.

- Synonym expansion: Can be used to automatically compare among similar collocates. Synonym expansion is provided by `getassoc` (Abekawa et al., 2010).
- No expansion (default): The standard view, only provides genre information for the selected collocation.
- Click expansion: Can be used to manually compare genre information for collocates.

7.2 GENRE COMPARISON INTERFACE

With the genre comparison view, users can visually compare a collocation's usage across different genres. The frequency numbers visible within the genre comparison interface are the relative frequencies of occurrence of a particular collocation per 100,000 collocations. This calculation is carried out to help make frequencies more comparable, even when corpus sizes differ, as is the case here. Moreover, Natsume uses data from the χ^2 tests to color-code genres as pink if the frequency of occurrence is significantly larger than the average across all genres, and blue if the frequency is significantly lower (using a significance threshold of 0.05). Thus, for genres that are not color coded, the frequencies are not significantly different from the mean. When even more information is desired, the user can bring up the example sentence view which shows example sentences from the selected corpora. The primary use of this interface is to utilize the genre comparison feature to help in deciding whether one's word choice is indeed appropriate for one's chosen register.

This method is more formally defined in §8.2.1

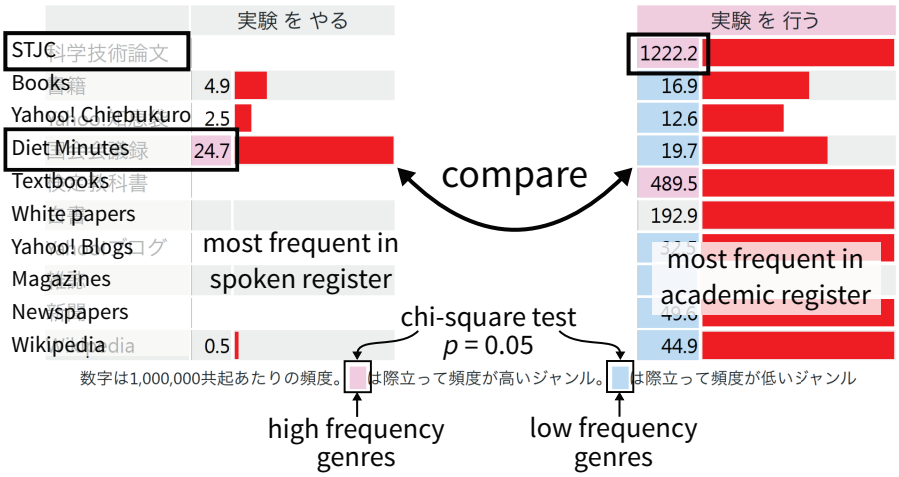


Figure 24: Comparing the collocations of "jikken" with "yaru", "suru", and "okonau" across genres.

7.3 EXAMPLE SENTENCE INTERFACE

The example sentence interface, which is triggered as a pop-up by the user clicking on a collocate in the genre comparison interface, is shown in Figure 25. Examples are grouped by genre, with up to 6 randomly-chosen sentences displayed per genre. Every part of the collocation is highlighted, including verb suffixes, to better convey the sense of the complete meaning unit. To the right of every sentence, the sentence's publication source appears, which includes the title of the source, its author, and publication year. The example sentence view can also be used when the user desires more contextual information concerning a collocation which is not already provided by one of the other views.

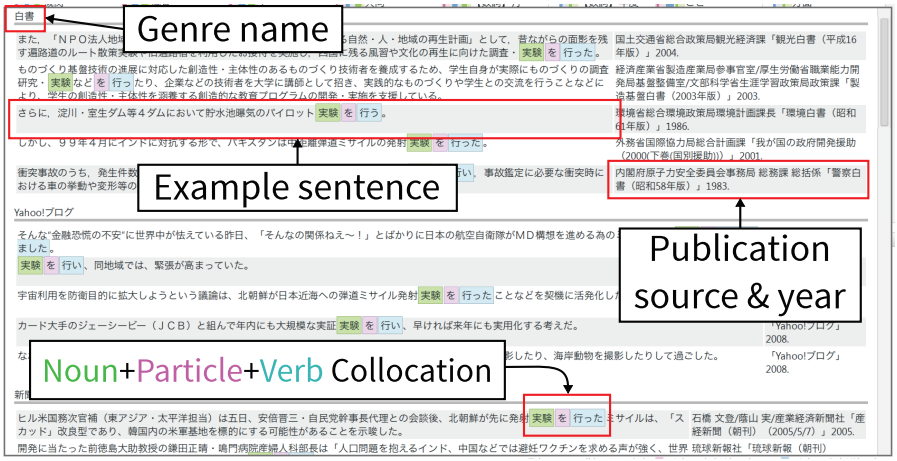


Figure 25: Example sentences for the collocate "jikken o okonau".

7.4 USAGE SUMMARY

The user can judge if a collocation is suitable for one's writing context by comparing its frequency across genres, the differences with similar collocations, and actual usages, as presented within the example sentences sampled from the different corpora. The defining feature of Natsume is the ability to differentiate between expressions that are suitable for writing in academic contexts and those that are not. Consider the following example from the STJC corpus:

- 代謝により発生した二酸化炭素は水に溶解し、・・・

  /Taisha ni yori hasseishita nisankatanso wa mizu ni yōkaishi, …/

  "The CO2 produced through the metabolism dissolves in water, …"[1]

Comparing the expression /nisankatanso ga mizu ni yōkaisuru/ "CO2 dissolves in water", taken from the example below, with the expression /satō ga mizu ni tokeru/ "sugar dissolves in water", one can say that native Japanese speakers with experience in writing scientific and technical Japanese can judge the former as written in an academic, technical style, while the latter as being of a more informal, spoken variety. For learners without such native language intuitions, but needing to arrive at the same conclusion, Natsume provides a data-driven way of helping them to move a step closer towards gaining such intuition.

7.5 NATSUME EVALUATION

7.5.1 *Sentence Rewriting Task*

The first evaluation was conducted to measure the improvements in terms of register usages for collocations within a sentence rewriting task. The experiment was limited to eliciting rewrites from participants, mainly, in areas where using Natsume could aid the user in selecting an appropriate collocation and did not tackle any issues relating to how use of Natsume could potentially improve writing skills in general.

7.5.1.1 *Experimental Procedure*

The experimental procedure, as shown in Figure 26, was used to test for the effects of using Natsume to assist with sentence rewrites. The experiment was conducted in January 2011 with 40 undergraduate participants enrolled in a

1 Excerpt from Terajima, R, Shimada, S., Oyama, T., & Kawasaki, S. (2009) "Fundamental Study of Siliceous Biogrout for Eco-Friendly Soil Improvement", Doboku Gakkai Ronbunshū C, Vol. 65 No. 1, p. 120-130.

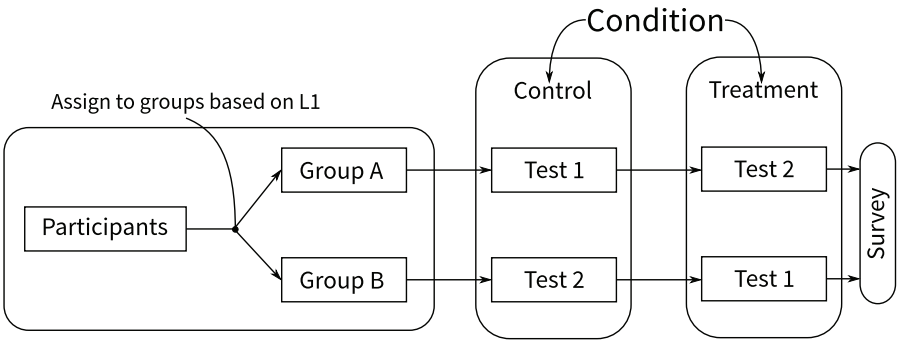


Figure 26: Experimental design for Experiment 1.

4-year university in Tokyo. The Japanese language proficiency level of participants was high enough for the participants to be able to conduct studies in Japanese, corresponding to levels 1 or 2 of the (pre-2010) Japanese Language Proficiency Test (JLPT). The first phase of the experiment involved Group A solving Task 1, and Group B solving Task 2 as the control condition. The second phase of the experiment involved both groups solving the opposite task, as the treatment condition of using Natsume. After completion of both tasks, the participants were asked to fill out a survey concerning their experiences using Natsume.

→ *Figure 26*

More than half of the participants had Mandarin Chinese as their L1 (22), followed by Korean (9), Indonesian (2), Thai (2), Mongolian (2), and one participant for Vietnamese, Khmer, and Bengali, respectively. Participant group selection was carried out in order to roughly balance participant L1s across the groups.

The two tasks (Task 1, Task 2) each consisted of 15 sentences and one short passage of text written in a colloquial style. Care was taken to closely match the levels of difficulty for the two texts based on the numbers of words per each level of the JLPT.

Copies of the tasks with associated instructions are available in Appendix §10.2.1

Several restrictions were imposed on the participants during the experiment:

- Participants were allotted 30-60 minutes to complete each task.
- Use of electronic and online dictionaries[2] was permitted during the tasks, and most participants made use of them.
- Participants were instructed in the use of Natsume for 5 to 10 minutes prior to commencing the treatment task.
- For both tasks, the participants were instructed to rewrite the sentences into an academic style. During the control phase, participants had to rely solely on their intuitions about academic style. During the treatment

2 Dictionaries were restricted to non-scientific and technical Japanese, i.e. general Japanese-Japanese and L2 learner dictionaries.

phase, the participants were instructed to look at the STJC genre frequency when deciding on about the appropriateness of expressions for academic writing.

All the instructions provided to the participants were in Japanese.

7.5.1.2 *Scoring Method*

The same scoring method was used for both tests and conditions.[3] Each sentence involved either one or two collocations, with the total number being 23 for each task. Points were awarded according to the appropriateness of the rewrites as examples of the academic register and according to whether the type of collocation was searchable with Natsume; with maximums of 1 point for unsearchable and 2 points for searchable (i.e. NPV). The number of such expressions was approximately 20 per task. That aspect of the scoring method was included to try and identify cases where Natsume would be less helpful and to try and obtain some sense of what kinds of rephrasing strategies such L2 learners might employ. The types of collocations that were unsearchable with Natsume at the time this study was conducted included, for example, ones involving adverb, adjective, and adverbial adjective words.

Example (1) is a sentence taken from Test A (Appendix 10), and can be roughly translated as "Even if meteorites collide, it is unlikely to affect us in a great way". Example (2) is an actual rewrite taken from the experiment, while Example (3) is a more appropriate rewrite for the target academic register.

(1) 天体の中の 小惑星がぶつかって も、私たちは 大きい 影響をもらわない でしょう。

/tendai no naka no shōwakusē ga butsukatte mo, watashitachi wa ōkī ēkyō o morawanai deshō/

(2) 天体の中の 小惑星がぶつかる が、私たちは 大きい 影響を受けない だろう。

/tendai no naka no shōwakusē ga butsukaru ga, watashitachi wa ōkī ēkyō o ukenai darō/

(3) 天体の 小惑星が衝突して も、我々は 重大／深刻な 影響を受けない。

/tendai no shōwakusē ga shōtotsushite mo, wareware wa jūdai/shinokuku na ēkyō o ukenai/

In Example (2), 影響をもらわない /ēkyō o morawanai/ "to not receive an affect" is rewritten as 影響を受けない /ēkyō o ukenai/ "to not be affected" which

3 For this particular study, scoring was conducted by the authors of Hodošček and Nishina (2011a).

Table 25: The results of the type III ANOVA conducted on score with L1 as a covariate and condition, test, and group as independent variables.

	Sum Sq	Df	F value	Pr(>F)	
(Intercept)	5089.7	1	30.33	5.035×10^{-7}	***
L1	1114.4	2	3.32	.0416	*
Condition	2223.6	1	13.25	.0005	***
Test	14.2	1	0.08	.7719	
Group	147.1	1	0.88	.3522	
Residuals	12418.9	74			

is awarded 2 points, while 小惑星がぶつかる /shōwakusē ga butsukaru/ "the meteorite impacts" and 大きい /ōkī/ "big" are not rewritten and so receive 0 points. Thus, the total points awarded to Example (2) are 2.

Example (3) shows the more appropriate rewrite in this context, where the collocation of 小惑星が衝突して /shōwakusei ga shōtotsushite/ "the meteorite impacts" and one of the word-level rewrites 重大な /jūdaina/ "severe" or 深刻な /shinkokuna/ "serious" are also employed, to receive an additional 2 and 1 points, respectively. Thus, the total number of points awarded to (3) are 5.

7.5.1.3 *Results*

Figure 27 shows boxplots for the scores as a function of group, test, and condition. Given that direct comparisons of the possible interactions between all levels of the three independent variables (condition, test, and group) are not feasible with the experimental design used, a ANOVA was conducted with L1 as the covariate. Prior to conducting the analysis, the L1 factor was recoded to include only three levels: Chinese ($N = 24$), Korean ($N = 6$), and Other ($N = 10$).

The results of the ANOVA, presented in Table 25, indicated a significant effect of condition ($F = 13.25$, $p < .001$) and a significant effect of L1 ($F(2, 1) = 3.32, p < .05$) Post-hoc comparisons using the Tukey HSD test indicated that the mean score for the treatment condition was significantly higher than the mean for the control condition. However, the post-hoc tests for L1 did not detect any statistical differences between the means for Mandarin, Korean, and Other languages (Table 26).

Finally, even though more than half of the participants were Chinese and the average scores for the Chinese and Korean participants were higher than the averages for the other groups, the highest overall scoring participant was from Mongolia (coded as "Other" in the ANOVA analysis).

