JAPANESE LEARNING SUPPORT SYSTEMS: HINOKI PROJECT REPORT

Bor HODOŠČEK
Tokyo Institute of Technology,
Graduate School of Decision Science and Technology,
Department of Human System Science
hodoscek.b.aa@m.titech.ac.jp

Kikuko NISHINA
Professor Emeritus,
Tokyo Institute of Technology
knishina@m06.itscom.net

Abstract

In this report, we introduce the Hinoki project, which set out to develop web-based Computer-Assisted Language Learning (CALL) systems for Japanese language learners more than a decade ago. Utilizing Natural Language Processing technologies and other linguistic resources, the project has come to encompass three systems, two corpora and many other resources. Beginning with the reading assistance system Asunaro, we describe the construction of Asunaro's multilingual dictionary and its dependency grammar-based approach to reading assistance. The second system, Natsume, is a writing assistance system that uses large-scale corpora to provide an easy to use collocation search feature that is interesting for its inclusion of the concept of genre. The final system, Nutmeg, is an extension of Natsume and the Natane learner corpus. It provides automatic correction of learners errors in compositions by using Natsume for its large corpus and genre-aware collocation data and Natane for its data on learner errors.

Keywords

CALL; reading assistance system; writing assistance system; scientific and technical Japanese corpus; learner corpus; genre

Izvleček

V poročilu predstavljamo projekt Hinoki, ki je bil zastavljen pred več kot desetimi leti za izdelavo spletnih sistemov za računalniško podprto učenje japonščine kot tujega jezika. Z uporabo jezikovnih tehnologij in drugih jezikovnih virov so bili v okviru projekta razviti trije sistemi, dva korpusa in veliko drugih virov. V nadaljevanju predstavljamo sistem Asunaro za podporo branju, izgradnjo njegovega večjezičnega slovarja in pristop k podpori branju, ki sloni na odvisnostni slovnici; sistem za podporo pisanju Natsume s preprostim vmesnikom za iskanje žanrsko določenih kolokacij v obsežnih korpusih; ter sistem Nutmeg za samodejno popravljanje napak. Nutmeg je nadgradnja sistema Natsume in učnega korpusa Natane, ponuja samodejno

popravljanje napak med samim pisanjem z uporabo žanrsko določenih kolokacijskih informacij iz obsežnih korpusov preko sistema Natsume in informacij o napakah piscev, ki se učijo japonščine kot tujega jezika, iz korpusa Natane.

Ključne besede

računalniško podprto učenje jezika; sistem za podporo branju; sistem za podporo pisanju; korpus znanstvene in tehnične japonščine; učni korpus; žanr

1. Preface

According to a 2009 report from the Japan Foundation, there are over three and a half million people learning Japanese outside Japan.[1] Fortunately, access to good general 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 unfortunately not improved as much.

The Japan Student Services Organization (JASSO) 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 reports that there are 40,799 international students studying Japanese in Japan.[2] For these students who are pursuing specialized study at institutions of higher education in Japan, the following are just some of the skills they will have to master:

- read textbooks
- write reports and papers
- listen to lectures and take notes
- present at conferences or seminars

Because it is hard to tailor the Japanese language class to meet the specialized language needs of each learner's field of specialization, an alternative is needed. One way of approaching this 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 material from their field of specialization. The construction of such self-learning, individualized learning systems

[1] Detailed statistics are available in the Japan Foundation's 2009 "Survey Report on Japanese Language Education Abroad", available at https://www.jpf.go.jp/j/japanese/survey/result/dl/survey_2009/gaiyo2009.pdf

[2] While JASSO provides numbers for international students in their 2011 report available at http://www.jasso.go.jp/statistics/intl_student/data11_e.html, they do not provide information on the number of students who are required to take Japanese classes. MEXT offers an independent report with slightly different numbers, available at http://www.bunka.go.jp/kokugo_nihongo/jittaichousa/h23/gaikoku_6_03.html, that does contain the number of Japanese language learners.

has been the goal of the Hinoki project. The following report describes three systems and several linguistic resources, the results of pursuing this goal for over a decade.

1.1 Report Overview

After providing an overview of the Hinoki project, Chapter 2 describes the linguistic resources in use by the project. Chapter 3 introduces the Asunaro reading support system, which features courseware designed for science and engineering students, as well as a multilingual dictionary that includes several commonly underrepresented Asian languages. Chapter 4 describes Natsume, a writing assistance system that is backed by large-scale corpora and provides an easy to use search interface for collocations. Chapter 5 introduces the search system for our Natane learner corpus, which has applications to second language acquisition research and machine-learning applications for automatic learner error detection and correction. Chapter 6 introduces Nutmeg, an automatic error correction system for learner's writing. Finally, Chapter 7 concludes this report by offering a summary and our perspectives for future work.

2. Linguistic Resources

The Hinoki project relies heavily on linguistic resources, though it is also a producer of such. Linguistic resources used in the project are native and learner corpora, as well as dictionaries. To meet the goals of the project, some linguistic resources had to be developed: multi-lingual dictionaries, purpose-specific corpora, as well as learner corpora.

In the earlier systems, emphasis was put on native resources, as they enable a Data-Driven Learning approach to learning Japanese.

However, to really know where and why learners make mistakes, a learner corpus is also essential and is where more recent efforts have been focused on.

2.1 Native Resources

As part of the *Nihongo kōpasu* ("Japanese Corpus") project led by the National Institute for Japanese Language and Linguistics (NINJAL) for 4 years, the main goal of our group was to explore the ways in which the project's new Balanced Corpus of Contemporary Written Japanese (BCCWJ) could be applied to Japanese language education. As we are focused on finding ways to assist Japanese language learners in writing academic reports and papers, it was necessary to compile another corpus containing this genre, in addition to using the BCCWJ. For several other reasons explained below, the Japanese version of Wikipedia was also used.

2.1.1 BCCWJ

The National Institute for Japanese Language and Linguistics (NINJAL) created the Balanced Corpus of Contemporary Written Japanese (BCCWJ) in the span of five years between 2006 and the end of 2011 (Maekawa, 2007b, 2007a, 2012). The objective of the project was to compile a tagged corpus of contemporary written Japanese that had sufficient scale and coverage of sub-varieties of written language to offer a representative sample of Japanese written language. Such a language resource had not previously existed for Japanese, and its creation was seen as important for the future development of any research with a need for representative Japanese language data, including official government language policy.

The BCCWJ consists of the Publication, Library and Special-Purposes sub-corpora, each of them accounting for roughly one-third the size of the BCCWJ. The Publication sub-corpus includes books, magazines, and newspapers and is sampled from all published material in Japan between 2001 and 2005. The Library sub-corpus includes books sampled from several library holdings within the Greater Tokyo Metropolitan area. The Special-Purposes sub-corpus differs from the other two in that it should not be considered a representative sample of written Japanese, but rather serve as useful comparison material for the others.

2.1.2 Scientific and Technical Japanese Corpus (STJC)

Unfortunately, the data needs of providing writing and reading assistance in an academic context are not fully satisfied by the BCCWJ. While some sub-corpora are close in subject matter (topic) and writing style (register), the lack of inclusion of genuine research papers from academic journals precludes their ability to serve as a representative sample of written science and engineering discourse. It was thus necessary to build a new corpus that contained a representative and authentic sample of academic writing. The new corpus was named the Scientific and Technical Japanese Corpus (STJC), and consists of papers from several scientific and technical journals written in Japanese. The following criteria were used when choosing which journals to collect papers from:

1. The journal must specialize in some scientific or engineering field.
2. The journal is published by a society with at least a thousand members.
3. The journal has reasonable review standards.
4. The journal allowed us to use the text from the papers as example sentences in our system.

Currently, papers are included 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.

2.1.3 Wikipedia

The decision to include the Japanese version of Wikipedia[3] was made for several reasons. For many tasks, the quantity of data provided by the BCCWJ is sufficient (Maekawa, 2011). However, due to the nature of the data used in Natsume, which includes triplet combinations of nouns, case particles and verbs, the amount of extractable data for any but the most common expressions quickly becomes insufficient. Additionally, other NLP technologies deployed in the Natsume system, such as getassoc[4], are more precise at scales of data in the range of Wikipedia. Another requirement of the project is that text data from corpora should be legally available to be displayed online. The permissive license of Wikipedia allows all us to show all sentences as example sentences in Natsume.

