8 Analysis of correctness in adverb use in the Japanese composition support system Nutmeg

Bor HODOŠČEK, NISHINA Kikuko, YAGI Yutaka, ABEKAWA Takeshi
Osaka University, Tokyo Institute of Technology, Picolab Co., Ltd., National Institute of Informatics

Abstract
Nutmeg (http://hinoki-project.org/nutmeg/) is a writing support system for Japanese language learners. It can identify probable mistakes in learner writing by classifying expressions based on their frequency distribution across several native Japanese corpora representing various registers. Namely, it divides the corpora into a positive group representing the target register and a negative group representing registers considered to contain inappropriate stylistic features. The purpose of this study is to examine adverb usage within the Japanese academic register and to evaluate the classification results of the system. The system classified 2,919 adverbs extracted from the electronic dictionary UniDic into 'acceptable', 'unacceptable' or 'unknown' classes. These results were compared to an independent classification by an L2 education expert and revealed differences, especially in the low recall performance of the system. Furthermore, adverbs that had a relatively high frequency in the positive corpus set were incorrectly classified as unacceptable. An investigation into these problems revealed that the classification of a lemma according to its different orthographic forms resulted in some of the differences between the human and system evaluations. Because the system classification works at the level of single morphemes, it could not arrive at the right conclusion in instances where the correct unit of classification was a morpheme compound. Other future tasks include classifying the multiple usage and meanings of a single lemma as separate items.

Keywords: academic writing, Japanese language learner, writing support system, register, large-scale Japanese corpora, Scientific and Technical Japanese Corpus, adverbs

1 Introduction

Foreign students enrolled in undergraduate programs in science, technology, engineering, and mathematics (STEM) fields in Japan are often required to write homework assignments, experimental results, graduation theses as well as research papers in academic Japanese. However, courses geared towards beginner and intermediate level learners of Japanese as a second language tend to emphasize the acquisition of spoken language. As a result, learners are inadequately prepared for academic writing and often struggle over how to correctly write academic texts. A common example is choosing the more appropriate writing form between *de aru* or *da* forms (copular verbs corresponding to 'is/are'). The following short passage, extracted from the Natane learners corpus[1], is written by a native Chinese first-year science student and illustrates the use of spoken words (*sugoku, totemo*) where more appropriate semantically-compatible replacements exist (*kiwamete, hijōni*):

Ex. 1) 日本の婚姻制度は中国と大体同じである。ふるい時代にくらべて、すごく自由、平等になった。日本の法律によると、未成年も結婚できることを了解して、とてもびっくりした。 *Nihon no kon'inseido wa Chūgoku to daitai onaji de aru. Furui jidai ni kurabete, sugoku jiyū, byōdō ni natta. Nihon no hōritsu ni yoru to, miseinen mo kekkondekiru koto wo ryōkaishite, totemo bikkurishita.* 'The marriage system of Japan is almost the same as that of China. It has become much more free and fair when compared to previous eras. I was very surprised to learn that adolescents were allowed to marry under Japanese law.'

An earlier survey based on the Natane error annotations revealed that adverb related errors, similar to those of Ex. 1, were among the most frequent (Yagi et al. 2014a; Yagi et al. 2014b). Adverbs are also an advantageous research target because a relatively smaller set of them are used in academic writing compared to spoken language. Also, the variety of adverb usage greatly differs along register lines.

Furthermore, while sentence-final expressions and function words that connect phrases and sentences are perhaps more indicative of register differences (Srdanović, Hodošček, Bekeš, & Nishina 2009), making them a valid subject of such a study, they have the undesirable property of transcending morpheme and phrasal (*bunsetsu*) boundaries, the latter form of which are not supported in the error classification API used in this research. In most cases adverbs are formed from one morpheme and are thus a more immediately tractable target, with the exception of the adverbs examined in Section 4.2.

Nutmeg's main focus is to assist the process of writing academic Japanese. It analyzes the user's text input and points out any expressions that are inconsistent with the academic writing register, thereby forcing users to reflect on their word choice and in the process, hopefully improve their writing (Yagi, Hodošček, Abekawa, Nishina, & Murota 2014). The current focus is on the identification of errors and not the automatic suggestion of alternative expressions, the development of which are left to future research.

Our research uses language-processing techniques on large-scale corpora, and aims to provide automatic corrections that are appropriate for the register required by the learner. Ng et al. (2014) describe a recent task on grammatical error correction in which many teams made use of machine learning, statistical machine translation and

[1] The Natane Learners Corpus contains over 200 essays collected or elicited from L2 Japanese learners and is available from https://hinoki-project.org/natane/. Example 1 is available from http://hinoki-project.org/natane/document/151_a/show.

rule-based approaches to various degrees of success. At the present time, our research only aims for the identification of errors and does not aim to give automatic corrections. Indeed, Hodošček (2011) previously showed that classifying words or expressions for suitability to a genre using only frequency data from large-scale corpora is a feasible and simple approach.

From the viewpoint of the overall effectiveness of writing assistance systems, Yagi et al. (2014a; 2014b) conducted experiments on how learners react to errors shown by the Nutmeg system. The results recommend showing learners only a few example sentences when they correct texts by themselves. Abekawa et al. (2015) analyze tendencies in learner errors related to adverbs from the viewpoint of the academic register by comparing learner errors and adverbs listed in the official vocabulary of the pre-2010 JLPT (Japanese Language Proficiency Test) in order to help develop methods of correction.