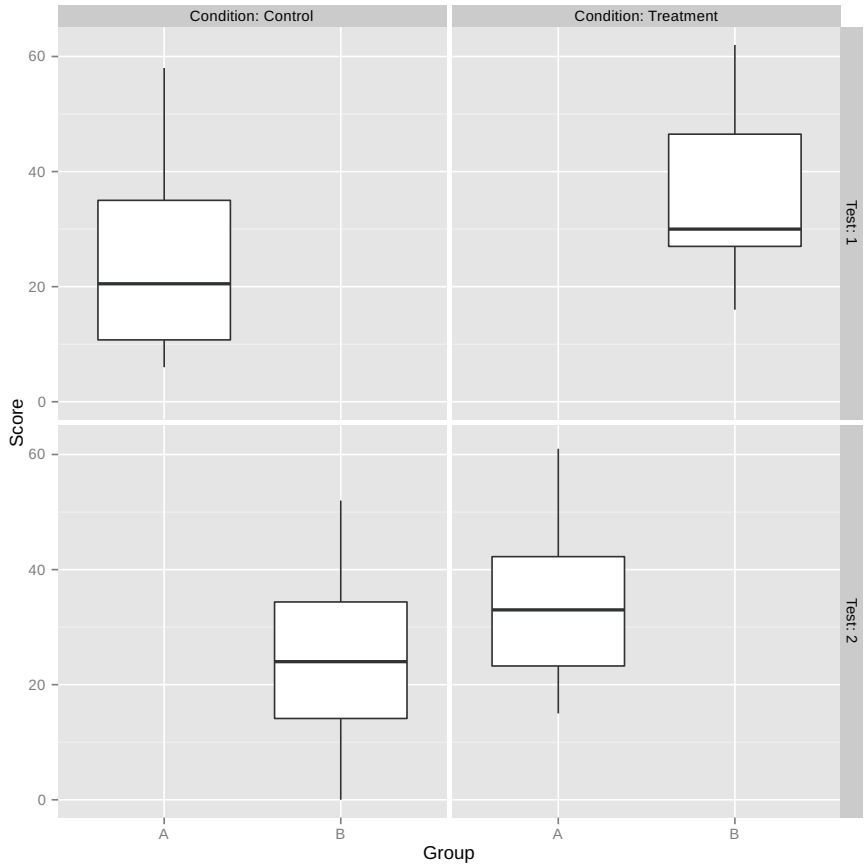


Figure 27: Boxplots for the control and treatment scores as a function of group and test.

Table 26: The results of the post-hoc Tukey HSD tests.

Factor	Comparison	$\mu_1 - \mu_2$	Lwr bound	Upr bound	p_{adj}
L1	Mandarin-Korean	0.39	-8.10	8.88	.9934
	Other-Korean	8.45	-1.35	18.25	.1047
	Other-Mandarin	8.06	-0.42	16.55	.0661
Condition	Treatment-Control	10.51	4.74	16.28	.0005

7.5.1.4 *Discussion*

Overall, scores were found to be higher for the treatment than for the control condition, suggesting that even first-time users could successfully navigate the Natsume interfaces and make appropriate decisions based on the data presented to them.

One can also gain useful insights into the participants' impressions and preconceptions towards the Japanese academic register and about the rewriting strategies that L2 learners may employ by examining the rewriting results from the control task. Although in the minority, there were five participants who scored lower under the treatment condition than for the control tests, and further examination of these rewrite sentences revealed that they were generally higher-scoring participants on the control test, possibly indicating higher levels of language proficiency and deeper knowledge of the academic register. Certainly, the rewrite strategies that they employed were rather advanced, including conjunctive particles, adverbs, adjectives, and written style expressions, as well as various paraphrasing strategies, such as adjusting for abstract/concrete concepts and verb nominalizations.

→ *Table 27*

These points suggest a few possible reasons for the lower scores under the treatment condition of using Natsume. One is that the tested version of the system did not provide sufficient collocational patterns to support the participants' existing rewrite strategies, as shown in Table 27. Another possible reason could be that for such highly-proficient learners, it may not have been particularly time-effective for them to undertake the rewrite task using an unfamiliar system, which did not necessarily provide any better rewrite suggestions than they could already generate themselves based on their own fields of experience, which would already reflect their exposures to academic topics and the Japanese written academic register.

Returning to the larger pattern of results, however, while the findings do indicate that many of the study participants possibly had some awareness of the academic register, which was also indicated by responses to the questionnaire survey, they also highlight a lack of practical knowledge concerning techniques and steps for finding and determining whether collocation candidates are actually appropriate for the academic register. For instance, the main comments regarding the participants' own perceived knowledge of academic writing from the post-experiment survey included the following:

- There are written words that should be used instead of their spoken equivalents.
- Higher usage of kanji is appropriate.
- Use of the /de-aru/ form is preferred.
- There are certain noun phrases, predicates, conjunctive particles, and adjective usages that are unique to the academic register.

Table 27: Examples of rewriting not supported by Natsume.

Type of rewrite	**Before**	**After**
Lower→higher concept	プラスチックなどのゴミ	不燃ゴミ
	/purasuchikku nado no gomi/	/funen gomi/
	“garbage such as plastic”	“incombustible garbage”
Noun phrase→noun	水の深さ	水深
	/mizu no fukasa/	/suishin/
	“depth of water”	“water depth”
Verb nominalization	決まった条件	先決条件
	/kimatta jōken/	/senketsujōken/
	“terms that were decided on”	“predetermined terms”
Adverb	いくつかの	若干
	/ikutsuka no/	/jakkan/
	“some” (hiragana)	“some” (kanji)
Adjective	いろいろな観点	様々な観点
	/iroirona kanten/	/samazamana kanten/
	“various viewpoints”	“various viewpoints”
Noun	大きい違い	大きい差
	/ōkī chigai/	/ōkī sa/
	“large difference”	“large difference”
Compound particle	家族について	家族に関して
	/kazoku ni tsuite/	/kazoku ni kanshite/
	“concerning the family”	“related to the family”
Modality	条件が合うようにしなければならない	条件を満たすべきだ
	/jōken ga au yō ni shinakerebanaranai/	/jōken o mitasu beki da/
	“the terms must be made to fit”	“the terms must be fulfilled”
Conjunctive particle	考え出したので	提出したがゆえに
	/kangaedashita node/	/teishutsushita ga yue ni/
	“Because … thought of …”	“For that reason …”

- Having seen an academic word or phrase before, it is still hard to use it correctly in a sentence.

Thus, in summary of the survey results, while some participants were aware of the differences between non-academic and academic writing styles, they are unable to determine if the particular expressions that they use are appropriate for academic writing. Finally, the significant effect for L1 within the ANOVA analysis would seem to suggest a need for further investigations of L1 effects within each of the coded language groups.

7.5.2 *Report Writing Task*

University-level report assignments require the student to observe and analyze data, form objective conclusions based on that data, as well as to refer to previous work within the problem space (Swales, 1990). In order to measure the effectiveness of Natsume within an authentic writing environment, a second evaluation experiment was conducted in June 2011. The experiment consisted of a report writing task with 20 undergraduate participants of intermediate-advanced Japanese language proficiency from technical and scientific fields attending a 4-year university in Tokyo.

7.5.2.1 *Experimental Design*

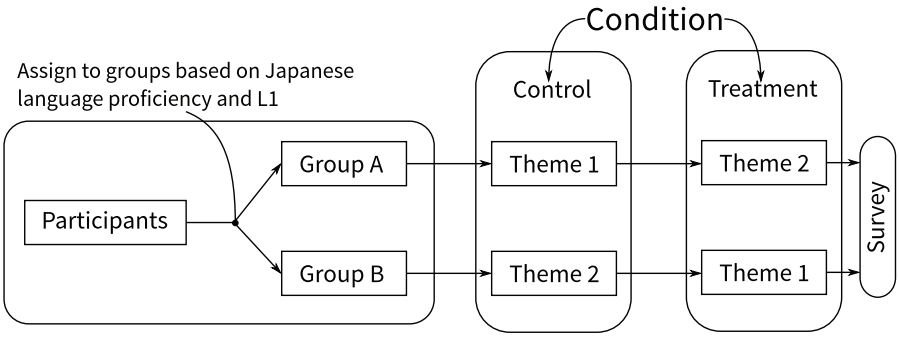


Figure 28: Experimental design for Experiment 2.

The experimental design for Experiment 2 closely follows that of Experiment 1, but rather than having the students rewrite sentences, they were given themes to write reports about. Moreover, as shown in Figure 28, participant group selection was carried out randomly based on the students' language proficiency level[4] and L1. Table 28 presents the results of a paired t-test carried out on language proficiency scores by group, which fails to reject the null hypothesis. In other words, the results supported the appropriateness of the assignments.

4 Based on their language proficiency test scores at the beginning of the semester.

Table 28: The result of the t-test for language proficiency scores between groups A ($\bar{\mu} = 35.10$, $SD = 6.20$, $N = 10$) and B ($\bar{\mu} = 34.8$, $SD = 7.40$, $N = 10$).

t-test type	Test statistic	df	p value
Paired	0.145	19	.886

The following two topics were chosen and assigned to the different groups:

Full introductory text and instructions are available in Appendix §10.3.1

1. Should Japan prohibit animal experiments? (「日本は動物実験を全廃にすべし」)
2. Should English be made an official language of Japan? (「日本は英語を第二公用語にすべし」)

Both topics came with a short introduction setting out the problem by providing relevant basic vocabulary and directions to help the participants argue for their positions.

While writing, the participants had to observe the following guidelines:

- The participants were provided with laptops, wrote in MS Word 2010, and used the Firefox 5 browser (when using Natsume).
- The participants had to write for at least half an hour but no longer than one hour.
- In their reports, they had to decide to either reject or agree with the theme and to back up their decision with reasoning. A short summary of the topic for each theme was provided and is available in Appendix §10.3.1.
- Each report had to be at least 200 characters long.
- Use of the /de-aru/ form was mandated.
- When using Natsume, the participants would refer to the STJC genre frequency when deciding if a collocation was suitable for the report.
- Use of general, non-specialized electronic or online dictionaries was allowed for both conditions (control and treatment).
- Before the experiment, all participants were introduced to Natsume and trained in how to used it for 5-10 minutes.

During the experiment, the amount of time each participant spent on the writing task, as well as the usage log for Natsume during the treatment condition was recorded.

7.5.2.2 *Data Extraction*

Collocation data was extracted from the written essays using Natsume. In order to compare the collocation productions between the two conditions, all extracted

collocations were matched with the Natsume usage logs for each participant, so that only those collocations that contained keywords searched for with Natsume were retained. In total, there were 146 keywords contained in the search logs, 100 of which were unique. Furthermore, the collocations were required to contain the top 8 most frequent keywords. Based on these criteria, 66 collocations were identified for the control condition, while 36 collocations were found for the treatment condition.

7.5.2.3 *Evaluation Procedure*

In this experiment, collocations were evaluated from three viewpoints (register, semantics, and syntax) and two levels (collocation and sentence).

The first level only focuses on the collocation without considering its connection to the context sentence, while the second level rather focuses on how the collocation fits within the sentence. This distinction allows for the following measurements of the collocation productions: if a collocation was correctly copied from Natsume and if a collocation was also then correctly integrated into the sentence.

The first viewpoint is that of register and captures how well the collocation fits into the academic register. For register, this research project only considers the appropriateness of the collocation and not the whole of the sentence. The second is that of semantics and captures the appropriateness of the collocation with respect to either the cohesive meaning of the two collocates or the meaning of the sentence. Finally, the third is syntax and captures the syntactic appropriateness of either the two collocates or the collocates as part of the larger syntactic structure of the whole sentence.

Three Japanese language teachers with experience in writing technical papers were asked to grade the collocations based on five criteria on a 5-point Likert scale. The five criteria are shown below with examples of high and low evaluations from the experiment:

1. **Register**: Is the collocation suitable for the academic register?

   High evaluation: 実験を行う /jikken o okonau/ “to conduct an experiment”

   Low evaluation: 実験をやる /jikken o yaru/ “to do an experiment”

2. **Collocation semantics (C-Semantics)**: Is the collocation semantically correct?

   High evaluation: 薬を飲む /kusuri o nomu/ “drink medicine”

   Low evaluation: 薬をとる /kusuri o toru/ “take medicine”

3. **Collocation syntax (C-Syntax)**: Is the collocation grammatically correct?

Table 29: Inter-annotator agreement, mean, and SD values for the recoded 3-point scale evaluations for each evaluation item.

	Register	S-Semantics	C-Semantics	S-Syntax	C-Syntax	Overall
Agreement (%)	34.2	29.5	56.8	34.9	64.4	44.0
$\bar{\mu}$	0.38	0.27	0.66	0.36	0.68	0.47
SD	0.79	0.84	0.67	0.83	0.67	0.78

High evaluation: 被害にあう /higai ni au/ “to meet with damage”

Low evaluation: 被害をあう /higai o au/ “to meet damage”

4. **Sentence semantics (S-Semantics)**: Is the usage of the collocation semantically correct with respect to the whole meaning of the sentence?

   High evaluation: 科学者が実験に命をかける /kagakusha ga jikken ni inochi o kakeru/ “a scientist risks his life in an experiment”

   Low evaluation: 動物の命をかけて実験をするのは動物にとって残酷すぎる /dōbutsu no inochi o kakete jikken o suru no wa dōbutsu ni totte zankokusugiru/ “to risk an animal’s life for the sake of an experiment is too cruel to the animal”

5. **Sentence syntax (S-Syntax)**: Is the usage of the collocation syntactically correct with respect to the syntactic structure of the sentence?

   High evaluation: 文章を英語で書いた /bunshō o eigo de kaita/ “[subject] wrote text in English”

   Low evaluation: 英語圏を中心とする今の科学界には、多くの文献は英語で書いた /eigoken o chūshin to suru ima no kagakukai ni wa, ōku no bunken wa eigo de kaita/ “The English-centered nature of the present scientific community is the reason that much writing is in English”

Analyses conducted on the 5-point Likert scale evaluations revealed a low level of inter-annotator agreement of 20.3% across all collocations and evaluation items. As actual differences in the magnitude of agreement or disagreement within the evaluation was deemed less important than the general direction of the evaluations, after recoding the scale to 3 points, the agreement level increased to 44% (see Table 29). This recoding of the evaluation scale was done by replacing evaluation scores of 4 and 5 with 1, 1 and 2 with -1, and 3 as 0. These recoded scores correspond to “suitable usage” (1), “neutral” (0), and “inappropriate usage” (-1).

Agreement percentages computed using the `agree` *function in the* `irr` *package for* `R`

The results show lower mean scores and agreement percentages for register and sentence-level semantics and syntax. By using Natsume, one may expect that clear semantic and syntactic violations would be limited by choosing from

Table 30: Pearson correlations between the responses (upper diagonal part contains correlation coefficient estimates, lower diagonal part contains the corresponding p-values).

	Register	S-Syntax	C-Syntax	S-Semantics	C-Semantics
Register		0.29	0.32	0.34	0.39
S-Syntax	< .001		**0.64**	0.57	0.45
C-Syntax	< .001	< .001		0.47	**0.67**
S-Semantics	< .001	< .001	< .001		**0.63**
C-Semantics	< .001	< .001	< .001	< .001	

the collocations provided, assuming that they are sufficiently frequent, which is one possible reason for the higher mean at the collocation level compared to the sentence level. Furthermore, evaluations at the collocation level are inherently more limited than those at the sentence level, as there are fewer factors that would impinge on judgment of appropriateness.