One unfortunate side effect of including Wikipedia is that for many less frequent collocations, the only information on them is available in Wikipedia, making any genre comparisons impossible. Another demerit of including Wikipedia is the inclusion of grammatical mistakes and comparatively long sentences, which average at around 51 characters compared with an average of 35 for newspapers (see Table 1).

Table 1: Character counts and average sentence length for all corpora.

Corpus	Subcorpora	Characters	Average sentence length
STJC		6,108,143	58.46
Wikipedia		372,901,202	51.10
BCCWJ	Books	53,801,124	37.59
	Yahoo! Q&A	9,763,298	30.69
	Diet minutes	8,712,108	75.06
	Textbooks	1,818,571	28.44
	White papers	8,443,965	58.25
	Yahoo! Blogs	5,246,121	23.26
	Magazines	455,634	31.41
	Newspapers	1,188,355	34.90
Total		**468,438,521**	**48,04**

[3] Currently a snapshot from 2008. Wikipedia data dumps are available from http://dumps.wikimedia.org/.

[4] Available from http://getassoc.cs.nii.ac.jp/.

2.2 Natane – Learner Corpus

2.2.1 Introduction

Natane is a Japanese language learner corpus annotated with learner errors. The main benefit of learner corpora in the context of writing assistance, when compared to native corpora, is that they enable insights into the kinds of errors learners make. For example, in the case of Natane, comparing learner error tendencies based on their first language might guide customizations to lesson plans based on the learner's first language.

Compared to native corpora containing the writings of native speakers, learner corpora are often smaller in size and variety. This is due to the difficulty of obtaining learner writing, which in most cases is elicited for the construction of the corpus and not collected from readily-available sources as in the construction of the BCCWJ. Another common differentiator is the inclusion of error annotations and background information on the learners who produced the material used.

The end goal of the construction of this corpus is the construction of well-formed and sufficient machine-learnable data for automatic writing error correction. It should be noted that while relatively simpler things like the construction of a spellchecker, co-occurrence checker, or writing style checker are possible, features that hinge on an understanding of semantics and discourse are hard to make practical even in state-of-the-art NLP systems.

As an ongoing joint project with several Japanese language teachers, the collection and annotation of the corpus initially proceeded along the following stages:

1. Collection of learner essays and their transcription.
2. Pilot annotation of learner errors using Excel (Cao & Nishina, 2010; Cao, Kuroda, Yagi, & Nishina, 2010).
3. Analysis of pilot annotation and definition of final error classification framework (Cao, Kuroda, Yagi, & Nishina, 2011; Cao, Yagi, & Nishina, 2012).
4. Use of the multipurpose annotation tool Slate for error tagging.

2.2.2 Collection

The essays were collected from undergraduate and graduate students as well as students attending Japanese language schools. All essays were written to a specific topic, though not all topics are the same. Each learner's age, nationality, university level, first language, major and Japanese language learning experience, as well as other background information, were recorded with the essay. Additionally, learners signed a waiver authorizing the anonymized usage of their essay in our project.

Although more than 5000 sentences have been collected, currently only around 3500 have been annotated[5]. 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 is biased towards Mandarin Chinese speakers, who account for more than half of learners and essays. The remaining languages are predominantly from Asia.

Table 2: Distribution of essays by first language.

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**

2.2.3 Pilot Annotation

While error classification frameworks for languages such as English and French already exist (Díaz-Negrillo & Fernández-Domínguez, 2006; Granger, 2003; L'Haire & Faltin, 2003), there were no preexisting comprehensive error annotation scheme or descriptive framework for Japanese language learner errors. Because of the lack of such a framework, the project decided to construct one itself, drawing from previous research as well as the annotator's teaching experience (Cao & Nishina, 2010). During the pilot annotation process, it became clear that there were two kinds of error annotations. The first were ordinary, unambiguous errors and the second kind were errors where the annotator felt the particular language usage was unnatural. Ordinary errors include deviations from standard orthography, syntactic function (voice, tense, aspect, modality), conjugation, and subject-predicate incongruity. They are typically easy to annotate and occur frequently. Unnatural errors include word choice, addition or omission of text units (phrase, paragraph, etc.), and are typically less frequent and harder to annotate, leading to lower agreement between annotators.

[5] An up-to-date breakdown of data included in Natane, including the nationalities of learners, is available at http://hinoki.ryu.titech.ac.jp/natane/stats.

2.2.4 Error Classification Framework

The feedback gained from the pilot annotation process was crucial for refinements in the error classification framework (Cao, Kuroda, Yagi, & Nishina, 2011). The resulting error annotation framework is hierarchical, able to take into account different viewpoints regarding learner errors, as well as enable the systematic annotation of such errors (Yagi & Suzuki, 2012).

The hierarchy consists of at most four levels, with higher levels corresponding with courser, more abstract categories, and branches out in three principal dimensions:

1. Error level – the linguistic level of the error (i.e. phoneme, word, phrase, ..., discourse; the word tag is further classified into word classes like noun, verb, etc.)
2. Error category – type and form of error
   - type: addition, omission, word order, deviation from standard orthography, etc.
   - form: conjunction, conjugation, collocation, (Japanese letter) script
3. Error source – reason or background for error (i.e. annotator's subjective opinion on source of error: register and style mismatch, coherence, first language interference, etc.)

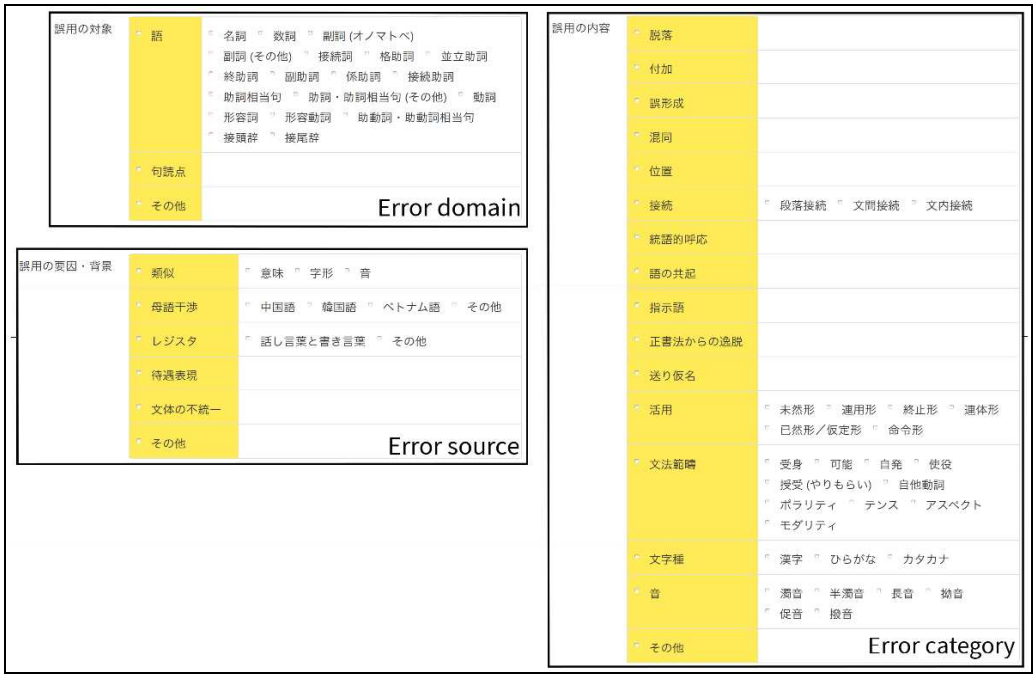

Figure 1: The hierarchical error classification framework used in Natane

2.2.5 Error Annotation Process with Slate

After the error classification framework was decided on, the choice had to be made between continuing to use Excel to annotate the corpus or finding another solution. Though Excel's free-form nature served the formative stage of the annotation process, significant drawbacks related to its ad-hoc usage became clear. The choice was then made to use the web browser-based Slate corpus annotation and management system[6], as it offers the following advantages over Excel: higher data integrity and greater data diversity (Kaplan, Iida, Nishina, & Tokunaga, 2012). Slate decreases the chance for inconsistent annotation by eliminating the chance for errors with respect to formatting differences between annotators and misplacement of annotations into the wrong table cell, among other problems. Using Slate also increases the diversity of possible annotations, by enabling more than one annotation per segment (sentence) as well as annotations that overlap or span multiple sentences. Previously the format of the Excel table limited the amount of possible error annotations to one per sentence. Slate also provides an overhead view of the hierarchical error classification framework that - coupled with an interface that allows the user to see all annotations at a glance - enables efficient and speedy annotation.