Pearson correlations between the evaluation items for all annotators revealed strong correlations within the two levels (collocation and sentence). Register was found to be the least correlated item, which suggests that it is more independent. This is fairly reasonable, as the spoken collocation 実験をやる /jikken o yaru/ "to do an experiment", is both syntactically and semantically correct, but is not appropriate for the academic register. → *Table 30*

7.5.2.4 *Report Writing Results*

Looking at the mean scores of all evaluation items reveals higher scores in the treatment condition. A Welch t-score test confirmed that the mean score for the treatment was greater than that for the control condition ($N(control) = 66$, $N(treatment) = 36$, $p < 0.05$) with all evaluation items combined. Individual items are presented in Figure 29. → *Figure 29*

Register evaluation scores were lower compared to semantics and syntax items indicating the difficulties of using appropriate collocations for the academic register, compared to the other items. Moreover, the fact that the inter-annotator agreement for register items was low would seem to reflect the difficulties to reliably annotating for this item.

The following three examples of collocation usage were given different evaluation scores.

(4) グローバル化が非常に進んでいる現在、多くの領域での国際的な協力が期待される。

/**gurōbaruka ga** hijō ni **susundeiru** genzai, ōku no ryōiki de no kokusaiteki na kyōryoku ga kitaisareru./

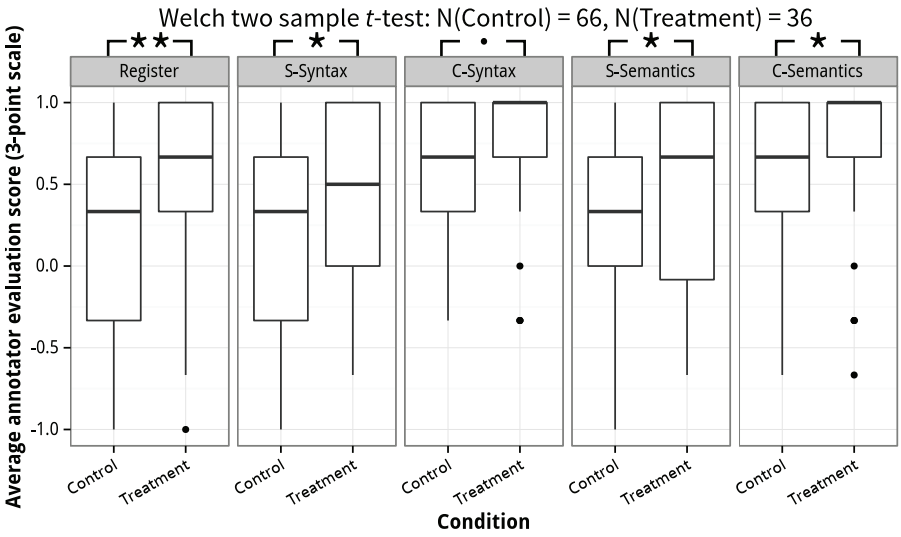


Figure 29: Boxplots for the control and treatment scores for each evaluation item ($. = p < .1, * = p < .05, ** = p < .01$).

> "Presently, as globalization rapidly advances, international cooperation may be expected in many areas"

Example (4) was assigned high evaluation marks from all the annotators and is an example of correct usage in terms of all the evaluation items.

(5) 小学生１年生から英語授業を導入したら、小さい子から英語能力を養うことを通して、日本人全民の英語能力が向上できる。

/shōgakusē ichi-nensē kara ēgo jugyō o dōnyūshitara, chisai ko kara ēgo **nōryoku o yashinau** koto o tōshite, nihonjin zenmin no ēgo **nōryoku ga kōjōdekiru**./

"If English language classes are introduced from the first grade of elementary school, and as English proficiency is fostered from childhood, the English proficiency of all Japanese people will increase."

In example, Example (5), the usage of the first collocation 能力を養う /nōryoku o yashinau/ "to foster proficiency" received high scores, while that of 能力が向上できる /nōryoku ga kōjōdekiru/ "proficiency can improve" was graded lower.

(6) 子供の考えはどこから影響をとるかというと、日常生活なのである。

/kodomo no kangae wa doko kara **ēkyō o toru** ka to iu to, nichijōsēkatsu na no de aru./

"If anything, everyday life influences childrens' thinking"

Finally, Example (6) used the expression 影響をとる /ēkyō o toru/ instead of the correct 影響を受ける /ēkyō o ukeru/ "to be affected by ~", leading to an overall low score for all evaluation criteria.

Another possible way of assessing the effect of the treatment condition on collocation productions is to look at the variety of verbs used for some common nouns contained in both essay types. For instance, the variety of verbs used with the noun 実験 /jikken/ "experiment", was greater in the treatment condition (4 vs. 3). There was also a greater variety of nouns used with the verb 進んでいる /susundeiru/ "keep going" in the treatment condition compared to the control condition (3 vs. 1). While it is hard to draw definitive conclusions from this small sample, given the flexible nature of searches within Natsume, where users can search for any word, and sort based on importance, it is possible that the users were able to utilize it to find a greater variety of expressions, perhaps only because they would be simply exposed to more collocates than one would encounter from using a dictionary lookup.

→ *Table 31*

7.5.2.5 *Survey Results*

After the experiment, the participants were also asked to complete a questionnaire survey about their experiences of using Natsume that consisted of ten 5-point items ("Strongly disagree", "Disagree", "Neither agree nor disagree", "Agree", and "Strongly agree"). The survey results are shown in Figure 30. The results of the post-experiment survey were in general fairly positive about the usability of the interfaces and about the usefulness of the system for report writing, although attitudes were more pessimistic about using the system for class preparation or e-mail writing. However, the lower evaluations of the system for e-mail writing are also fairly reasonable given that Natsume does not contain an e-mail corpus, although the lower evaluations about using the system for class preparation would seem to warrant further investigation in the future. Finally, responses to the "not tiring" item would seem to suggest that there are still some aspects of the interfaces or the usage patterns that could be further improved to make the experience of using the system even smoother.

While not attempting to systematically tackle all the issues that the less positive comments to the "not tiring" item would seem to justify, this thesis has also conducted a multiple linear regression analysis for the amounts of time that each participant in the two experimental conditions spent to complete the reports. While the regression analysis included participants' L1, language proficiency scores, and report length in characters as predictors, it did not consider the possible interactions between the independent variables (condition, theme, and group) as the experimental design was insufficient (cf. discussion in §7.5.1.3).

The results of the multiple regression analysis as shown in Table 32, only revealed a significant effect of report theme on the time taken to complete the

Table 31: Variety of collocations containing common nouns and verbs for both control and treatment conditions.

Keyword	Collocates	Control	Treatment
実験 /jikken/ 'experiment'	を 行う /o okonau/ 'to perform'	7	8
	を 実施する /o jisshisuru/ 'to carry out'	0	2
	を やる /o yaru/ 'to do' (colloquial)	2	1
	を する /o suru/ 'to do' (neutral)	10	0
	が 必要 /ha hitsuyō/ 'is necessary'	0	2
進んでいる /susundeiru/ 'advancing'	科学研究 が /kagakukenkyū ga/ 'scientific research'	0	1
	発展 が /hatten ga/ 'growth'	0	1
	グローバル化 が /gurōbaruka ga/ 'globalization'	0	1
	グローバル が /gurōbaru ga/ 'global'	1	0

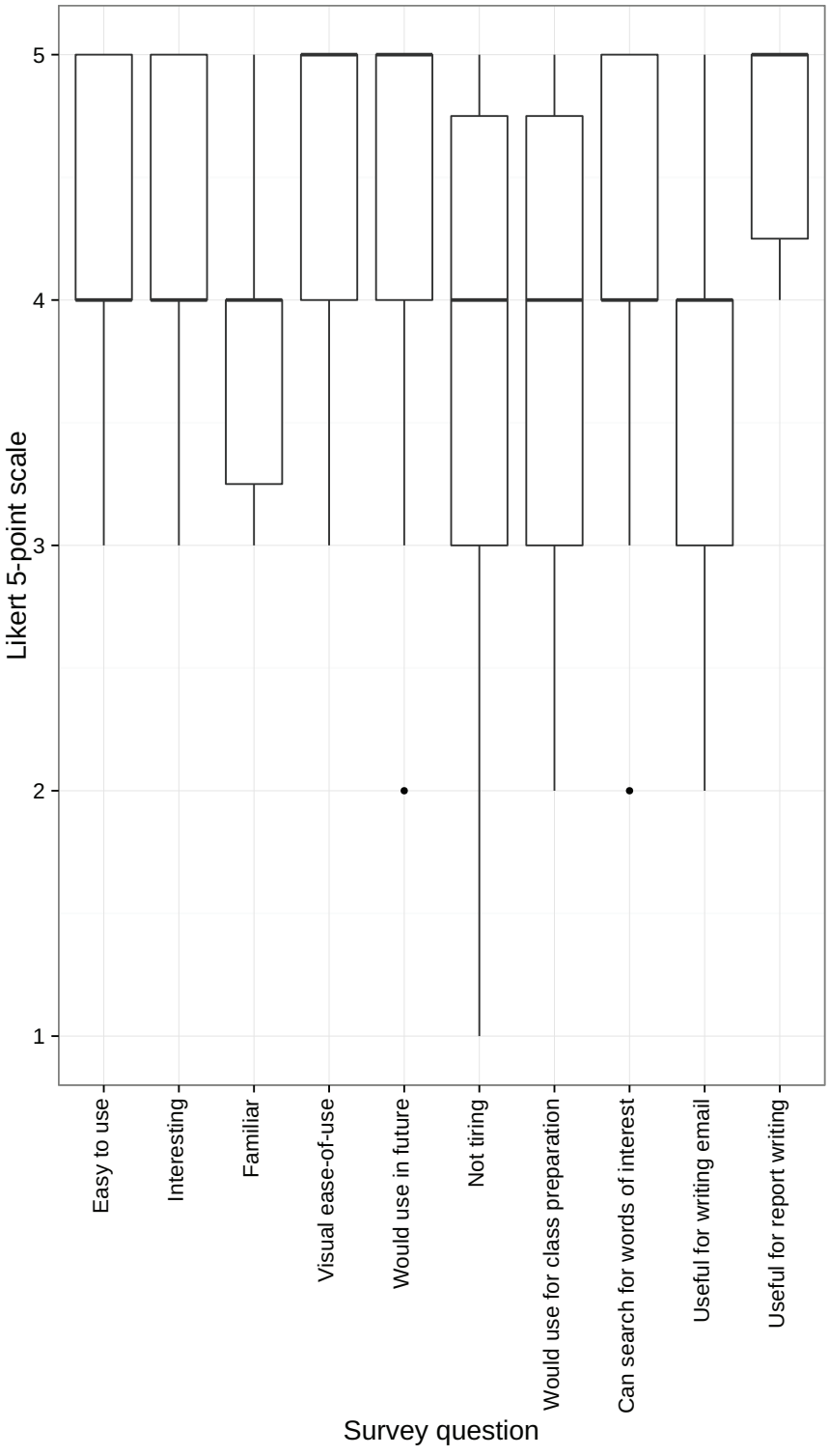


Figure 30: The results of the post-experiment survey (5-point Likert scales: 1 = "Strongly disagree", 2 = "Disagree", 3 = "Neither agree nor disagree", 4 = "Agree", and 5 = "Strongly agree"; Cronbach's alpha = 0.604).

Table 32: The results of the multiple linear regression on the time taken to write the reports.

Coefficients	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	47.525	9.782	4.86	3e-05	***
L1 (Korean)	6.653	3.978	1.67	.104	
L1 (Other)	-1.407	3.983	-0.35	.726	
Language proficiency score	-0.306	0.240	-1.28	.211	
Characters	0.004	0.007	0.62	.539	
Condition (treatment)	2.508	2.443	1.03	.312	
Theme (2)	5.769	2.471	2.33	.026	*
Group (B)	3.506	2.740	1.28	.210	

Residual standard error: 7.7 on 32 DF
Multiple R^2: .303, Adjusted R^2: .151
F-statistic: 1.99 on 7 and 32 DF, p-value: .0879

Significance codes: *** $\rightarrow$ 0.001, ** $\rightarrow$ 0.01, * $\rightarrow$ 0.05, . $\rightarrow$ 0.1

reports ($t = 2.33, p = .026$), with a multiple comparison test using Tukey contrasts for the differences in mean times between report themes revealing an estimated 5.77 minute difference where theme 2 was longer ($t = 2.33, p = .026$). It should be mentioned that the overall fit of the model was low and that none of the predictor variables adequately explained all of the variation in time observed within the data. However, the different themes would naturally impose different kinds of constraints on the word choices that the respective participants might make, which highlights the need for careful control to match for theme within similar studies in the future.

7.5.2.6 *Discussion*

While the evaluations indicated improvements for certain collocation usages in terms of their appropriateness of register, syntax, and semantics, further examinations of the usage logs for Natsume and of the participant essays revealed that many of the collocations could have been more effectively rewritten or corrected if the Natsume system had been used. However, given that the participants were not aware that they were making a mistake, naturally, they did not attempt to use Natsume in those cases.

Moreover, collocations that were not covered by Natsume's Noun-Particle-Verb and Noun-Particle-Adjective patterns were also found to be a common source of errors. Obviously, it is important to try and expand the range of collocation types covered by including other common types, such as those involving adverbs.

7.6 RELATED WORK

While predominantly researcher-focused, the ongoing work at NINJAL's Center for Corpus Development[5] has led to the development of two corpus search systems for the BCCWJ. One is Shonagon[6], which is a freely accessible corpus search system that offers basic KWIC search features, while the other is Chunagon[7], which is a subscription-based system that allows for searching with regular expressions over both short and long unit words.

Another system that shares Natsume's focus on Japanese language education is NINJAL-LWP[8], a lexical profiler for a subset of the BCCWJ (Pardeshi et al., 2012). It contains features similar to Natsume, but differentiates itself by providing many different kinds of collocations. Perhaps the most sophisticated collocation query system, which has also been applied to Japanese, is the Sketch Engine (Kilgarriff, Rychly, Smrz, & Tugwell, 2004; Srdanović-Erjavec, Erjavec, & Kilgarriff, 2008). The Sketch Engine is described as a "Corpus Query System incorporating word sketches, one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour" ("Sketch Engine: SketchEngine," 2013). It supports multiple languages, including Japanese. Two Japanese corpora are available: a 400 million token web-based corpus (JpWaC) that was released in 2008 (Srdanović-Erjavec et al., 2008), and a larger JpTenTen web corpus comprised of 8.43 billion tokens, released in 2013 (Srdanović, Suchomel, Ogiso, & Kilgarriff, 2013). More than 50 collocational and grammatical relations are used for the word sketch grammar. The Sketch Engine also contains a unique word comparison feature, called word sketch difference, which is effectively quite similar to searching for several words at the same time using Natsume, though it is also more sophisticated. An experimental version of JpWaC that includes tagged modality is also available (Srdanović, Hodošček, Bekeš, & Nishina, 2009).

7.7 CONCLUSION AND FUTURE WORK

Natsume was developed to assist a specific part of the writing process—finding the right words in a particular writing context, which, for the purposes of evaluation, was set to the scientific and technical Japanese register.

Two evaluation experiments were conducted to measure the effectiveness of the Natsume system for both a rewriting task and a report writing task. The first evaluation experiment was performed to assess the effectiveness of the system in a sentence rewriting task. Overall, mean scores under the treatment

5 An overview of research activities is provided at http://www.ninjal.ac.jp/english/organization/chart/06/.
6 Accessible from http://www.kotonoha.gr.jp/shonagon/.
7 Accessible from https://chunagon.ninjal.ac.jp/.
8 Accessible from http://nlb.ninjal.ac.jp/.

condition were higher than under the control condition, suggesting that the genre comparison interface provided by Natsume helped participants to select more appropriate collocations for academic writing. However, further examination of the rewrite strategies employed by five participants who scored higher under the control condition also revealed the use of word classes, collocational patterns, and expressions not searchable with Natsume. Expanding Natsume to encompass these patterns should be the focus of future work on the system.

Regarding the usability of the system, the effect of using the system with minimal training was not shown to be detrimental; as evidenced by the participants' ability under the treatment conditions, on average, to outperform on the rewriting tasks in the first experiment or the writing tasks in the second experiment. A detailed examination of the writing times required during the second experiment also revealed an effect of the report theme on writing time. Future evaluation experiments should take special steps to carefully control for theme by pre-screening the participants in terms of the knowledge of the candidate topics before they commence some writing task.

In addition to emphasizing the need for an expanded set of collocational patterns and the careful selection of themes, participant feedback from the report writing experiment was also helpful to identifying some necessary improvements to the interface and system features:

- One common source of confusion concerning the interface was the need to select a specific collocational pattern (NPV or NPAdj) and the part of the collocation that one searched from (noun, verb, or adjective) when searching. Based on morphological analysis of the search input, an automatic inference mechanism could be devised that would make such selections unnecessary.
- The addition of more grammatical relations to the search. Commonly requested relations include relations with adverbs and sentence-final modality form, among others.
- Currently, the example sentences are displayed in random order. The ability to show the sentences most appropriate to the learner's level and interests would be highly preferable. Tailoring the example sentence view to display examples that are appropriate to the learner's proficiency level is an important future goal (Hodošček et al., 2012).

See §6 for such an attempt

In order to reach the final goal of improving academic writing, it will be necessary to develop some strategies for closing the gap between standard academic report writing and scientific and technical Japanese writing. This would involve detailed specifications of the scientific and technical Japanese register and its relation to the standard academic register of essays and reports. Similarly, it would also important to develop curricula that incorporate elements specifically designed to enlighten learners about the register differences that exist for language.

Finally, there is also a need for methods that can foster greater self-awareness in learners regarding the mistakes that they commonly make while writing, in order to effectively help them overcome these problems.

8 NUTMEG

Natsume, while useful for finding collocations, does not automatically detect and correct the learner's writing. As presented in the previous chapter, the on-going evaluation of Natsume has identified a class of learner errors that are not being covered by Natsume: namely, that the learners themselves are not aware that they are making mistakes. In order to address this problem, a project was undertaken to develop a system that can check learner writing and provides feedback on any errors that it detects. The Nutmeg writing assistance system provides basic feedback for learner writing using automatic error identification (Yagi et al., 2012). The system realizes this correction using two sources: native and learner corpora.

8.1 NUTMEG ERROR INTERFACE

The writing interface consists of a survey section for user information and a writing section. When users fill in the requisite survey forms, which include the user's L1 and Japanese language proficiency level, they can proceed to write or copy and paste some text into the text box. Clicking on the 添削 /tensaku/ "correct" button will then underline possible errors and provide feedback on mouse hover. The feedback consists of a short explanation of why the system is indicating a candidate error, but does not provide the user with any solution suggestions. It is, thus, more a form of indirect feedback, which prompts the user to reflect further by applying their existing knowledge of language rules in order to correct the error themselves (Godwin-Jones, 2008).

→ *Figure 31*

→ *Figure 32*

Currently, there are several different error types checked for, including:

- Stylistic inconsistency

  This refers to the mixing of "dearu-cho" with "desu/masu-cho" in the same text.

- Orthographic inconsistency

  This refers to when a writer is using different orthographic variations for the same word within the same text (当たり前, 当り前, and あたりまえ for /atarimae/ "obvious, usual").

- Register mismatch

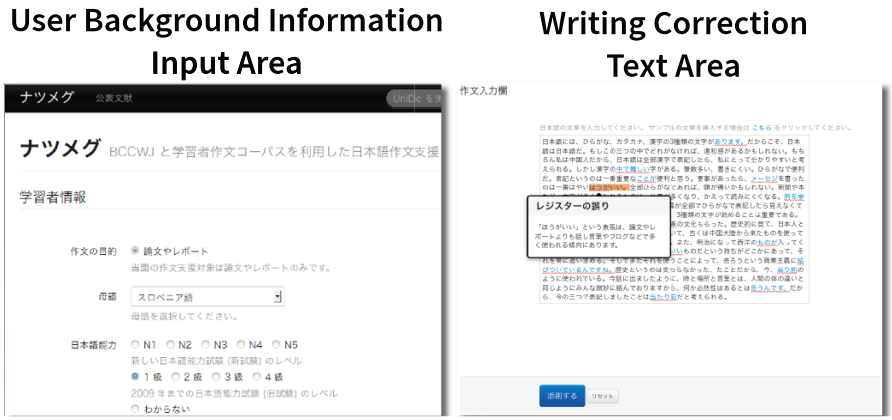


Figure 31: Input area for information about the user's background and the interface for writing text to be corrected within Nutmeg.

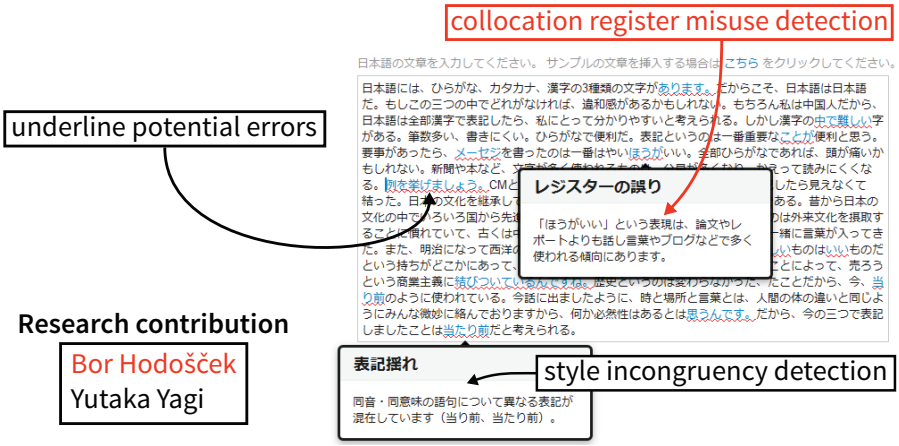


Figure 32: The Nutmeg writing interface.

The next section focuses on the register correction feature that makes use of the Natsume API. The other features are described in Yagi et al. (2012).

8.2 REGISTER MISUSE CORRECTION METHOD

A relatively obvious, but nonetheless, vital observation to make about register is that L2 learners are typically exposed to far fewer L2 registers than the variety of L1 registers they would have encountered. Moreover, it is a matter of debate as to whether knowledge or familiarity with some L1 register can be transferred to what could be assumed to be a comparative L2 register.

Despite these reservations, this thesis essentially adopts the assumption that it is, at least, possible to foster substantial awareness concerning the differences between certain L1 registers and their corresponding L2 registers. If a learner is able to visualize the differences between related L1 and L2 registers, it could open up a path that could allow the learner to assimilate appropriate knowledge regarding the differences and, ultimately, achieve a kind of workable understanding that would be essential to writing in the relevant L2 register. Another way of conceptualizing this awareness is as an expansion of the writer's field of experience that enables them to write texts that are appropriate to the target L2 register. Drawing on an example of register contrasts noted in the previous chapter (Figure 24), for instance, if an English native writer could consciously perceive the register difference between "to do an experiment" and "to conduct an experiment", it could potentially help them in realizing that a parallel contrast also exists between the spoken Japanese of 実験をやる /jikken o yaru/ and the equivalent written Japanese of 実験を行う /jikken o okonau/.

Though the language contained in the BCCWJ and STJC should, in principle, be considered correct, this does not preclude the use of native corpora as instruments in identifying learner errors. One example is making use of the various genres available in Natsume, through which it is possible to correct collocation usage from the genre perspective. For example, using data from Natsume it is possible to automate the process of checking if a collocation is appropriate for an academic report as shown in Figure 24. If a learner uses the collocation /jikken o yaru/ in an academic report, the system will be able to identify the inappropriate usage and offer the replacement collocation /jikken o okonau/ as a correction. This is possible because of the existence of relatively incompatible genres, such as those that lean towards a formal writing style (STJC and White papers) and those that lean towards an informal or spoken writing style (Yahoo! Blogs, Yahoo! Q&A, and Diet minutes) (Hodošček & Nishina, 2011b). The corpora in Natsume can thus be divided into so-called positive and negative genres and the relative frequencies of collocations in those genre sets can be tested using the χ^2 test. In fact, this is exactly how the genre comparison interface in Natsume is used.

Following the example from Figure 24, when an expression like /jikken wo yaru/ is used it is possible to determine that it is incorrect because its frequency in the negative genres is significantly high, while its frequency in positive genres is significantly low. A replacement collocation could be found by searching through similar collocations and testing them in the same manner or by using thesaura such as WordNet to expand the available search space (Bond et al., 2009; Isahara et al., 2012).

8.2.1 *Register Identification Method*

As a formalization of the manual comparison process outlined in Figure 24, the following method was proposed that can classify a collocation as either right or wrong (Hodošček & Nishina, 2011b):

1. Perform χ^2 tests for variance on the relative frequency f of a word or a collocation for all corpora $C_0, C_1, \cdots, C_N$, with the average frequency across all corpora denoted by $\bar{f_K}$

$$\Delta f(k) = \begin{cases} f(k) - \bar{f_K}, & \text{if } \frac{(f(k)-\bar{f_K})^2}{\bar{f_K}} > \chi^2(\text{df} = N-1; \alpha = .05) \\ 0, & \text{otherwise} \end{cases} ; \forall k \in 0, \cdots, N \tag{2}$$

2. Assign the selected corpora to either the positive (C_+) or negative (C_-) corpus groups, compare their sums, and compute the score s for each chosen word

$$s = \begin{cases} 1, & \text{if } (\sum_{j \in C_+} \Delta f(j) \geq 0) \wedge (\sum_{j \in C_-} \Delta f(j) < 0) \rightarrow \text{correct usage} \\ -1, & \text{if } (\sum_{j \in C_+} \Delta f(j) \leq 0) \wedge (\sum_{j \in C_-} \Delta f(j) > 0) \rightarrow \text{incorrect usage} \\ 0, & \text{if neither is statistically significant} \rightarrow \text{no result} \end{cases} \tag{3}$$

Finally, given that the distributions for most collocations tend to be sparse, even with the amount of data currently available within Natsume, it is most crucial, for all but the most common words, to avoid classifying a collocation as an error solely on the fact that the collocation in question does not appear in any of the system corpora. Thus for collocations or words not included within Natsume's database, a score of -2 is returned in the API provided to Nutmeg.

8.2.2 *Word Register Identification*

From the 'register' domain of the hierarchical error classification hierarchy of Natane (Figure 10)

In order to evaluate the effectiveness of the method, learner errors and corrections that are closely related to register misuses were extracted from Natane annotations (Hodošček & Nishina, 2011b). From the total of 416 initially extracted annotations, some were excluded, such as simple stylistic corrections like correcting ですから /desu kara/ "because" for だから /da kara/ "because", which resulted in a set of 195[1] annotations, which were then broken down by word class. The most frequent error annotation was for conjunctive particles (94), followed by adverbs (65). In the next step, uncorrected and corrected words were separated from the annotation, and the register identification method outlined in the previous section was applied[2].

For positive corpora (C_+), the STJC and White papers media were selected, while for negative corpora (C_-), Yahoo! Q&A, Yahoo! Blogs, and the Minutes of the Diet media were selected.[3] For the uncorrected words, counts were made for how many instances were correctly identified as being incorrect (hits: $s = -1$) and how many instances were misidentified as being correct (misses: $s = 0/1$). Similarly, for the corrected words, counts were taken of how many instances were correctly identified as correct (hits: $s = 1$) and how many instances were misidentified as being correct (misses: $s = -1/0$).

Summarizing the results of Table 33, while it was possible to detect probable mistakes (79%), the corrections made by the annotators were harder to detect as correct (only 15% were identified as such). For example, in one annotation, correcting for use of the adverb 多分 /tabun/ "probably" changing to the adverb 恐らく /osoraku/ "probably", /tabun/, was correctly identified as an error, but the method failed at correctly identify /osoraku/ as being a correct answer. This is because neither adverbs appear in the positive corpus set, because they express uncertainty, which tends to be avoided within scientific and technical writing. As the Natane corpus mainly contains essays, which were gather together under less stringent criteria, such expressions were generally not regarded as critical by the annotators, and rather than deleting them, the annotators proposed rewrites that involved more formal expressions. Another example of an unidentified error and the unidentified correction set is the sentence final modality form ～かもしれない /kamoshirenai/ "may be ~". While the use of expressions with this sentence final modality is conceivable in essays where learners are required to state their opinion on topics, the use of such modal expressions is rarely observed within

1 Duplicate words were removed.

2 As this experiment evaluates words and expressions not available in Natsume at the time of the experiment, all frequencies were extracted using regular expressions with the `Sphinx` full-text search engine (http://sphinxsearch.com/).

3 The choice of positive and negative corpora was made based on previous experience with the corpora and because the selected corpora roughly correspond to opposite ends of the formal-informal and written-spoken language scales.

Table 33: The results for register-related classifications based on word-level evaluations of annotations within Natane.

	Uncorrected		**Corrected**	
Identified	44	(79%)	10	(15%)
Unidentified	12	(21%)	57	(85%)
TOTAL	56	(100%)	67	(100%)

Table 34: The results of identification performance for the NPV, NPAdj, and AdjN collocational patterns extracted and expanded from Natane error annotations (Reproduced from Yagi et al., 2012).