As there was considerable data included in the existing Excel tables, it was not re-annotated but rather converted for inclusion into Slate. All new annotations are being recorded using Slate. Three teachers specializing in Japanese language education at different universities separately annotated all essays using the Slate corpus annotation and management system.

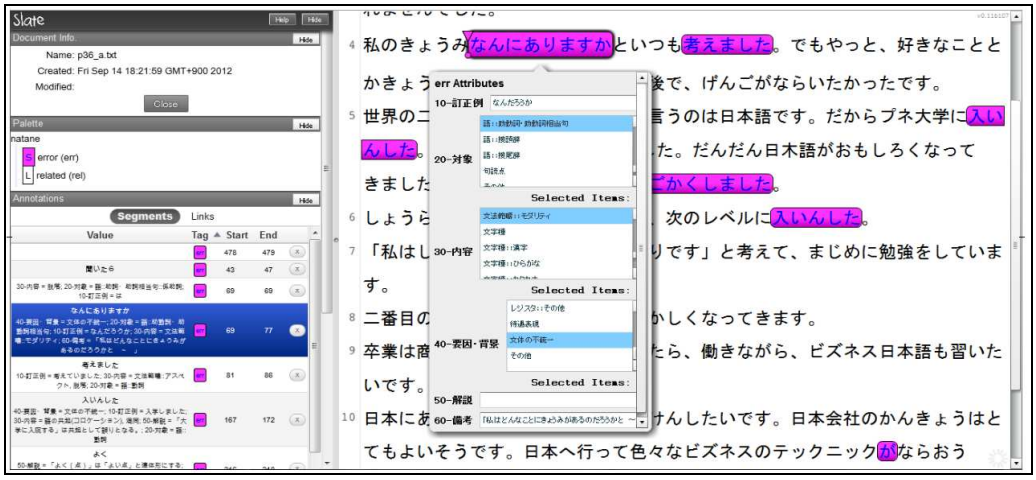

Figure 2: An example of composition errors annotated with Slate; marked areas represent errors, with the left pane providing detailed information including all error annotations.

[6] More information is available at Slate's homepage: http://www.cl.cs.titech.ac.jp/slate/.

2.3 Conclusion

The Hinoki project depends on the existence of many large-scale corpora, most of which are already available to the research community. For more specialized needs, such as the inclusion of a representative sample of scientific and technical Japanese, no corpora existed, so one had to be constructed. The available Japanese language learner corpora are still few, although recent developments have increased the number available: Learner's Language Corpus of Japanese[7], Teramura corpus[8], NINJAL's learners corpus[9], JC Corpus[10] are just some of the corpora available now. The existing major differences between Natane and these learner corpora is that they are more focused on the annotation of grammatical errors and thus have a less comprehensive error classification framework than the one used in Natane.

Though not mentioned in this chapter, without the availability of high-quality Natural Language Processing tools for Japanese, it would be hard to impossible to make use of much of these linguistic resources. The specific tools used in each system are detailed in the explanations of each system separately.

3. Asunaro: Multilingual Reading Assistance System

3.1 Introduction

The first system developed under the Hinoki project was the Asunaro multilingual reading assistance system (Nishina, Okumura, Yagi, et al., 2002; Nishina, Okumura, Abekawa, et al., 2004). Development of Asunaro began in 1999 and the system was first released online in 2002.

At its inception, Asunaro was unique in that it integrated a multilingual reading and learning environment into one online system accessible to anyone with an Internet connection. At the time, most systems targeted English language learners, while Asunaro incorporates several Asian languages. This was important because the number of international students from neighboring Asian countries studying at universities in Japan is greater than that of students from English-speaking countries.

The main goal of the system was to help Japanese learners read and understand academic material in Japanese. The main target of the system is Japanese language learners enrolled in Japanese universities majoring in the fields of science and engineering. Many of them are expected to be able to read academic papers and textbooks in their field, but it is often difficult to provide for their specialized learning

[7] http://cblle.tufs.ac.jp/llc/ja/

[8] http://teramuradb.ninjal.ac.jp/

[9] http://jpforlife.jp/taiyakudb.html

[10] http://www34.atwiki.jp/jccorpus/pages/21.html

needs in university Japanese language classes. The use of Asunaro was seen as a way to enable personalized learning for those learners.

3.2 Main Features

Users accessing the Asunaro system are presented with the main screen containing a text box into which they can paste or directly enter Japanese language text for analysis. The main screen is split into three areas consisting of the user input area in the top left, the translation and example sentence area in the top right, and a detailed word and phrase view of single sentences at the bottom.

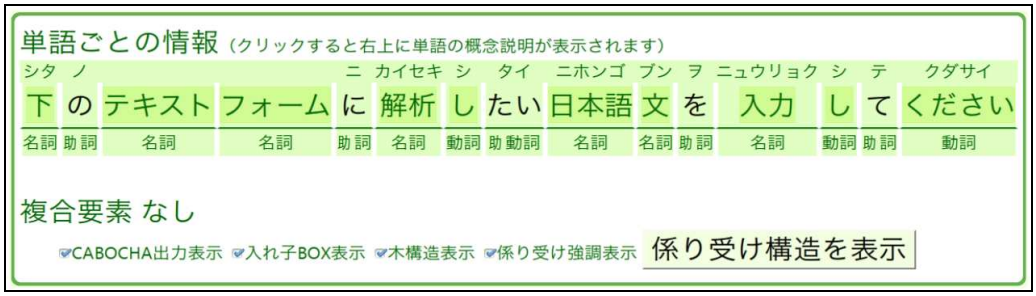

Figure 3: Bottom area containing morphologically analyzed user input with readings and word class information provided by MeCab (Kudo, 2012).

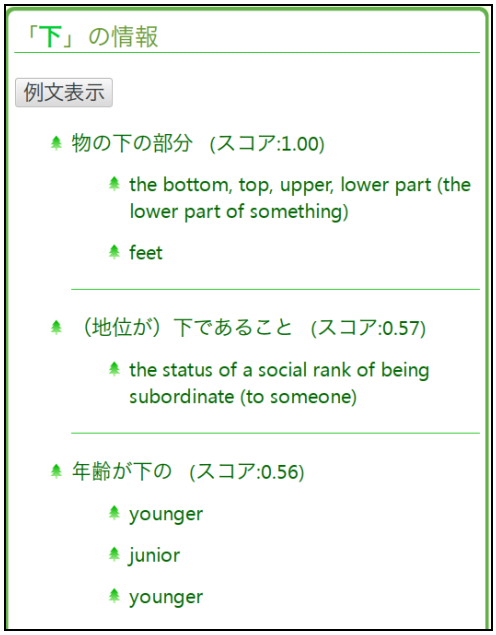

Figure 4: Top right area containing translations and example sentences of user selected words or phrases.

Users click words or phrases[11] in the bottom area to update translations in the top right area. Translations appear in order of importance, based on the application of meaning disambiguation using the surrounding word context.

Finally, clicking on the arrow at the beginning of each sentence takes the user to the secondary screen where they can see the dependency structure of the sentence.