	NPV		**NPAdj**		**AdjN**	
Not found	598	(%54)	16	(%43)	11	(%20)
Failure	41	(%4)	6	(%16)	17	(%30)
Not enough data	468	(%43)	14	(%38)	26	(%46)
Success	3	(%0)	1	(%3)	2	(%4)
TOTAL	1,110	(100%)	37	(100%)	56	(100%)

scientific and technical Japanese papers. Thus, this aspect of the Natane data, which includes more "colloquial" essays and reports, is undoubtedly one factor that contributed to the lower detection scores.

8.2.3 *Collocation Register Identification*

A similar experiment conducted on collocations is described in (Yagi et al., 2012). In that experiment, three types of collocations, available within Natsume, were first extracted from the Natane corpus and matched to the annotations: Noun-Particle-Verb, Noun-Particle-Adjective, and Adjective-Noun. The extracted 117 collocations were then expanded with data from Natsume's database and WordNet (Bond et al., 2009; Isahara et al., 2012). The expanded set of collocations (1,203 sets) was then evaluated using the register identification method described in §8.2.1.

→ *Table 34*

The following are some examples of collocations that the method was able to correctly identify as errors:

- ことがある /koto ga aru/ "there is ~"
- 問題が起きる /mondai ga okiru/ "a problem will occur"
- 結論を出す /ketsuron o dasu/ "to conclude"
- 一緒にする /issho ni suru/ "to put together"

- いい経験 /ī kēken/ “good experience”

While all of these expressions represent correct Japanese usages in less formal registers, they stand out as inappropriate for the scientific and technical Japanese register. On the other hand, the following are some examples of collocations that were incorrectly identified as being incorrect:

- 子供がいる /kodomo ga iru/ “there are children”
- 仕事をする /shigoto o suru/ “to do work”
- 大学に行く /daigaku ni iku/ “go to university”

These collocations were incorrectly classified for the same reason noted in the precious section: the Natane corpus is biased towards more opinion-eliciting essay writing and towards topics not commonly found within the STJC.

8.3 RELATED WORK

Compared to other existing systems of automatic composition correction for the Japanese language, Nutmeg strives to incorporate both native and learner corpora within its correction model. An example of a similar, but narrower, application system is Chantokun[4], developed at the Nara Advanced Institute of Science and Technology (NAIST). More specifically, Chantokun is a system that seeks to detect and correct for misuses of case particles based on corrected Japanese language sentences from the Lang-8 website[5], a language-exchange social networking website, where users with different first languages correct each other’s writing (Mizumoto & Komachi, 2012).

→ *Figure 33*

In contrast, as an example of a system that emphasizes the employment of a native corpus for automatic error correction there is also the Japanese proof-reading system Tomarigi[6] (Oono & Inazumi, 2011), while yet another example is the jcorrect tool, which uses the dependency structure of a sentence to revise complex sentences into ones that should be easier to understand (Oosaki, 2006).

8.4 CONCLUSION AND FUTURE WORK

The ultimate aim for the Nutmeg system is to develop a compositional tool that is capable of automatically warning learners of potential writing mistakes as they are making them. The two types of data that power Nutmeg’s error correction facilities are native and learner corpora that correspond to the data provided by Natsume and Natane, respectively.

4 Accessible from http://cl.naist.jp/chantokun/.
5 Accessible from https://lang-8.com/.
6 More information and download links available at http://www.pawel.jp/outline_of_tools/tomarigi/.

Figure 33: The correction interface of the Chantokun system.

Although the size of the available native corpora is much greater than that of learner corpora, the learner corpora provides unique information concerning the common errors made by different groups of learners. For example, future work could also utilize information specific to Chinese speakers and provide detection features that are tailored for the errors that are common made by this learner subgroup.

One avenue that could potentially greatly enhance the correction of collocation errors is to provide candidate replacement expressions for the learner errors, perhaps by using WordNet. Within the present system, feedback on the possible errors is indirect in nature and does not explicitly assist the user find a suitable correction. The potential benefits of incorporating more direct feedback mechanisms, perhaps in the form of providing candidate replacement expressions, should also be considered in the future.

It is also clear that it would be most advantageous to greatly expand the scale of Natane, in order to realize more comprehensive evaluations. Specifically, the corpus should be expanded to contain more advanced forms of writing produced by L2 learners, such as drafts of research papers sent for publication.

Naturally, one must exercise some caution over how systems of automatic error correction are actually implemented, because the performance of the system will be highly dependent on the objectivity of the annotations employed within the system. It is, for instance, particularly difficult to maintain appropriate levels of objectivity when annotators are seeking to correct for learner errors at the semantic and discourse levels, where some degree of subjectivity may seep

in. Accordingly, a more effective first step would seem to focus on relatively straightforward kinds of errors, such as orthographic errors (Yagi et al., 2012).

Finally, greater effort could beneficially be directed towards obtaining or constructing specific-purpose corpora, so that the system can also cover a wider range of writing genres, such as business or e-mail writing.

Part IV

CONCLUSION

9 CONCLUSION

Part I of this thesis provided some of the background knowledge necessary for understanding the attempts undertaken within the thesis to implementing various writing assistance systems for L2 Japanese writers. More specifically, Part I provided background on the following: (1) the four scripts of the Japanese writing system were introduced, with explanations about how they are commonly used together in representing Modern Japanese in writing, (2) the current state of computer-aided writing assistance with a focus on Japanese writing, (3) the current situation of L2 Japanese learners both within Japan and abroad, and (4) the importance of collocations for L2 learning. Part I also framed the two research questions that shaped and informed the various research projects conducted for this thesis.

Part II of the thesis turned to introduce the key concept of writing context and, by taking inspiration from Schramm's model of communication, to trace out how the linguistic concepts of register, topic, and readability interact together in forming the writing context. The primary objective for Part II was to demonstrate the feasibility of utilizing quantitative methods for drawing on corpora to realize systems of writing assistance. In that connection, three distinct, but highly integrated models of register, topic, and readability were proposed and individually evaluated in terms of their respective potential contributions to the overall performance of systems of writing assistance.

As attractive solutions to some of the problems outlined in Part I, Part III of the thesis examined in detail the construction, evaluation, and evolution of the two writing assistance systems, Natsume and Nutmeg. Part III also sought to demonstrate the potential performance improvements that can be realized through the incorporation of the various models developed in Part II.

Finally, Part IV of the thesis provides a summary of the various research contributions proposed throughout the thesis and reflects on some interesting suggestions for future research.

9.1 SUMMARY OF CONTRIBUTIONS

This thesis proposed a novel approach for applying writing context to the development of systems of writing assistance for L2 Japanese learners when writing reports and papers.

1. Modeled the writing context on the language variation inherent in corpora.

   The contextual model comprises of three distinct models relating to different aspects of context: register, topic, and readability. The topic and register models were able to successfully discriminate between differences within corpora relating to both register and topic. The readability model, while not totally successful at predicting readability at the level of the sentence, should be capable of assessing the feasibility of several predictors of readability in future studies, including those corresponding to syntactic complexity and vocabulary that were proposed within this thesis.

2. Applied the concept of register to two writing assistance systems, based on a framework that emphasizes how users themselves can learn to discover differences in collocation usages across registers, with a special focus on academic writing.

   The framework seeks to reduce the need for L2 users to rely on intuitions concerning L1 language variations, which L2 learners are less likely to possess, at least in the earlier stages of their L2 acquisition. Within Natsume, the user can discover such differences by themselves, by searching for and comparing candidate collocations, and by observing their frequency variations within different corpora, as well as by consulting with further example sentences, as necessary. The results of evaluation experiments of Natsume point to some potentially interesting ways of improving collocation usages. Within Nutmeg, the user relies on the system to identify inappropriate uses of register, which the user may not be aware of. The register identification method underlying the Nutmeg system, that utilizes collocation data from Natsume, showed some promise in terms of identifying probable inappropriate uses of words and collocations within L2 academic writing.

Taken together, the contributions obtained through this research provide some interesting answers for the two research questions posed in Part I of this thesis.

9.2 RESEARCH QUESTIONS REVISITED

This section summarizes these answers to the research questions posed in §1.6.

- *Research Question 1*

  > Is it possible to utilize linguistic resources, such as corpora, to realize more effective systems for computer-assisted writing?

Given that the process of L2 writing typically entails the user making use of various dictionaries and other forms of references, one may ask whether it is possible to utilize corpora to go beyond the inherent limitations associated with these time-tested tools. The type of corpus-driven writing assistance systems proposed in this thesis focus on helping L2 learners of Japanese find collocations suitable to their writing purposes. Formed from commonly cooccurring words, the number of possible collocations is far greater than the number of words themselves, and, thus, requires a different approach for the effective presentation of such quantities of information. The interfaces for Natsume provide user-friendly ways of visualizing the information content relating to collocations, by taking advantage of knowledge about the kinds of writing contexts that collocations commonly occur in. As a consequence, Natsume is able to provide a form of surrogate knowledge about register differences within Japanese for the L2 users. In this way, the system is effectively able to not only sidestep the problems that L2 learners face due to a lack of appropriate intuitions about register differences, but it also effectively offers richer learning opportunities for the users to make discoveries about collocations by themselves.

Taking this concept even further, the Nutmeg system can detect register misuses relating to collocations within the user's writings. This approach allows the system to not only detect grammaticality issues, but also paves the way towards realizing a correction system for academic writing style.

- *Research Question 2*

  > Is it possible to provide writing style guidance based on corpora alone?

  Evaluations of Natsume revealed that when users have been explicitly instructed to prefer collocations that have higher frequencies within the STJC corpus, they are generally able to produce more appropriate collocations within their report writing and sentence rewriting. However, the gathered evaluations did not examine the possibilities of using the system to deepen learner understanding of L2 registers, so that remains an open question to be taken up in future studies. On the other hand, evaluations concerning the detection of register misuses relating to words and collocations within the Nutmeg system highlighted some interesting possibilities of providing writers with feedback about possible errors. Currently, the Nutmeg system provides indirect feedback about such possible learner errors, which can be beneficial in prompting the writer to reflect further on their writing. There are, however, still a number of unanswered questions that could be addressed. For instance, it would be interesting to

investigate and compare which feedback strategies could be more effective, such as specific scaffolding strategies designed to guide users to appropriate decisions but to later gradually fade back as the learner gains deep insights into the mechanisms for their own mistakes.

9.3 FUTURE WORK

Having briefly summarized the main contributions of the thesis and revisited its motivating research questions, this section offers some structured guidelines concerning possible directions for future research, which are informed by both some still unresolved and some newly-uncovered issues that have been touched upon in previous sections. While these directions are tentatively organized around some relevant research issues, they also seek to emphasize the potential synergies wherever they exist.

9.3.1 *Incorporation of Models of Context into Writing Assistance Systems*

As the three proposed models, when taken together, provide even more accurate approximations of the writing context, it should become increasingly possible to handle even more nuanced types of corrections. With the enhanced integration of such systems, it will be possible to provide two levels of assistance for each component of the writing context:

1. Word and collocation level
    - Register: identify register misuses for words and collocations.
      This feature is already a part of Nutmeg.
    - Topic: identify words and collocations with similar topics.
      This feature would provide similar capabilities to the `getassoc` word similarity feature currently available within Natsume. The benefit of the topic model perspective could, however, be further extended to offer suggestions not only for words but also for related topics, offering the writer even more flexibility in their word choices.
    - Readability: identify word level.
      The ability to identify rare words is useful for writers that are concerned with the level of language proficiency among their target readers, such as L2 educators. Such a feature, based on JLPT word scores, as available within Kawamura, Kitamura, and Hobara (2013), could be provided.
2. Text level

- Register: identify closest register(s) to the input text.

  This would allow the writer to see, at a glance, the perceived register of the text. Furthermore, the writer could request sample documents from the system that closely match to his target resister.

- Topic: identify topics similar to the input text.

  This feature would allow the writer to browse documents related to the input text and is a typical application of topic models.

- Readability: identify sentences with low readability, and provide a readability report for the whole text.

  This would allow for similar uses to the word and collocation level case, but as it relates to the text as a whole, would be a form of indirect feedback, which the writer can use to reflect on. Additionally, this feature could allow more targeted feedback at the sentence or paragraph levels.

  Another use for readability assessment in writing assistance systems is that the writer could request documents or sentences at the writer's preferred readability level.

In other words, all aspects of the system would employ knowledge about the writer's proficiency level, chosen writing topic, and target audience. In a very real sense, the system would effectively *know* part of the field of experience possessed by the writer, and given the system's *knowledge* of the target audience, could also include some knowledge that reflects the field of experience possessed by the target audience. And, in the case of academic writing, such knowledge can be modeled in terms of differences between the STJC corpus and other corpora. Thus, having information about both parties to the communication, the system could guide the writer in narrowing the gaps between their own fields of experience and that to the audience's fields of experience.

One possible improvement to the method described in §8.2.1, that could effectively utilize the contextual model, could relate to the selection and classification of positive and negative corpora, which has traditionally been rather "ad-hoc" in nature, typically based on a researcher's perceptions about differences between corpora. The contextual model could represent a more principled approach towards the assessment of distances between candidate corpora and the weighting for occurrence of some word or collocation depending on the actual writing context.

9.3.2 *Evaluation of Models and Correction Methods with Learner Corpora*

More extensive evaluations of the register, topic, and readability models could be performed using the Natane Japanese learner corpus introduced in §3.4.

The Natane data on learner errors relating to register misuse have already been successfully used in the development and evaluation of methods for identifying register misuse. The application of topic models would, rather, allow gaining greater insight into the effect of writing topic to possible errors. A secondary, but possibly more salient benefit, would be the ability to measure the topic of learner essays and control for the report theme imbalance present in Natane. Finally, the readability of essays in Natane could also be used to uncover ways in which feedback about the readability of the learners writing might potentially lead to compositional improvements.

9.3.3 *Strategies for Improving Writing*

Most of the writing assistance features introduced into the Natsume and the Nutmeg systems focus on sentence-level assistance. However, given that these features are operating more at the micro-level, these are less relevant to the provision of more effective ways for users to structure their writings or improving the cohesion of their writing styles at more macro-levels of linguistic analysis. Accordingly, future work could potentially benefit greatly from paying more attention to the integration of writing assistance features at higher macro-levels, such as offering the writer of a report or academic paper a template or structural queues that can assist them.