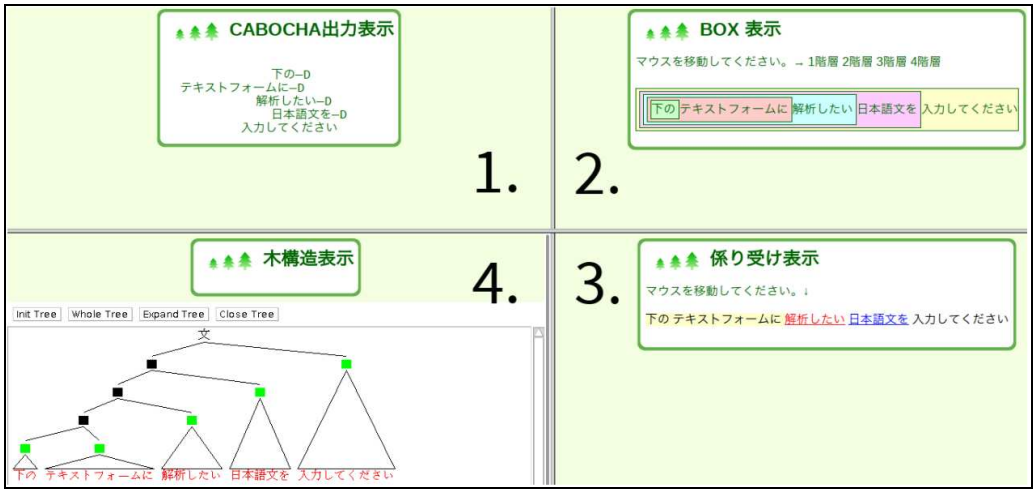

Figure 5: Different views of dependency structure in secondary screen. (Clockwise from top left)
1. raw output from the Japanese dependency parser CaboCha (Kudo & Matsumoto, 2012),
2. embedded dependency structure ("ireko") display,
3. mouse-over dependency link-like display, and
4. tree structure representation of sentence.

3.3 Courseware

However, the usage outlined above is in many ways too difficult for non-advanced learners. For beginning-to-intermediate learners who are studying in the fields of science and engineering, the provided courseware is more appropriate. Learning a language through reading is best when the material read is learner-level appropriate. Asunaro makes use of a textbook (Nishina, 2001) which is written specifically for intermediate level undergraduate science and engineering students. The main goal of the courseware is to help science and engineering students achieve proficiency in technical communication to be able to read papers and discuss research in seminars. All courseware in Asunaro was checked for parsing mistakes and manually corrected.

[11] Idioms and phrases like *te wo tunagu* and *kao ga hiroi* are automatically recognized as such and also marked as phrases.

Additionally, the courseware contains an audio playback feature so users can listen to the courseware material while learning to read it.

3.4 Multilingual Dictionary

Although electronic Japanese-English dictionaries have been available since the beginning of the 1990s, for many Asian languages such as Malay, Thai or even Chinese, no electronic dictionary was available at the time of Asunaro's inception. As more than half of all international students in Japan come from other Asian countries, support of languages other than English was seen as a high priority. The EDR Electronic Dictionary is used for its translations between English and Japanese, as well as its concept ID, which links every Japanese word to a concept.[12] Enabling translations of Japanese words into languages other than English required the construction of a new multilingual dictionary that would map words from the target language to EDR's concept ID. This differed from many similar systems at the time that used English as the intermediary language. Excluding the Japanese and English entries from the EDR dictionary, the multilingual dictionary contains around 25,000 entries for Chinese, and around 5,000 each for Thai, Indonesian and Malay.

Another unique feature of the system was that it provided a common language-independent framework for handling compound expressions. This is important as compounds and phrasal units are language-dependent and must be handled on a per-language basis. For Japanese, phrasal units and compounds are detected using CaboCha and the EDR electronic dictionary.

3.5 Related Work

Reading Tutor, which is a reading assistance system widely used in Japan and abroad, also contains a multilingual dictionary (Kawamura, Kitamura, & Hobara, 2012, 2000). Additionally, Reading Tutor's "Kyozai Banku" (Kawamura & Kitamura, 2001) is a similar effort to the courseware feature in Asunaro to provide leveled reading material.

Rikai.com is a popular website that provides hiragana readings and English translations of online text[13]

3.6 Conclusion

Asunaro was constructed to assist students from the fields of science and engineering to read and understand technical Japanese. For beginning- to intermediate-

[12] More information on the EDR Electronic Dictionary is available at https://www2.nict.go.jp/out-promotion/techtransfer/EDR/index.html.

[13] Rikai.com is accesible from http://www.rikai.com.

level learners, it includes courseware aimed at assisting them to eventually be able to read authentic texts from their field. For advanced learners, the copy and paste nature of the system allows them to focus on learning just the sentences they do not yet fully understand. It uses the EDR Electronic Dictionary as a basis for constructing a larger multilingual dictionary, presenting learners with glosses into their native language: English, Chinese, Malay, Thai and Indonesian.

4. Natsume: Writing Assistance System

4.1 Introduction

Natsume is an online writing assistance system that began operating in 2009 (Hodošček, in press, 2012; Abekawa, Hodošček, & Nishina, 2011). The initial focus during the development of Natsume was to enable users to not just be able to search for collocations, but also be able to convince themselves of the correct usage of a collocation in several ways. Thus, Natsume was to not just provide raw collocation information, but was to enable users to look for similar collocations and compare collocational tendencies between different genres.

While Asunaro assists international students in reading, Natsume focuses on assisting them in writing technical Japanese. For example, writing reports or papers at universities can be hard if the students cannot differentiate between what words or expressions are spoken and what are written Japanese. As a study and writing aid, the use of conventional (non-corpus-based) electronic dictionaries is prevalent among international students. However, these dictionaries seldom contain information on a word's usage with respect to written and spoken language. Natsume, by virtue of having access to corpora from various genres, contains information that can be used to determine if a word is appropriate for spoken or 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 frequently co-occurring patterns of words. These frequently co-occurring patterns of words are called collocations and are important because they offer more contextual information about a word than what is found in conventional dictionaries. Moreover, knowledge of collocations has been shown to be essential to achieving high second language proficiency (Pawley & Syder, 1983).

Users can use the system to find collocations of a word, check the correct use of a word or collocation by looking at example sentences, and compare observed frequencies in various genres. This follows the philosophy of data-driven language learning by giving users access to authentic information which they can then use as the basis for any decisions with respect to writing and word choice.

Natsume's current target users are intermediate to advanced learners of Japanese, as well as Japanese native speakers.

4.2 Main Features

The interface can be divided into three views:

1. Collocation view – where users search for the collocate words of any noun, adjective or verb.
2. Genre comparison view – looking at the genre frequency distribution of a collocation reveals that collocation's genre tendencies.
3. Example sentence view – authentic examples enable the learner to see how the collocation is used at the sentence level.

Users must select the particular collocation pattern they want to search for and a matching noun, verb or adjective into the search box to start the search. Searching for a word will present several lists, grouped by case particle and sorted by frequency, of the searched word's collocates. The sorting scheme is user selectable and one can choose from the default frequency, Dice's coefficient, t score, Jaccard similarity coefficient, Log-likelihood ratio, Chi-square coefficient, and Mutual-Information score for different types of collocations. The color bars at the right of every collocate indicate the relative frequency (or 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 on which one is more suited for them. Using this feature, users can additionally resort on any input word, which makes it easy to see at a glance which words collocate with which input words.

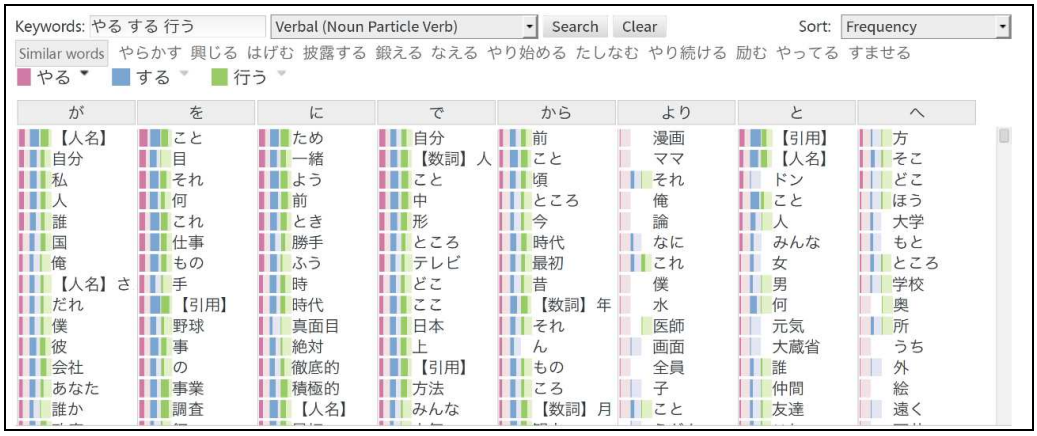

Figure 6: Main interface containing word search input area, similar words feature and collocates of the three input words. Frequency information for each verb is uniquely color coded.

When the user is interested in seeing more information on a particular collocation triplet, clicking on the collocate will load the genre comparison view to the bottom of the main collocation view. The behavior of the click can be set to one of:

- particle/conjugation expansion – can be used to compare among different grammatical uses of collocates
- synonym expansion – can be used to automatically compare among similar collocates
- no expansion (default) – standard view, only provides genre information of selected collocation
- click expansion – can be used to manually compare genre information of collocates

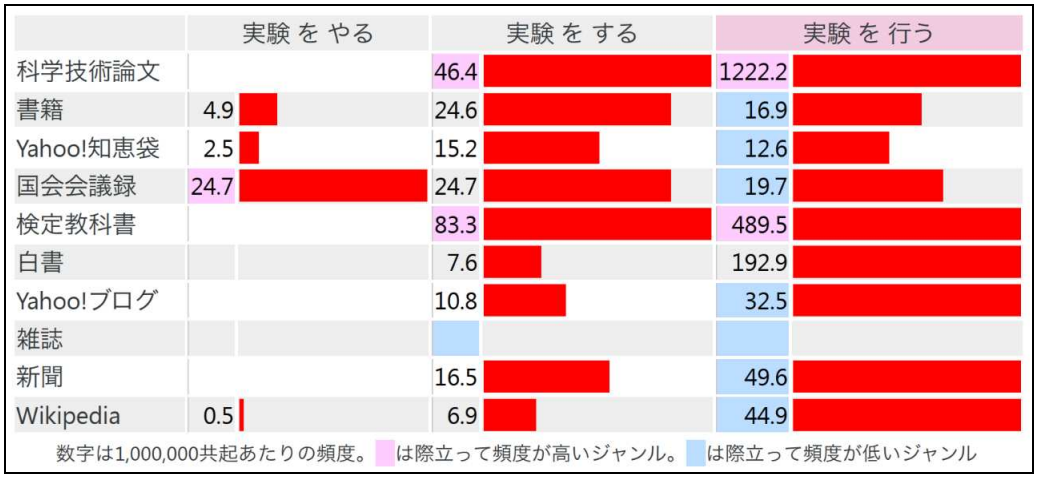

Figure 7: Comparing three collocates of /jikken/ “experiment” taking the /wo/ case particle: /yaru/ “to do” (colloquial), /suru/ “to do”, and /okonau/ “to conduct, carry out”.

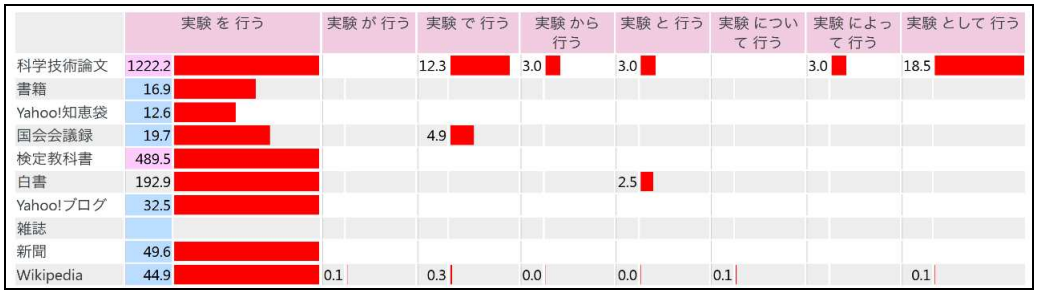

Figure 8: Comparing genre frequencies between different patterns including /jikken/ and /okonau/ using the case particle and conjugation expansion feature.

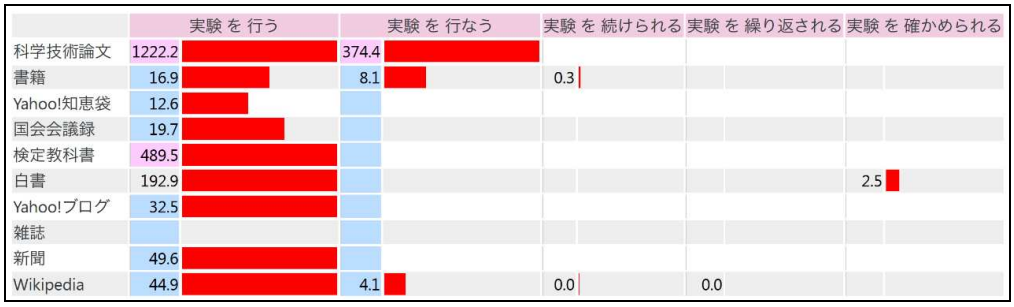

Figure 9: Comparing genre frequencies between similar collocates of /jikken/ using the similarity expansion feature.

In the genre comparison view, users can visually compare a collocations usage across different genres. The frequency numbers visible in the genre comparison interface are the relative frequency of occurrence of a collocation per 100,000 collocations. This is done to ensure the frequencies are comparable even if the corpus sizes differ, as is the case here. Additionally, Natsume uses the chi-square test to color-code genres as blue if the frequency of occurrence is significantly larger than the average across all genres, and pink if the frequency is significantly lower. Genres that are not color coded do not significantly differ 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. Sentences are displayed randomly by genre up to a limit of six sentences per genre. One can judge if a collocation is suitable for one's writing context by comparing its frequency across genres, its differences with similar collocations, and the actual usage as seen in example sentences from different corpora.

Figure 10: Examples sentences of /jikken wo okonau/.

4.3 Related Work

In parallel to the construction of the BCCWJ, NINJAL commissioned the construction of two search systems, one that is freely available and offers basic KWIC search features, called Shonagon, and another subscription-based one that allows searching with regular expressions over short and long unit words, called Chunagon[14]. Another system that shares Natsume's focus on Japanese language education is NINJAL-LWP[15], a lexical profiler for a subset of the BCCWJ (Pardeshi, 2012). It contains features similar to Natsume, but differentiates itself by providing many different kinds of collocations.

Perhaps the most sophisticated collocation query system for Japanese is the Sketch Engine, a "Corpus Query System incorporating word sketches, one-page, automatic, corpus-derived summary of a word's grammatical and collocational behaviour" (Sketch Engine, 2012; Kilgarriff, Rychly, Smrz, & Tugwell, 2004). The Sketch Engine supports multiple languages including Japanese through a 400 million token web-based corpus (JpWaC) that was first released in 2008. More than 50 collocational and grammatical relations are in use in the word sketch grammar (Srdanović-Erjavec, Erjavec, & Kilgarriff, 2008). The Sketch Engine also contains a unique word comparison feature, called word sketch difference, which is in some aspects similar to searching for several words at the same time using Natsume, though it is also more sophisticated.

4.4 Collocation Data Extraction

The Japanese dependency analyzer CaboCha was used to extract the dependency structure of all sentences in the corpora, from which noun, particle and verb or adjective dependent patterns were extracted. Post-processing was performed on verbs to differentiate passive (/iwareru/, passive of "say") or potential (/ieru/ "be able to say")[16] with causative (/iwasu/ "to make talk") voice usage, as well as combine verbal compounds into single units (/kaki + hajimeru/ "begin to write"). Nouns were post-processed to normalize numbers, dates and personal names.