APPENDIX

10

APPENDIX

10.1 READABILITY MODEL

10.1.1 *Tuning Parameters*

- Radial SVM
  - *sigma*
  - *C*: .25, .5, 1, 2, 4, 8, 16, 32, 64, 128, 256, 512, 1020, 2050, 4100

  The final values used for the model were
  - Sources: *C* = 4096 and *sigma* = 0.127. Tuning parameter *sigma* was held constant at a value of 0.1274519.
  - Paragraphs: *C* = 4 and *sigma* = 0.1. Tuning parameter *sigma* was held constant at a value of 0.1058.
  - Sentences: *C* = 1 and *sigma* = 0.1. Tuning parameter *sigma* was held constant at a value of 0.1058.
- C5.0
  - *model*: rules/tree
  - *trials*: 1, 10, 20, 30, 40, 50, 60, 70, 80, 90, 100
  - *winnow*: TRUE/FALSE

  The final values used for the model were
  - Sources: *model* = tree, *trials* = 30, and *winnow* = FALSE.
  - Paragraphs: *model* = rules, *trials* = 100 and *winnow* = FALSE.
  - Sentences: *model* = tree, *trials* = 40 and *winnow* = FALSE.
- Random forest
  - *mtry*: 1-15

  The final value used for the model was
  - Sources: *mtry* = 11.
  - Paragraphs: *mtry* = 7.
  - Sentences: *mtry* = 3.
- Neural net

- *size*: 1-10
- *decay*: 0, .1, 1, 2

The final values used for the model were

- Sources: *size* = 7 and *decay* = 0.1.
- Paragraphs: *size* = 9 and *decay* = 0.1.
- Sentences: *size* = 10 and *decay* = 0.1.

- glmnet
  - *alpha*: 0-1 in .1 increments
  - *lambda*: 40 segments between .01-.2

The final values used for the model were

- Sources: *alpha* = 0.9 and *lambda* = 0.0197.
- Paragraphs: *alpha* = 0.1 and *lambda* = 0.01.
- Sentences: *alpha* = 0.2 and *lambda* = 0.01.

10.1.2 *Model Differences*

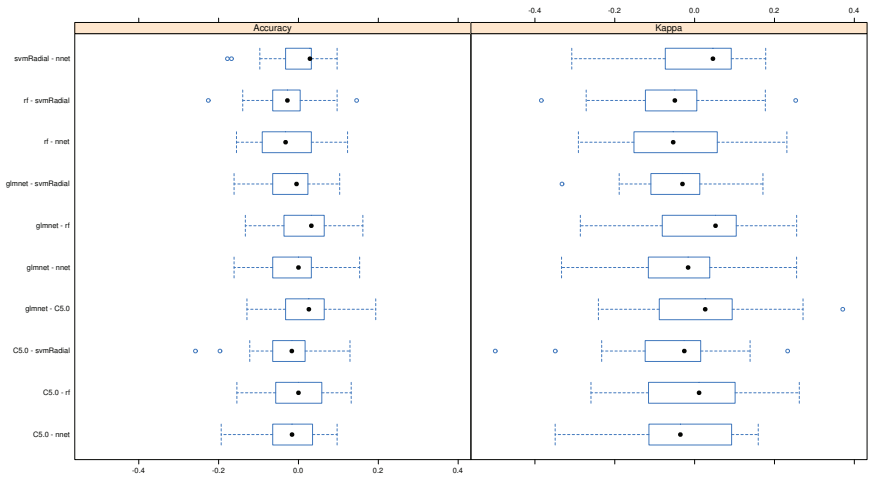


Figure 34: Pairwise comparisons between resampled models in terms of accuracy and associated kappa values for the sources level. Results indicate no significant differences between models.

Table 35: Model diagnostic statistic by E, M, and H class on the sources, paragraphs, and sentences held-out test sets.

Source															
	E					M					H				
	SVM	C5.0	RF	NNet	glmnet	SVM	C5.0	RF	NNet	glmnet	SVM	C5.0	RF	NNet	glmnet
Sensitivity	.957	.957	.957	.957	.957	.688	.563	.688	.438	.375	.919	.968	.968	.919	.984
Specificity	.962	.949	.962	.936	.962	.953	.988	.988	.965	1.000	.897	.872	.897	.821	.769
Pos Pred Value	.880	.846	.880	.815	.880	.733	.900	.917	.700	1.000	.934	.923	.938	.891	.871
Neg Pred Value	.987	.987	.987	.986	.987	.942	.923	.944	.901	.895	.875	.944	.946	.865	.968
Prevalence	.228	.228	.228	.228	.228	.158	.158	.158	.158	.158	.614	.614	.614	.614	.614
Detection Rate	.218	.218	.218	.218	.218	.109	.089	.109	.069	.059	.564	.594	.594	.564	.604
Detection Prevalence	.248	.257	.248	.267	.248	.149	.099	.119	.099	.059	.604	.644	.634	.634	.693
Paragraph															
	E					M					H				
	SVM	C5.0	RF	NNet	glmnet	SVM	C5.0	RF	NNet	glmnet	SVM	C5.0	RF	NNet	glmnet
Sensitivity	.722	.738	.695	.716	.715	.320	.304	.364	.268	.169	.910	.888	.900	.874	.921
Specificity	.925	.902	.924	.907	.913	.935	.932	.920	.915	.965	.649	.668	.675	.641	.551
Pos Pred Value	.727	.676	.715	.680	.692	.589	.564	.570	.480	.587	.766	.772	.778	.755	.722
Neg Pred Value	.924	.926	.917	.921	.921	.825	.821	.832	.811	.799	.851	.826	.843	.801	.846
Prevalence	.216	.216	.216	.216	.216	.226	.226	.226	.226	.226	.558	.558	.558	.558	.558
Detection Rate	.156	.159	.150	.155	.154	.072	.069	.082	.061	.038	.508	.496	.503	.488	.514
Detection Prevalence	.215	.236	.210	.227	.223	.123	.122	.144	.126	.065	.663	.643	.646	.646	.712
Sentence															
	E					M					H				
	SVM	C5.0	RF	NNet	glmnet	SVM	C5.0	RF	NNet	glmnet	SVM	C5.0	RF	NNet	glmnet
Sensitivity	.477	.521	.526	.533	.432	.063	.170	.170	.049	.000	.944	.902	.916	.919	.945
Specificity	.937	.921	.930	.908	.926	.985	.948	.959	.981	.999	.329	.450	.432	.378	.268
Pos Pred Value	.615	.581	.614	.551	.551	.524	.462	.520	.399	.000	.693	.725	.721	.703	.674
Neg Pred Value	.894	.901	.903	.902	.885	.799	.812	.814	.796	.791	.786	.741	.762	.745	.751
Prevalence	.175	.175	.175	.175	.175	.209	.209	.209	.209	.209	.616	.616	.616	.616	.616
Detection Rate	.083	.091	.092	.093	.076	.013	.036	.036	.010	.000	.582	.555	.564	.566	.582
Detection Prevalence	.136	.157	.150	.169	.137	.025	.077	.068	.026	.000	.839	.767	.782	.805	.863

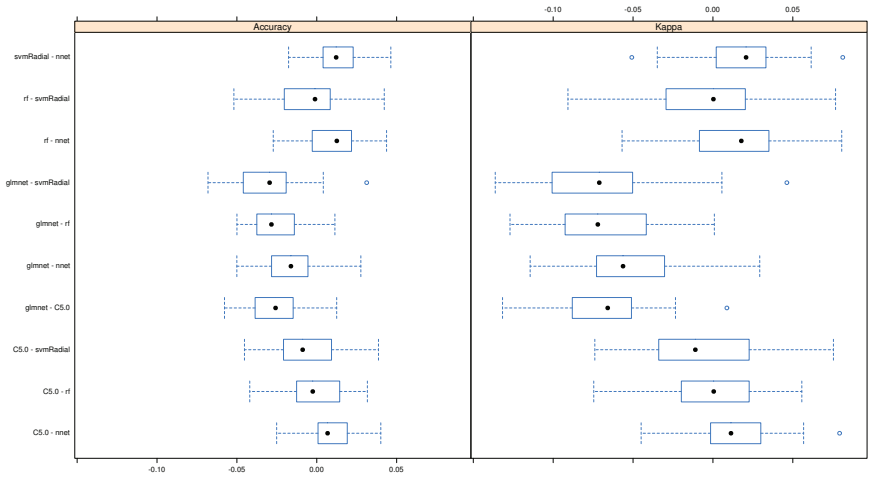


Figure 35: Pairwise comparisons between resampled models in terms of accuracy and associated kappa values for the paragraphs level. Results indicate no significant differences between models.

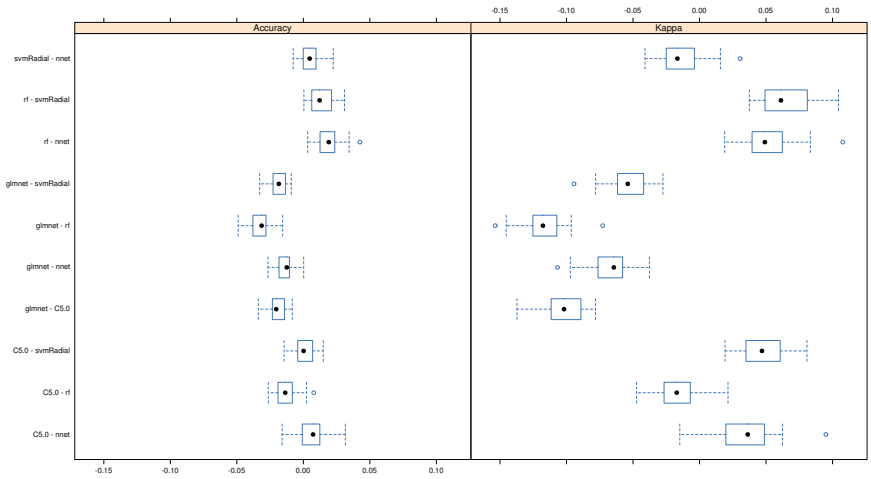


Figure 36: Pairwise comparisons between resampled models in terms of accuracy and associated kappa values for the sentences level. Results indicate significant differences between most models, excluding kappa estimate for the C5.0-svmRadial pair.

10.2 NATSUME EXPERIMENT 1

10.2.1 *Test Content*

The instruction given to both groups in the control and treatment cases was:

- Treatment case:

  次の問題文 1—15 の文をあなたが論文として書くつもりでコンピュータ上の「なつめ」を見ながら書き換えてください。名詞、動詞、副詞、述語などを書き言葉らしくするように注意してください。

- Control case:

  次の問題文 1—15 の文をあなたが論文として書くつもりで書き換（か）えてください。名詞、動詞、副詞、述語などを書き言葉らしくするように注意してください。

10.2.1.1 *Test A*

例題と解答例
　この方式が十分な効き目を発揮（はっき）するためには，エンジンの精度が悪くないという条件がいいと思うんです。→この方式が十分な効果を発揮するためには，エンジンの精度が悪くないという条件が必要である。

1. そこで働いている人たちに意見を聞いてデータを集めるよい手法を考え出したので、その新しい考えをみなさんに言う。
2. 若い人達に大きい影響をあげる人々の意見は大切である。
3. グラフ化するための簡単な解析をやってみた。(解析 analysis)
4. 大学四年生 100 人とその両親を対象に家族関係について調査をやったら、考え方に大きい違いがあることがわかった。
5. A 社は、いい品質の品物を売って多くの人からとても高い評価をもらっている。
6. 大気汚染によって、都市に住んでいる人々に大きな被害が来ている。
7. 試験管は試料を取り出して操作（そうさ）したり貯蔵したりするために使われる。
8. 自然保護について色々な観点から考えて、とうとうその計画をやめることになった。

9. 次の章でこの問題の見解を書く。
10. 天体の中の小惑星がぶつかっても、私たちは大きい影響をもらわないでしょう。（小惑星　miniplanet）
11. このデータから二国間の関係が悪くなって行ったことがわかる。
12. 作業項目を編集するときには、次のリストにあるきまりが適用される。
13. 電子工学についてもっと詳しいことを知りたい人は、研究に関係のある本を参照することができる。
14. このプログラムを作ったことで、経営がうまくいったことが検証できた。
15. エアコンとは冷暖房両方のはたらきを備えたものを指す。

日本のたいていの技術者は、してよいことと、してはいけないことの区別がつけられます。技術者は倫理的にだめだから倫理教育がいるのだという考えだったら、教育はできません。技術者たちは、学校で勉強して、会社にはいるための試験を受けて、選ばれてから会社に入って、まじめに働いています。そういう人達だからこそ、短い時間で、効率のよい教えができるのです。

技術者の倫理の勉強には、知識を勉強することと意識を定着させるという二つがあります。一度定着したら、その後何度も勉強しなくてもいいです。技術倫理の内容としては、これだけはどうしても要るというものがあります。そんなテキストを作っておいたら、それが学習者の手元に長く残って、問題を解くのに役立ちます。大事なのは学習者に倫理とは何かということが分かることです。

人々が仕事をするときには、事故を起こさないようにする義務があります。こんな義務を安全管理につなげて考えると、倫理についての考えがよく分かるようになるでしょう。（倫理　moral　公衆 public）

＜論文らしい文章を以下に書いて下さい＞

10.2.1.2 *Test B*

例題と解答例

この方式が十分な効き目を発揮（はっき）するためには，エンジンの精度が悪くないという条件がいいと思うんです。→この方式が十分な効果を発揮するためには，エンジンの精度が悪くないという条件が必要である。

＜問題開始＞

1. 新しい実験の方法を考え出したので、それを論文の中で言うつもりです。
2. はじめに実験結果の分析をやってから、それをもとに仮説を証明した。(仮説 hypothesis)
3. 高校に入学するには、本人の気持ちや希望だけではなくて、いくつか決まった条件を全部合うようにしなければなりません。
4. 二つのシステムの設計仕様を比べてみると，次のような違いがある。(設計仕様 design specification)
5. その学者は論文を書いて、とてもすばらしいと多くの人から高い評価をもらい、ノーベル賞をもらった。
6. 海にプラスチックなどのごみが捨てられて、海に住む生物に大きな被害が来ています。
7. 平和国家で武器を使うということは変則の事態である。
8. 第 5 章でこのことに対する見解を言いましたが、むずかしい問題です。
9. 近頃女性が外で働くようになってから、その家族は大きい影響をもらいました。
10. このグラフはタンクの水の深さが変わったことを示しています。
11. パートタイム労働者にも働き方のきまりが適用される。
12. ファクシミリを使って、ホームページで見られる知りたいことを手に入れるシステムがある。
13. 電子工学の研究のために必要なことが色々書いてある本を参照することができる。
14. このシステムが本当に役に立つということを検証するため，われわれは次のように調べた。
15. 指紋自動認識システムは、短時間で犯人を見つけ出す働きを備えたものである。(指紋　fingerprint　犯人 crime)