[14] Available at http://www.kotonoha.gr.jp/shonagon/ and https://chunagon.ninjal.ac.jp/, respectively.

[15] NINJAL-LWP is accessible from http://nlb.ninjal.ac.jp/.

[16] Passive and potential usage is not always discriminated by the underlying MeCab morphological analyzer and IPA electronic dictionary.

Table 3: Collocation token and type count per genre.

Genre	NPV tokens	NPV types	NPAdj tokens	NPAdj types	AdjN tokens	AdjN types
STJC	323,182	164,803	16,936	8,755	20,153	9,794
Wikipedia	17,935,354	6,818,612	614,75	203,547	950,112	338,833
Books	2,837,802	1,547,893	117,309	59,793	236,534	121,21
Yahoo! Q&A	394,228	241,74	32,848	18,005	35,508	20,328
Diet minutes	404,819	193,774	17,555	8,318	30,092	13,809
Textbooks	96,007	66,72	3,185	2,225	6,925	4,857
White papers	393,881	165,535	15,978	6,934	27,567	11,016
Yahoo! Blogs	184,122	136,134	10,418	7,678	18,759	13,877
Magazines	20,144	17,132	922	797	1,91	1,766
Newspapers	60,447	49,683	2,006	1,665	3,652	3,179
Total	**22,649,986**	/	**831,907**	/	**1,331,212**	/

4.5 Genre

The defining feature of Natsume is the ability to differentiate between expressions that are suitable for writing in an academic context and those that are not. Consider the following example from the STJC corpus:

- /Taisha ni yori hasseishita nisankatanso wa mizu ni yōkaishi, .../ "The CO2 produced from metabolism dissolved in water, ... "[17]

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", it is clear that the former is written in an academic, technical style, while the latter is of a more informal, spoken variety. For learners without the native language intuition needed to arrive at the same conclusion, Natsume provides a data-driven way of helping them to take a first step towards gaining this kind of intuition.

4.6 Conclusion and Future Work

Natsume is primarily a system to assist a specific part of the writing process: finding the right words for a particular writing context, which in this case is technical

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

Japanese. Currently, example sentences in Natsume are displayed randomly. Tailoring the example sentence view to display examples appropriate to the learner's proficiency level is a future goal of the project (Hodošček, Abekawa, Murota, & Nishina, 2012).

5. Natane: Error Search System

5.1 Introduction

Having introduced the corpus Natane in Chapter 2, this chapter focuses on the search capabilities of Natane and how they might help the researcher or Japanese language teacher in analyzing their own students' errors or identify particular areas where learners have tendencies to make errors.

5.2 Interface

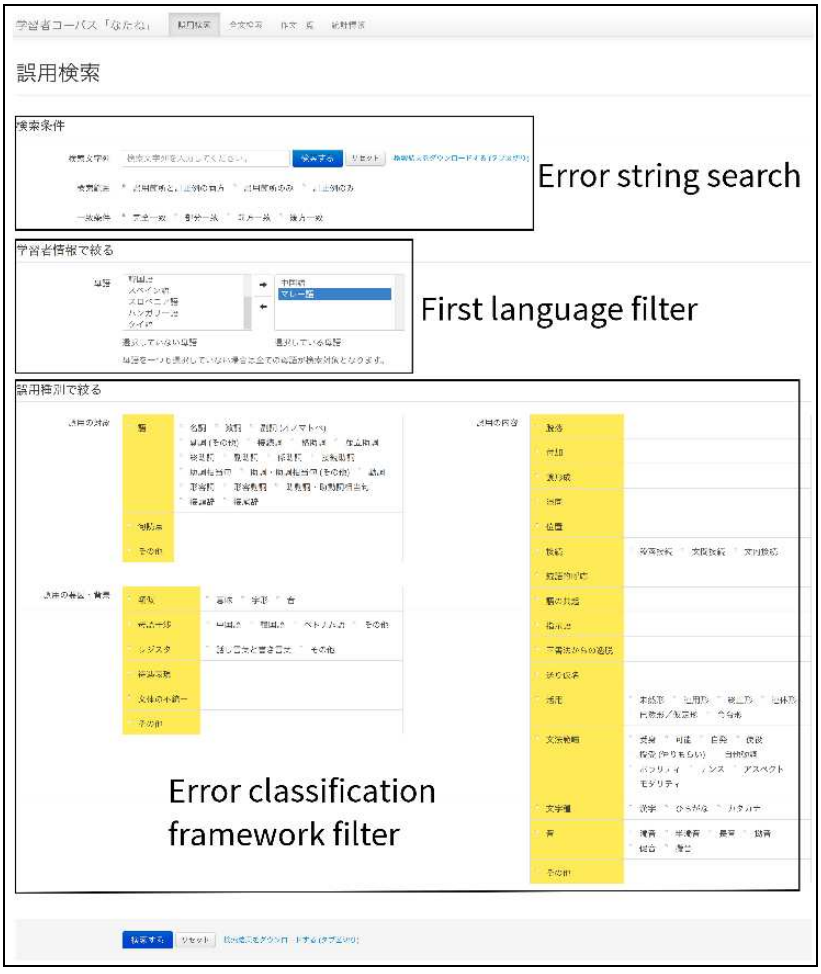

Figure 11: Natane interface: search for learner errors and filter based on first language and specific error types.

Figure 12: Searching for /yaru/ “to do” will return all learner errors containing the word. Here the correct way of writing the second sentence is to replace /yaru/ with its polite version /suru/.

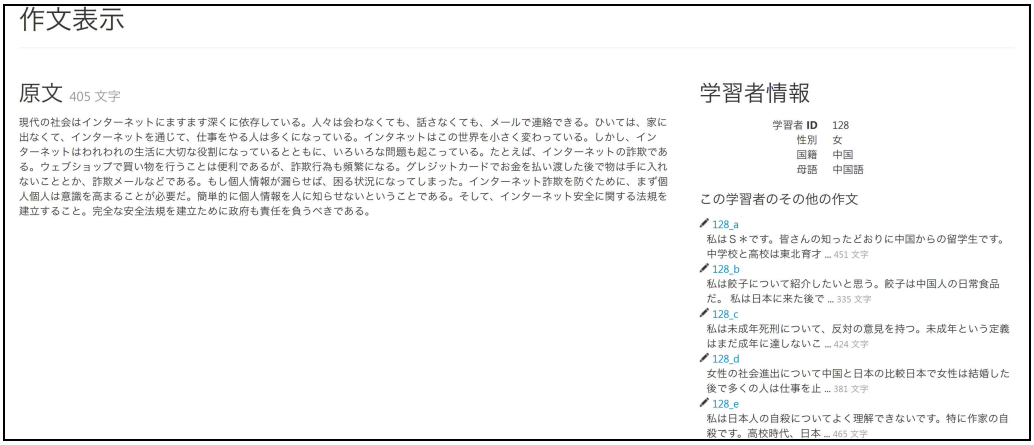

Figure 13: Situating the previous error in the learner essay.