学校を卒業した後で、本を読んでも知識をもらうことが出来ます。一方で、学校で教えられたことが意識にしっかり入っていれば、実際の仕事で役に立ちます。どうやって意識にしっかり入れるか、いろいろなやり方があるはずです。技術者倫理の教育では、もらった知識を測るとい

う評価方法が合いません。でも、教える人は学生を評価しなくちゃいけません。

学習者は事例を使って、社会に本当にあった事から倫理に関係がある問題を取り出していく過程を勉強します。新聞記事などは近くにあって、けっこう面白いものだと学習者に思わせるのもいい効果になります。

授業では何を学生にあげることができるのでしょうか。一番目に、授業全部を忘れても一つだけ残ればよいということがあります。二番目には、卒業した後で、先生から教えてもらったことがノートのどこかに書いてあることを思い出して、探し出せるテキストがあることが大事です。でも、キャンパスの中に倫理を大事に思う考えがないと、学生たちに技術者の倫理について教えるのはばからしいですね。(倫理　ethics)

（論文らしい文章）

10.3 NATSUME EXPERIMENT 2

10.3.1 *Topic Introduction Text*

- 課題 1:「日本は動物実験を全廃にすべし」

  医学の進歩、化粧品や薬の開発などさまざまな分野で動物実験は欠かせないものと考えられてきた。動物実験はこれまでに、様々な感染症 (infectious disease) に対するワクチン開発の副作用の有無確認、あるいは化粧品の開発などに利用されている。

  一方では、その倫理上の責任の大きさから、あるいは動物愛護の立場から欧米では厳しい法的な規制を設けている。ある人々は、動物実験を全面的に止めるべきという。日本ではそのような法的規制はない。

  この両方の立場について意見を述べなさい。

- 課題 2:「日本は英語を第二公用語にすべし」

  英語の第二公用語化とは、英語を日本の第二公用語とする構想のことである。

  国際共通言語として機能している英語を「公用語」とすることで、その習得・利用を促進し、日本人の英語力および非日本語話者とのコミュニケーション能力の向上を目的とする。小学校 1 年生から英語を必修科目とする動きもある。

  一方では、日本独自の歴史的・文化的な資産が失われることへの危惧によって導入を反対する意見もある。英語を導入することによって国語教育が疎か (おろそか) になるという人々もいる。

  この両方の立場について意見を述べなさい。

Table 36: Contingency table for 2-gram collocations.

	a	$\neg a$	$*$
b	f_{ii}	f_{oi}	f_{xi}
$\neg b$	f_{io}	f_{oo}	f_{xo}
$*$	f_{ix}	f_{ox}	f_{xx}

10.4 STATISTICAL MEASURES OF COLLOCATION

Collocations are commonly defined as words that cooccurr in close proximity more than would be expected by chance. Statistical measures of collocations will use the following naming conventions for 2-gram (a, b) collocations:

a A unit.
$\neg a$ Any unit not including a.
b Another unit.
$\neg b$ Any unit not including b.
$*$ Any (all) units.
f_{ii} Frequency of a cooccurring with b.
f_{oi} Frequency of $\neg a$ cooccurring with b.
f_{xi} Frequency of all cooccurrences with b.
f_{io} Frequency of a cooccurring with $\neg b$.
f_{oo} Frequency of $\neg a$ cooccurring with $\neg b$.
f_{xo} Frequency of all cooccurrences with $\neg b$.
f_{ix} Frequency of all cooccurrences with a.
f_{ox} Frequency of all cooccurrences with $\neg a$.
f_{xx} Total number of units.

The relation between the provided terms can be seen in the contingency table presented in Table 36. Collocational measures are introduced following the classification of Evert (2004).

10.4.1 *Asymptotic Hypothesis Tests*

- t-test

$$t = \frac{f_{ii} - \frac{f_{ix} f_{xi}}{f_{xx}}}{\sqrt{f_{ii}}} \tag{4}$$

- χ^2

$$\chi^2 = \frac{f_{xx}(f_{ii}f_{oo} - f_{io}f_{oi})^2}{(f_{ii} + f_{io})(f_{ii} + f_{oi})(f_{io} + f_{oo})(f_{oi} + f_{oo})} \tag{5}$$

- Log-likelihood ratio (Dunning, 1993)

$$\begin{aligned} LLR = & \\ & 2(xlx(f_{ii}) + xlx(f_{ix} - f_{ii}) + xlx(f_{xi} - f_{ii}) \\ & +xlx(f_{xx}) + xlx(f_{xx} + f_{ii} - f_{ix} - f_{xi}) \\ & -xlx(f_{ix}) - xlx(f_{xi}) - xlx(f_{xx} - f_{ix}) \\ & -xlx(f_{xx} - f_{xi})), \\ & \text{where } xlx(f) \text{ is } f\,ln(f) \end{aligned} \tag{6}$$

10.4.2 *Point Estimates of Association Strength*

- Mutual Information (MI) score

$$MI = log_2 \frac{f_{ii}f_{xx}}{f_{ix}f_{xi}} \tag{7}$$

- Dice coefficient

$$Dice = \frac{2f_{ii}}{f_{ix} + f_{xi}} \tag{8}$$

- Jaccard coefficient

$$Jaccard = \frac{f_{ii}}{f_{ii} + f_{oi} + f_{io}} \tag{9}$$

Abekawa, T., Hodošček, B., & Nishina, K. (2010). Collocation search refinements in the natsume writing support system. In *Tokutei ryōiki `nihongo kōpasu' hēsē 21 nendo kōkai wākushoppu (kenkyū sēka hōkokukai) yokōshū* (pp. 243–244).

Abekawa, T., Hodošček, B., & Nishina, K. (2011a, March). Go no kyōki o kōritsuteki ni kensaku dekiru nihongo sakubun shien shisutemu Natsume no shōkai [Introduction to efficient collocation search in Japanese writing assistance system Natsume]. (Vol. 17, pp. 595–598). Proceedings of The 17th Annual Meeting of The Association for Natural Language Processing. Toyama: The Association for Natural Language Processing.

Abekawa, T., Hodošček, B., & Nishina, K. (2011b, August). Japanese writing support system Natsume using genre information. (pp. 774–775). International Conference on Japanese Language Education. Tianjin, China.

Abekawa, T., Hodošček, B., & Nishina, K. (2012, August). Nihongo sakubun shien shisutemu Natsume ni okeru aratana youhou no kumikomi. International Conference on Japanese Language Education. Nagoya, Japan.

Benjamin, R. G. (2012). Reconstructing readability: Recent developments and recommendations in the analysis of text difficulty. *Educ Psychol Rev*, *24*, 63–88. doi:10.1007/s10648-011-9181-8

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257.

Biber, D. (1995). *Dimensions of register variation: A cross-linguistic comparison*. Cambridge: Cambridge University Press.

Biber, D. & Conrad, S. (2009). *Register, genre, and style*. Cambridge: Cambridge Textbooks in Linguistics.

Blei, D. M. (2012, April). Probabilistic topic models. *Commun. ACM*, *55*(4), 77–84. doi:10.1145/2133806.2133826

Bond, F., Isahara, H., Fujita, S., Uchimoto, K., Kuribayashi, T., & Kanzaki, K. (2009, August). Enhancing the Japanese WordNet. (pp. 1–8). ACL-IJCNLP 2009. The 7th workshop on Asian Language Resources. Singapore.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. doi:10.1023/A:1010933404324

Cao, H., Kuroda, F., Yagi, Y., & Nishina, K. (2011, August). Analysis of error classfication framework for learner corpus. (Vol. 2, pp. 520–521). International Conference on Japanese Language Education 2011. Tianjin, China.

Cao, H., Kuroda, F., Yagi, Y., Suzuki, T., & Nishina, K. (2010, July). Gakushūsha sakubun shien shisutemu no tame no goyō dētabēsu sakusei: dōshi no goyō

bunseki o chūshin ni [Constrction of learner error database for composition assistance: Focus on error analysis of verbs]. (Vol. 2, pp. 1571–1579). International Conference on Japanese Language Education 2010. Taipei, Taiwan.

Cao, H. & Nishina, K. (2010). Establishment of error classification framework for error database. *Journal of Japanese language education methods*, *18*(1), 38–39.

Cao, H., Yagi, Y., Kuroda, F., & Nishina, K. (2012, August). Construction of learner corpus Natane and possible application. (pp. 1–4). 5th international conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J). Nagoya. Retrieved from http://2012castelj.kshinagawa.com/proceedings/Poster/Poster5_Cao.pdf

Dıaz-Negrillo, A. & Fernández Domınguez, J. (2006). Error tagging systems for learner corpora. *Revista española de lingüıstica aplicada*, (19), 83–102. Retrieved from http://dialnet.unirioja.es/descarga/articulo/2198610.pdf

Dunning, T. (1993, March). Accurate methods for the statistics of surprise and coincidence. *Comput. Linguist. 19*(1), 61–74. Retrieved from http://dl.acm.org/citation.cfm?id=972450.972454

Friedman, J. H., Hastie, T., & Tibshirani, R. (2010, February 2). Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, *33*(1), 1–22. Retrieved from http://www.jstatsoft.org/v33/i01

Fujiike, Y., Konishi, H., Ogura, H., Ogiso, T., & Koiso, H. (2011). Tyōtan'i ni motozuku 'gendai nihongo kakikotoba kinkō kōpasu' no hinsihiritu ni kansuru bunseki. In *Proceedings of The 17th Annual Meeting of The Association for Natural Language Processing* (Vol. 17, pp. 663–666). Toyohashi, Japan.

Godwin-Jones, R. (2008). Emerging technologies: web-writing 2.0: enabling, documenting, and assessing writing online. *Language Learning & Technology*, *12*(2), 7–13.

Gottlieb, N. (2008). Japan: language policy and planning in transition. *Current Issues in Language Planning*, *9*(1), 1–68.

Granger, S. (2003). Error-tagged learner corpora and CALL: a promising synergy. *The Computer Assisted Language Instruction Consortium (CALICO) Journal*, *20*(3), 465–480.

Halliday, M. (2009). Methods - techniques - problems. In M. Halliday & J. J. Webster (Eds.), *Continuum Companion to Systemic Functional Linguistics* (pp. 59–86). Continuum International Publishing Group.

Halliday, M. & Matthiessen, M. C. (2004). *An introduction to functional grammar* (3rd ed.). Hodder Education.

Halliday, M. & Webster, J. J. (2009). Keywords. In M. Halliday & J. J. Webster (Eds.), *Continuum Companion to Systemic Functional Linguistics*. Continuum International Publishing Group.

Hasan, R. (2009). The place of context in a systemic functional model. In M. Halliday & J. J. Webster (Eds.), *Continuum Companion to Systemic Functional Linguistics*. Continuum International Publishing Group.

Heilman, M. J., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2007). Combining lexical and grammatical features to improve readability measures for first and second language texts. In *Proceedings of the NAACL Human Language Technology Conference* (pp. 460–467). Morristown, NJ: Association for Computational Linguistics.

Hodošček, B. (2010). *Development of a register-based writing assistance system for academic Japanese* (Master's thesis, Tokyo Institute of Technology).

Hodošček, B. (2011). Word class ratios and genres in written Japanese: Revisiting the Modifier Verb Ratio. *Acta Linguistica Asiatica*, *1*(2), 53–62. Retrieved from http://revije.ff.uni-lj.si/ala/article/view/28/37

Hodošček, B. (2012). Sakubun sien to rejisutā [Writing assistance and register]. In K. Nishina, M. Kamada, H. Cao, T. Utashiro, & T. Muraoka (Eds.), *Nihongo gakusyūsien no kōtiku: gengokyōiku kōpasu sisutemu kaihatu [Constructing Japanese Language Learning: Language education, corpus and system development]* (3, pp. 275–287). Tokyo, Japan: Bonjinsha.

Hodošček, B. (2013). Kōpasu no shūshū, setsumei, janru; jikken to bunseki [Corpus collection, explanation and genre; Experiment and analysis]. In Y. Sunakawa (Ed.), *Kōza nihongo kōpasu [Japanese corpus textbook series]* (Chap. 5, Vol. 5, Vols. 8). Asakura Publishing Co., Ltd.

Hodošček, B., Abekawa, T., Bekeš, A., & Nishina, K. (2011). Assisting co-occurrence production in report writing: Evaluation of writing assistance tool Natsume. *Journal of Technical Japanese Education*, *13*, 33–40. doi:10.11448/jtje.13.33

Hodošček, B., Abekawa, T., Murota, M., & Nishina, K. (2012, August). Readability of example sentences in writing assistance tool Natsume. (pp. 1–4). 5th international conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J). Nagoya. Retrieved from http://2012castelj.kshinagawa.com/proceedings/Poster/Poster8_Bor%20Hodo%C5%A1%C4%8Dekodoscek.pdf

Hodošček, B. & Nishina, K. (2011a, March). Learning effect on academic Japanese expression usage with writing support system Natsume. (pp. 1–2). 13th forum of the society for technical japanese education. Tohoku University, Sendai, Japan. Retrieved from http://stje.kir.jp/download/13STJE_discussion.pdf

Hodošček, B. & Nishina, K. (2011b, August). On the treatment of register in writing assistance systems. (Vol. 2, pp. 522–523). International Conference on Japanese Language Education 2011. Tianjin, China.

Hodošček, B. & Nishina, K. (2012a, March). BCCWJ ni okeru shuttenjōhō to topikku oyobi rejisutā to no kankei [Comparison of metadata with topic and register in the BCCWJ]. In *Dai ikkai kōpasu nihongo waākushoppu yokōshū [Proceedings of the First Workshop on Japanese Corpus Linguistics]* (pp. 339–342). Dai ikkai nihongo kōpasu wākushoppu [First Workshop on Japanese Corpus Linguistics]. Tokyo, Japan.

Hodošček, B. & Nishina, K. (2012b). Japanese learning support systems: Hinoki project report. *Acta Linguistica Asiatica*, *2*(3) Lexicography of Japanese as a Second/Foreign Language (Part 2), 95–124. Retrieved from http://revije.ff.uni-lj.si/ala/article/view/221

Japan Foundation: survey report on Japanese language education abroad. (2009). Retrieved from https://www.jpf.go.jp/j/japanese/survey/result/dl/survey_2009/gaiyo2009.pdf

Johns, T. (1991). Should you be persuaded: two samples of data-driven learning materials. *English language research journal*, *4*, 1–16.

Johns, T. (2002). Data-driven learning: the perpetual challenge. *Language and Computers*, *42*(1), 107–117.