誤用一覧

15 件の誤用タグが付与されています。

作文 ID	誤用 ID	前文脈	誤用箇所	後文脈	誤用の対象	誤用の内容	誤用の要因・背景	訂正例	解説
128_f	25084	現代の社会はインターネットにますます	深くに	依存している。人々は会わなくても、話さなくても、メールで連絡できる。ひいては、家	語::形容詞	付加, 混同, 誤形成		深く	形容詞の副詞的用法の誤り、
128_f	25085	現代の社会はインターネットにますます深くに	依存している	。人々は会わなくても、話さなくても、メールで連絡できる。ひいては、家に出なくて、	語::助動詞・助動詞相当句, 語::動詞	文法範疇::アスペクト		依存することになってきた	
128_f	25086	すます深くに依存している。人々は会わなくても、話さなくても、メールで連絡できる。	ひいては	、家に出なくて、インターネットを通じて、仕事をやる人は多くになっている。インタネ	語::副詞	混同		さらには	
128_f	25087	依存している。人々は会わなくても、話さなくても、メールで連絡できる。ひいては、家	に	出なくて、インターネットを通じて、仕事をやる人は多くになっている。インタネットは	語::助詞・助詞相当句::格助詞	混同		を	
128_f	25088	的に個人情報を人に知らせないということである。そして、インターネット安全に関する	法規を建立する	こと。完全な安全法規を建立ために政府も責任を負うべきである。				法規を確立する	
128_f	25089	お金を払い渡した後で物は手に入れないこととか、詐欺メールなどである。もし個人情報	が	漏らせば、困る状況になってしまった。インターネット詐欺を防ぐために、まず個人個人	語::助詞・助詞相当句::格助詞	混同		を	
128_f	25090	いということである。そして、インターネット安全に関する法規を建立すること。完全な	安全法規を建立ため	に政府も責任を負うべきである。	語::動詞	語の共起(コロケーション), 脱落, 誤形成	類似::意味	安全法規を確立するため	
128_f	25091	しまった。インターネット詐欺を防ぐために、まず個人個人は意識を高まることが必要だ	。簡単的に	個人情報を人に知らせないということである。そして、インターネット安全に関する法規	語::副詞	混同, 誤形成, 付加		簡単に	
128_f	25092	ても、メールで連絡できる。ひいては、家に出なくて、インターネットを通じて、仕事を	やる	人は多くになっている。インタネットはこの世界を小さく変わっている。しかし、インタ	語::動詞	混同	レジスタ::話し言葉と書き言葉	する	
128_f	25093	、困る状況になってしまった。インターネット詐欺を防ぐために、まず個人個人は意識を	高まる	ことが必要だ。簡単的に個人情報を人に知らせないということである。そして、インター	語::動詞	文法範疇::ヴォイス::自他動詞		高める	
128_f	25094	人に知らせないということである。そして、インターネット安全に関する法規を建立する	こと。	完全な安全法規を建立ために政府も責任を負うべきである。	語::助動詞・助動詞相当句	脱落	文体の不統一, レジスタ::その他	ことである。	
128_f	25095	欺である。ウェブショップで買い物を行うことは便利であるが、詐欺行為も頻繁になる。	グレジットカード	でお金を払い渡した後で物は手に入れないこととか、詐欺メールなどである。もし個人情	語::名詞	文字種::カタカナ, 音::濁音	類似::音	クレジットカード	
128_f	25096	ーネットを通じて、仕事をやる人は多くになっている。インタネットはこの世界を小さく	変わっている	。しかし、インターネットはわれわれの生活に大切な役割になっているとともに、いろい	語::助動詞・助動詞相当句, 語::動詞	文法範疇::アスペクト	類似::意味	している	
128_f	25097	存している。人々は会わなくても、話さなくても、メールで連絡できる。ひいては、家に	出なくて	、インターネットを通じて、仕事をやる人は多くになっている。インタネットはこの世界	語::助動詞・助動詞相当句, 語::動詞	接続::文内接続, 位置, 混同, 誤形成, 付加	類似::意味	出ることなく	
128_f	25098	は手に入れないこととか、詐欺メールなどである。もし個人情報が漏らせば、困る状況に	なってしまった	。インターネット詐欺を防ぐために、まず個人個人は意識を高まることが必要だ。簡単的	語::助動詞・助動詞相当句	文法範疇::アスペクト		なってしまう	

Figure 14: Viewing all learner errors in a given essay.

Searching for errors relating to the verb /yaru/ returns two errors. One example is the sentence /intānetto wo tsūjite shigoto wo yaru hito wa ōku natte iru/ “the number of people working on the Internet is increasing”, where the usage of /yaru/ is wrong because it is the colloquial form of /suru/.

5.3 Conclusion and Future Work

Natane is a learner corpus that has many potential uses, though we envision two main types of usages, one by Japanese language educators and the other by NLP researchers.

In this chapter, we described the search interface for the Natane corpus, which is targeted at the former. Japanese language educators can make use of Natane to find examples of learner errors. Also, the data provided is useful for analyzing error

tendencies due to first language interference, as well as for observing the language acquisition process.

The latter usage is primarily aimed at applications in NLP and machine learning, where Natane can be used to construct novel error correction systems. An example of one such system is introduced in the next chapter.

With the existence of several Japanese language learner corpora that all make use of different error classification frameworks, a movement towards a common standard is, perhaps, the most pressing issue.

6. Nutmeg: Writing Assistance and Automatic Error Correction System

6.1 Introduction

Natsume, while useful for finding collocations, does not automatically correct the learner's writing. In an evaluation of Natsume, it became clear that for every collocation the learner checked using Natsume, there were many more that went unchecked (Hodošček, Abekawa, Bekeš, & Nishina, 2011). The next obvious step was to develop a system that checks learners' writing and provides feedback on any errors they may have made. This writing assistance system was named Nutmeg and provides basic feedback for learners' writing using automatic error identification (Yagi, Hodošček, & Nishina, 2012). The system is unique in that it does this from two sources: native and learner corpora.

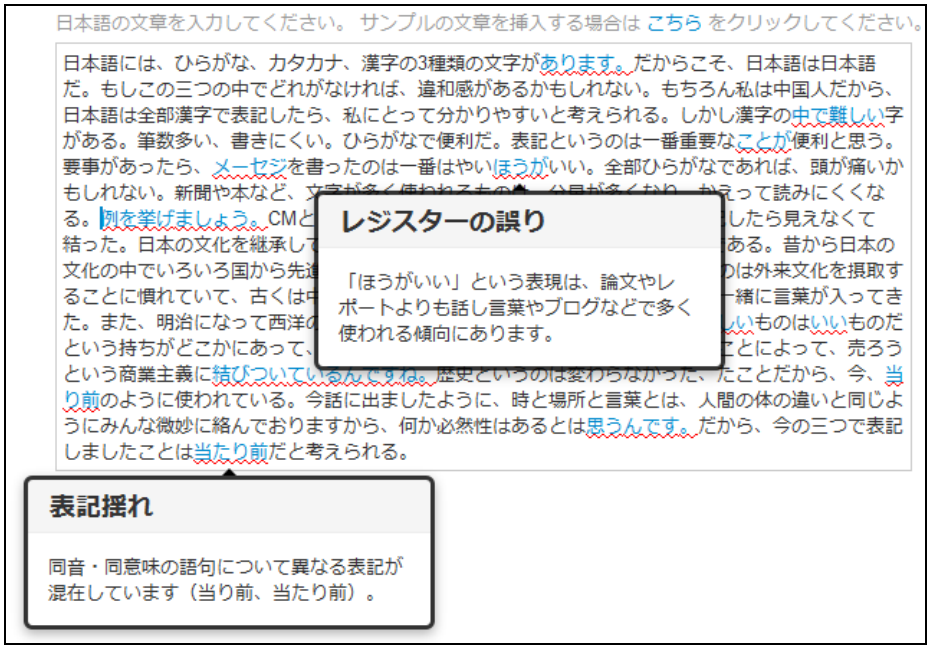

Figure 15: Nutmeg interface showing two correction suggestions.

6.2 Related Work

Compared to other existing Japanese language automatic composition correction systems, Nutmeg strives to incorporate both native and learner corpora in its correction model. An example of a more narrow application of a similar system is Chantokun. Developed at the Nara Advanced Institute of Science and Technology (NAIST), Chantokun[18] is a system that detects and corrects case particle misuse based on corrected Japanese language sentences from the Lang-8 website[19], a language-exchange social networking website where users with different first languages correct each other's writing (Mizumoto & Komachi, 2012).

An example of a system that focuses on the native corpus side of automatic error correction is the Japanese proofreading system Tomarigi[20] (Oono & Inazumi, 2011). Another example that uses the dependency structure of a sentence to revise complex sentences into easier to understand ones is the jcorrect tool (Oosaki, 2006).

6.3 Error Correction Method

In general, Nutmeg uses the native BCCWJ and STJC corpora as "correct data", whereas it uses learner errors from Natane as "incorrect data". Thus expressions that are tagged as errors in Natane become candidates for automatic correction.

Natane contains 386 orthographic errors. One way of detecting outright orthographic errors is if they go unrecognized in morphological analysis. Additionally, most such errors are found within two letters of a word. A word including an error is replaced with the corresponding word in the native word list. For example, suppose a learner were to mistaken the word /messeeji/ "message" as /meeseji/. If there is no prior learner error of the same word in Natane, then a morphological analysis reveals that it is an unknown word. The unknown word can then be matched to similar words contained in the morphological dictionary. Finally, the correct orthography can be presented to the learner.

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 outlined in Chapter 4. If a learner uses the collocation /jikken wo yaru/ in an academic report, the system will be able to identify the inappropriate usage and offer the replacement

[18] See http://cl.naist.jp/chantokun/ for more information.

[19] Accessible from https://lang-8.com/.

[20] More information and download links available at http://www.pawel.jp/outline_of_tools/tomarigi/.

collocation /jikken wo 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, 2011). The corpora in Natsume can thus be divided into so-called positive and negative genres and the relative frequencies of collocations in those genres can be tested using the chi-square test. When an expression like /jikken wo yaru/ is used we can 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 WordNet to expand the available search space (Bond et al., 2009; Isahara et al., 2012).

6.4 Conclusion and Future Work

The aim of Nutmeg is to become a compositional tool that is able to automatically warn learners of potential mistakes as they are making them. The two types of data backing Nutmeg's error correction facilities are native and learner corpora corresponding to the data provided by Natsume and Natane, respectively.

Though the available size of native corpora is much greater than that of learner corpora, an avenue for improvement to collocation error correction is to provide candidate replacement expressions for learner errors, perhaps using WordNet. More effort must be put into obtaining or constructing other specific-purpose corpora if other writing genres, such as business writing, are to be considered.

It is also clear that Natane should be expanded in scale in order to conduct a more comprehensive quantitative evaluation. The implementation of an automatic error correction system must be treated cautiously, because the results of automatic error correction depend on the annotations being objective. This is especially difficult for learner errors at the semantic or discourse levels, as it is here that annotators subjectivity most easily comes into play. Thus, as a first step, easier items such as orthographic errors should be considered (Yagi, Hodošček, & Nishina, 2012).

7. Conclusion and Future Work

In the span of just over a decade, the Hinoki project has produced the Asunaro, Natsume and Nutmeg systems as well as the Natane learner corpus. As the project is led by linguists, language teachers, computer engineers and educational engineering researchers, it has been able to synthesize ideas from these disciplines together into several multi-viewpoint CALL systems.

The construction of Asunaro resulted in the construction of a novel electronic multilingual dictionary that contains several often underrepresented Asian languages.

Asunaro also applied state of the art NLP research to provide a practical dependency grammar-based reading assistance system.

Natsume was developed as a corpus-backed collocation search tool that allows users to find new collocations that fit their writing style by enabling them to check the correctness of Japanese collocations they are not confident about. An immediate goal of the development of Natsume is the addition of new types of collocations to the search interface. Another goal is being pursued in ongoing work to channel Natsume's knowledge of genres and collocations into Nutmeg for use in automatic error correction. Finally, the extension of available native corpora to other learner-specific purposes, such as the writing of emails or business writing is also being considered.

The development of Natane has resulted in a unique Japanese learner corpus and an accompanying search system. It has applications for both language researchers and educators, as well as NLP applications. The future direction of Natane is closely aligned with that of Nutmeg, the usage of which will hopefully contribute to the development and further validation of Natane's error classification framework.

Nutmeg is an extension of both Natsume and Natane into automatic error correcting for learner writing. From the development of Natane it became clear that simpler orthographic and syntactic factors are easier to objectively annotate than semantic and discourse factors, which are more prone to subjective decision making on the part of the annotator. This subjective decision making also leads to greater difficulty in automating error correction at a reasonable precision. There is thus a need for a greater volume of annotations, that are objectively classified in the error classification framework of Natane. This is essential in order to realize more sophisticated error correction and composition assistance.

Finally, an effort should be made to move from the localized lexical writing assistance seen in Natsume and Nutmeg towards a more comprehensive discourse-level composition assistance. For this purpose, more inter-system collaboration with other projects is needed.

References

Abekawa, T., Hodošček, B., & Nishina, K. (2011). Go no kyōki wo 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.

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.

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., 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 wo 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., Yagi, Y., & Nishina, F. K. 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

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

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

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: Bonjinsha.

Hodošček, B. (in press). 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). Asakura Publishing Co., Ltd.

Hodošček, B., & Nishina, K. (2011, 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., 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.

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_BorHodoscek.pdf

Isahara, H., Bond, F., Kanzaki, K., Uchimoto, K., Kuroda, K., Kuribayashi, T., Torisawa, K. (2012). Japanese WordNet. Retrieved December 10, 2012, from http://nlpwww.nict.go.jp/wn-ja/index.en.html

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

Kasahara, S. (2012). Chantokun —tōkeiteki nihongo kōsei— [Chantokun —statistical correction of Japanese—]. Retrieved December 10, 2012, from http://cl.naist.jp/chantokun/

Kawamura, Y., & Kitamura, T. (2001, March). Development of a Japanese reading resource bank using the Internet. *Current report on Japanese-language education around the globe, 6*, 241–255. Retrieved from http://ci.nii.ac.jp/naid/110001046390/en/

Kawamura, Y., Kitamura, T., & Hobara, R. (2000, August). Development of a reading tutorial system for JSL and JFL learners using the EDR Electronic Japanese-English Dictionary. *Japan journal of educational technology, 24*, 7–12. Retrieved from http://ci.nii.ac.jp/naid/10008760744/en/

Kawamura, Y., Kitamura, T., & Hobara, R. (2012). Japanese language reading tutorial system. Retrieved December 10, 2012, from http://language.tiu.ac.jp/

Kilgarri ▪ , A., Rychly, P., Smrz, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of EURALEX*.

Kudo, T. (2012). MeCab: Yet Another Japanese Dependency Structure Analyzer. Retrieved December 10, 2012, from https://code.google.com/p/mecab/

Kudo, T., & Matsumoto, Y. (2012). CaboCha: Yet Another Japanese Dependency Structure Analyzer. Retrieved December 10, 2012, from https://code.google.com/p/cabocha/

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

Maebo, K. (2012). A survey of register labelling in Japanese dictionaries - Towards the labelling of words in dictionaries for learners of Japanese. *Acta Linguistica Asiatica, 3*(3), 9-26.

Maekawa, K. (2007a, March 1–3). 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

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.

Nishina, K. (2001). *Yasashi kagaku gijutsu nihongo dokkai nyūmon (kaitei-ban) [A gentle introduction to scientific and technical Japanese reading comprehension (revised edition)]*. International Student Center, Tokyo Institute of Technology.

Nishina, K., Okumura, M., Abekawa, T., Yagi, Y., Bilac, S., & Fu, L. (2004, March 8–9). 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.

Nishina, K., Okumura, M., Yagi, Y., Totugi, N., Sawaya, T., Fu, L., … Abekawa, T. (2002). Kōbun hyōji to tagengo intāfēsu wo sonaeta nihongo dokkai gakushū shien shisutemu no kaihatsu [Development of a reading assistance system for Japanese language containing grammar display and multilingual interface features]. In *Proceedings of the 8th conference of the Association of Natural Language Processing* (pp. 228–231). Association of Natural Language Processing.

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/

Oosaki, H. (2006). Tips for technical writing. Retrieved December 10, 2012, from http://www.ispl.jp/~oosaki/research/ tips-jcorrect/

Pardeshi, P. (2012). Compilation of Japanese Basic Verb Usage Handbook for JFL Learners: A Project Report. *Acta Linguistica Asiatica*, *2*(2), 37-63.

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.

Sketch Engine: SketchEngine. (2012). Retrieved December 10, 2012, from http://www.sketchengine.co.uk/

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

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: Bonjinsha.

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 [Proceeding of the first workshop on Japanese corpus linguistics]*. Dai ikkai nihongo kōpasu wākushoppu [First workshop on Japanese corpus linguistics]. Tokyo.

Acknowledgments

We thank Irena Srdanović for help with a revision. We also thank the editors for their useful comments and suggestions.