Joyce, T. (2011). The significance of the morphographic principle for the classification of writing systems. *Written Language & Literacy*, *14*(1), 58–81.

Joyce, T., Hodošček, B., & Nishina, K. (2012). Orthographic representation and variation within the Japanese writing system: Some corpus-based observations. *Written Language & Literacy*, *15*(2) Special Issue on Units of Language – Units of Writing, 254–278. doi:10.1075/wll.15.2.01rob

Kabashima, T. & Jugaku, A. (1965). *Buntai no kagaku*. Sogeisha.

Kaplan, D., Iida, R., Nishina, K., & Tokunaga, T. (2012). Slate - a tool for creating and maintaining annotated corpora. *Journal for Language Technology and Computational Linguistics*, *26*(2), 89–101. Retrieved from http://www.cl.cs.titech.ac.jp/publication/673.pdf

Karatzoglou, A., Smola, A., Hornik, K., & Zeileis, A. (2004). Kernlab – an S4 package for kernel methods in R. *Journal of Statistical Software*, *11*(9), 1–20. Retrieved from http://www.jstatsoft.org/v11/i09/

Kess, J. F. & Miyamoto, T. (1999). *The Japanese mental lexicon: Psycholinguistic studies of kana and kanji processing*. John Benjamins Publishing.

Kilgarriff, A., Rychly, P., Smrz, P., & Tugwell, D. (2004, July). The Sketch Engine. In *Proceedings of EURALEX* (pp. 105–116). Lorient, France.

Knuth, D. E. (2007). Computer programming as an art. In *ACM Turing award lectures* (p. 1974). ACM.

Kuhn, M. (2008). Building predictive models in R using the caret package. *Journal of Statistical Software*, *28*(5), 1–26. Retrieved from http://www.jstatsoft.org/v28/i05

Kuhn, M., Weston, S., Coulter, N., & Quinlan, R. (2013). *C50: C5.0 Decision Trees and Rule-Based Models*. R package version 0.1.0-15. Retrieved from http://CRAN.R-project.org/package=C50

Labov, W. (1994). *Principles of linguistic change, internal factors*. Blackwell Publishers.

Leacock, C., Gamon, M., & Brockett, C. (2009). User input and interactions on Microsoft Research ESL Assistant. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 73–81). Association for Computational Linguistics.

Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language Learning & Technology*, *5*(3), 37–72. Retrieved from http://llt.msu.edu/vol5num3/pdf/lee.pdf

L'Haire, S. & Faltin, A. (2003). Error diagnosis in the FreeText project. *The Computer Assisted Language Instruction Consortium (CALICO) Journal*, *20*(3), 481–495. Retrieved from https://calico.org/a-290-Error%20Diagnosis%20in%20the%20FreeText%20Project.html

Liaw, A. & Wiener, M. (2002). Classification and Regression by randomForest. *R News*, *2*(3), 18–22. Retrieved from http://CRAN.R-project.org/doc/Rnews/

Maekawa, K. (2007a, March 1–03). Design of a balanced corpus of contemporary written Japanese. In *Proceedings of the symposium on Large-scale Knowledge Resources (LKR2007)* (pp. 55–58). Tokyo Institute of Technology. Tokyo, Japan.

Maekawa, K. (2007b). KOTONOHA and BCCWJ: Development of a Balanced Corpus of Contemporary Written Japanese. In *Proceedings of the first international conference on Korean language, literature, and culture* (Vol. 2, pp. 158–177). Corpora and Language Research. Seoul.

Maekawa, K. (2011, October). Linguistics-oriented language resource development at the National Institute for Japanese Language and Linguistics. In *2011 International Conference on Speech Database and Assessments (Oriental COCOSDA)* (pp. 1–6). doi:10.1109/ICSDA.2011.6085971

Martin, J. (2009). Discourse studies. In M. Halliday & J. J. Webster (Eds.), *Continuum Companion to Systemic Functional Linguistics*. Continuum International Publishing Group.

Maruyama, T., Yamazaki, M., Kashino, W., Sano, M., Akimoto, M., Inamasu, S., … Oyauchi, Y. (2010). Outline of sampling method in the Balanced Corpus of Contemporary Written Japanese (4): Corpus design and the result of sampling. In *Tokutei ryōiki `nihongo kōpasu' hēsē 21 nendo kōkai wākushoppu (kenkyū sēka hōkokukai) yokōshū* (pp. 37–46).

Miller, R. A. (1967). *The Japanese language*. Charles E. Tuttle.

Mizumoto, T. & Komachi, M. (2012, February). Robust NLP for Real-world Data: 3. Why is Japanese so Hard to Learn?—A Preliminary Investigation on Realistic Japanese Learners' Corpus and Application of Natural Language Processing to Japanese Language Learning and Education—. *IPSJ Magazine*, *53*(3), 217–223.

Mudge, R. (2010). The design of a proofreading software service. In *Proceedings of the NAACL HLT 2010 workshop on computational linguistics and writing: writing processes and authoring aids* (pp. 24–32). CL&W '10. Los Angeles, California: Association for Computational Linguistics. Retrieved from http://dl.acm.org/citation.cfm?id=1860657.1860661

Muraoka, T., Chinami, K., & Nishina, K. (2009). Senmon bunshō sakusē shien hōhō no kaihatsu ni mukete: sukīma kēsē o chūshin ni [Towards the development of composition assistance methods for technical japananese writing: Focusing on schema construction]. *Journal of Technical Japanese Education*, (11), 23–30.

Naber, D. (2003). *A rule-based style and grammar checker* (Master's thesis, Bielefeld University).

Nation, I. (2001). *Learning vocabulary in another language*. Cambridge Applied Linguistics. Cambridge University Press. Retrieved from http://books.google.co.jp/books?id=sKqx8k8gYTkC

NINJAL [National Institute for Japanese Language and Literature]. (2011). Tokuteiryōiki kenkyū nihongo kōpasu kenkyū seika hōkoku [Priority-Area Research "Japanese Corpus": Research Report] [DVD media containing UniDic and Comainu]. Tokyo: General Headquarters, Priority-Area Research "Japanese Corpus".

Nishina, K., Doi, M., & Takano, T. (2007). *An Introduction to Technical Japanese*. 3anet.

Nishina, K., Okumura, M., Abekawa, T., Yagi, Y., Bilac, S., & Fu, L. (2004, March 8–09). Asunaro CALL system: combining multilingual with multimedia. In *International symposium on Large-scale Knowledge Resources LKR 2004* (pp. 69–72). Tokyo Institute of Technology. Tokyo, Japan.

Oono, H. & Inazumi, H. (2011, March). Support tool of Japanese document proofreading and polish : Tomarigi : overview of efforts to support and labor-saving, for the labor of correction. *Research report of JSET Conferences*, *2011*(1), 325–332. Retrieved from http://ci.nii.ac.jp/naid/10029781745/en/

Oshima, Y. (2009). Analysis of the discussions of case studies and historical document-based studies in social science. *Journal of Technical Japanese Education*, (11), 15–22. Retrieved from http://ci.nii.ac.jp/naid/40017331801/en/

Pardeshi, P., Imai, S., Kiryu, K., Lee, S., Akasegawa, S., & Imamura, Y. (2012). Compilation of Japanese Basic Verb Usage Handbook for JFL Learners: A Project Report. *Acta Linguistica Asiatica*, *2*(2), 37–64.

Pawley, A. & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In *Language and Communication* (pp. 191–226). London: Longman.

Quinlan, J. R. (1993). Combining instance-based and model-based learning. In *ICML* (pp. 236–243).

R Core Team. (2013). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria. Retrieved from http://www.R-project.org/

Römer, U. (2006). Pedagogical applications of corpora: some reflections on the current scope and a wish list for future developments. *Zeitschrift für Anglistik und Amerikanistik*, *54*(2), 121–134.

Rousseeuw, P. J., Ruts, I., & Tukey, J. W. (1999). The bagplot: a bivariate boxplot. *The American Statistician*, *53*(4), 382–387.

Sato, S., Matsuyoshi, S., & Kondoh, Y. (2008). Automatic assessment of Japanese text readability based on a textbook corpus. In *LREC-08* (pp. 654–660).

Saussure, F. d. (1959). *Course in general linguistics* (W. Baskin, Trans.). New York: Philosophical Library. Retrieved from http://www.archive.org/details/courseingenerall00saus

Schramm, W. (1955). The process and effects of mass communication, 3–10, 13, 17.

Schramm, W. (1997). How communication works. *Mass media & society*, 51–66. Repr. of. The process and effects of mass communication. (1997), 3–10, 13, 17

Shibasaki, H. & Hara, S. (2010). 12 gakunen o nan'ishakudo to suru nihongo rīdabiritī hanteishiki [Japanese readability formula for discriminating k-12 grade level texts]. *Mathematical Linguistics*, *27*(6), 215–232.

Shibatani, M. (1990). *The languages of Japan*. Cambridge Language Surveys. Cambridge University Press. Retrieved from http://books.google.co.jp/books?id=sD-MFTUiPYgC

Shimodaira, H. (2004). Approximately unbiased tests of regions using multistep-multiscale bootstrap resampling. *Annals of Statistics*, *32*, 2616–2641.

Sinclair, J. M. (1987). *Looking up: An account of the COBUILD project in lexical computing and the development of the Collins COBUILD English language dictionary*. HarperCollins Publishers Limited.

Smola, A. & Narayanamurthy, S. (2010). An architecture for parallel topic models. *Proceedings of the VLDB Endowment*, *3*(1-2), 703–710.

Srdanović, I., Hodošček, B., Bekeš, A., & Nishina, K. (2009). Extraction of suppositional adverb and clause-final modality form distant collocations using a web corpus and corpus query system and its application to Japanese language learning. *Journal of Natural Language Processing*, *16*(4), 29–46.

Srdanović, I., Suchomel, V., Ogiso, T., & Kilgarriff, A. (2013, March). hyaku oku go koopasu o mochiita nihongo no goi - bunpou jouhou no puro-

fairingu [Japanese Language Lexical and Grammatical Profiling Using the Web Corpus JpTenTen]. In *Dai san kai kōpasu nihongo waākushoppu yokōshū [Proceedings of the 3rd Workshop on Japanese Corpus Linguistics]* (pp. 229–238). Tokyo, Japan.

Srdanović-Erjavec, I., Erjavec, T., & Kilgarriff, A. (2008). A web corpus and word sketches for Japanese. *Information and Media Technologies*, *3*(3), 529–551.

Swales, J. (1990). *Genre analysis: English in academic and research settings*. Cambridge Applied Linguistics. Cambridge University Press.

Tateisi, Y., Ono, Y., & Yamada, H. (1988). A computer readability formula of Japanese texts for machine scoring. In *Proceedings of the 12th Conference on Computational Linguistics* (Vol. 2, pp. 649–654).

The National Language Research Institute. (1981). *Introduction to the linguistic atlas of Japan: Methodology*. Tokyo, Japan: Ōkurashō Insatsu. Retrieved from http://www6.kokken.go.jp/laj_map/

Tranter, N. (2008). Nonconventional script choice in Japan. *International Journal of the Sociology of Language*, *2008*(192), 133–151.

Venables, W. N. & Ripley, B. D. (2002). *Modern Applied Statistics with S* (Fourth). ISBN 0-387-95457-0. New York: Springer. Retrieved from http://www.stats.ox.ac.uk/pub/MASS4

Wolf, P. & Bielefeld, U. (2010). Aplpack: another plot package: stem. leaf, bagplot, faces, spin3r, and some slider functions. *R package version*, *1*(3).

Yagi, Y., Hodošček, B., & Nishina, K. (2012, March). BCCWJ to gakushūsha sakubun kōpasu o riyōshita nihongo sakubun shien [Japanese writing assistance using the BCCWJ and a learner corpus]. In *Dai ikkai kōpasu nihongogaku wākushoppu yokōshū [Proceedings of the First Workshop on Japanese Corpus Linguistics]*. Dai ikkai nihongo kōpasu wākushoppu [First Workshop on Japanese Corpus Linguistics]. Tokyo, Japan.

Yagi, Y. & Suzuki, T. (2012). Gakushūsha sakubun kōpasu no kōchiku to goyō no bunseki [Construction of learner corpus and error analysis]. In K. Nishina, M. Kamada, H. Cao, T. Utashiro, & T. Muraoka (Eds.), *Nihongo gakushūshien no kōchiku: gengokyōiku kōpasu shisutemu kaihatsu [Constructing Japanese Language Learning: Language education, corpus and system development]* (3, pp. 249–274). Tokyo, Japan: Bonjinsha.

ONLINE REFERENCES

After the Deadline - Spell, Style, and Grammar Checker for WordPress, Firefox, TinyMCE, jQuery, and CKEditor. (2013). Retrieved May 30, 2013, from http://www.afterthedeadline.com/

Agency For Cultural Affairs. (2011). 内訳図表―日本語学習者数の推移 [Tabulated data on Japanese language learner trends]. Retrieved May 30, 2013, from http://www.jasso.go.jp/statistics/intl_student/data12_e.html

Evert, S. (2004). Association measures. Retrieved from http://www.collocations.de/AM/index.html

Google. (2013). Japanese - Google Scholar Metrics. Retrieved July 20, 2013, from http://scholar.google.com/citations?view_op=top_venues&hl=en&vq=ja

Isahara, H., Bond, F., Kanzaki, K., Uchimoto, K., Kuroda, K., Kuribayashi, T., … Torisawa, K. (2012). Japanese WordNet. Retrieved May 30, 2013, from http://nlpwww.nict.go.jp/wn-ja/index.en.html

Japan Student Services Organization. (2013, February). International students in Japan 2012. Retrieved May 30, 2013, from http://www.jasso.go.jp/statistics/intl_student/data12_e.html

Kawamura, Y., Kitamura, T., & Hobara, R. (2013). Japanese language reading tutorial system. Retrieved May 30, 2013, from http://language.tiu.ac.jp/

Kudo, T. (2013). MeCab: Yet Another Japanese Dependency Structure Analyzer. Retrieved May 30, 2013, from https://code.google.com/p/mecab/

Kudo, T. & Matsumoto, Y. (2013). CaboCha: Yet Another Japanese Dependency Structure Analyzer. Retrieved May 30, 2013, from https://code.google.com/p/cabocha/

Oosaki, H. (2006). Tips for technical writing. Retrieved May 30, 2013, from http://www.ispl.jp/~oosaki/research/tips-jcorrect/

Sketch Engine: SketchEngine. (2013). Retrieved May 30, 2013, from http://www.sketchengine.co.uk/

UniDic Project Top Page. (2013). Retrieved May 30, 2013, from http://en.sourceforge.jp/projects/unidic/