From a narrower perspective of error-types, the Chantokun system (Mizumoto 2012) identifies and corrects Japanese case particle usage using a classifier trained by feeding in Japanese learner texts and their associated corrected versions constructed by native speakers on the Lang-8 website[2]. The major difference between Mizumoto (2012) and our present research is that the former employs the use of error-corrected learner corpora for misuse detection, whereas the latter uses native corpora representing varied registers for misuse detection.

From the perspective of research in second language education, Watanabe (2010) analyzes differences in adverb usage within academic reports written by learners and native speakers. The research shows that learners tend to use inappropriate degree adverbs such as the colloquial 一番 *ichiban* 'the most'. Watanabe's research is similar to our present study as both focus on adverb usage, but differs in methodology. Watanabe's research is based on a manual analysis, whereas our research is based on a predictive analysis using corpus data. Moreover, Watanabe (2010) addresses the issue of the L2 Japanese academic writing curriculum as part of the research aim, whereas we developed this system as a tool for self-study.

2 Academic Japanese and official orthographic policy

Textual genres are often described in terms of differences: spoken vs. written, colloquial vs. formal, objective and logical reasoning in essays vs. subjective emotional descriptions. This carries over into the expressions used in those registers. Varieties of language, directly connected to the situation of their use, are referred to as register (Halliday &

[2] Accessible from http://lang-8.com/.

Hasan 1976). The present work thus focuses on the appropriateness of learners' expressions to the academic register, with the goal of improving learner composition by conforming to the academic writing style.

Compared with several other languages with clearly encoded spelling rules, a compulsory orthographic policy for the Japanese language has not yet been established. The Ministry of Education, Culture, Sports and Technology (MEXT) has published a standardized manual called *Kōbunsho Sakusei Yōryō* 'Criteria for the writing of official documents' (Kōbunsho Manual) for various orthographies of official government documents (MEXT 2014). However, magazines, newspapers, and other media in Japan are not bound to its rules and tend to have their own more-or-less equivalent but differing internal style manuals. As a consequence of the lack of standardization across these groups and organizations, the orthography of the Japanese writing system remains complicated. In this research we assume that academic communities have their own writing rules that are relatively close to the standard orthographic rules set out by MEXT, but will examine the validity of this assumption in the succeeding sections.

3 Methods and materials

3.1 Corpus selection criteria

The selection of appropriate corpora is essential to realizing the goals of the system: namely, to provide feedback on learners' written errors within the genre of scientific and technical academic writing. The combination of the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al. 2013), which comprises a diverse range of registers including informal and spoken text, is combined with the Scientific and Technical Japanese Corpus[3] to satisfy these requirements.

Classifying a word as either appropriate or inappropriate for the academic register relies on quantifying its appropriateness with respect to a variety of corpora with well-defined situational characteristics. Therefore, taking those corpora from the BCCWJ and the STJC for which the situational characteristics most closely align with the academic register, one can attempt to infer a word's appropriateness. For words that do not appear or are rarely found within corpora belonging to the academic register, one approach is to just mark them as inappropriate. The approach outlined in this paper takes a different stance in which, in order to identify a word as inappropriate, it is not enough to simply find the word within the set of corpora closest to the academic register, but also necessary to have a separate set of corpora for which the situational

[3] The Scientific and Technical Japanese Corpus (STJC) is an ongoing project seeking to form a representative sample of scientific and technical Japanese. It is formed from Japanese language journals and proceedings in such fields as Natural Language Processing, Civil Engineering, Electrical Engineering, and Medicine.

characteristics differ enough from the academic register so that a significant presence of a word within them is taken to be a strong indicator of inappropriate use within the academic register.

For the positive corpus set, the STJC along with the White papers and Law documents media from the BCCWJ were selected, while for the negative corpus set, Yahoo! Q&A, Yahoo! Blogs, and the Minutes of the Diet media, all from the BCCWJ, were selected. While the White paper and Law documents sub-corpora are not strictly academic in nature, many of their situational characteristics are shared with or similar to the STJC. Indeed, previous research has shown that their writing style (Hodoščeк 2011), and specifically the fact that they can be considered to have undergone editing for consistency with other publications in their fields and are meant for an expert audience, are similar enough to the STJC that we are able to justify their inclusion in the positive set. As the choice of corpora for the negative set was constrained to the corpora available within the BCCWJ, sub-corpora that consistently contain either transcribed speech (Minutes of the Diet) or contain informal writing (Yahoo! Blogs and Yahoo! Q&A) were selected. Finally, the remaining corpora are essential for deciding whether an adverb's relative frequency is exceptionally high or low in the positive and negative corpus sets when compared to 'average' Japanese prose.

3.2 Adverb selection criteria

The list of adverbs examined within this study was compiled from the full list of adverbs within the UniDic morphological dictionary. UniDic was jointly developed alongside the BCCWJ and employs a hierarchical structure that captures the orthographic variation inherent within the Japanese writing system. Morphemes are organized under their lemma, word, and orthographic form to encode the structure shown in Figure 1 (Ogiso et al. 2010), where the adverb *yahari* 'well' is divided into 6 or more (see Table 5 in Section 4.1.2) orthographic forms (やはり, ヤハリ, 矢張り, やっぱり etc.). Unlike a traditional dictionary, the lemma is organized at the level of meaning so that polysemious words having identical word and orthographic forms can be organized under two or more different lemmas.

The question of which orthographic form of a word to choose from when writing is dependent on the particular writing context into which the word is to be inserted. For the purposes of this study, we hypothesize that for the academic register, in which clarity of communication is at a premium, standardization within most academic domains will mean that there is in general a single preferred way of writing a word. We therefore choose to prioritize the analysis of words at their orthographic form level first, and then to select several examples to be additionally analyzed from the lemma level. There are merits and demerits to both approaches. As mentioned in Section 2 above, we need to

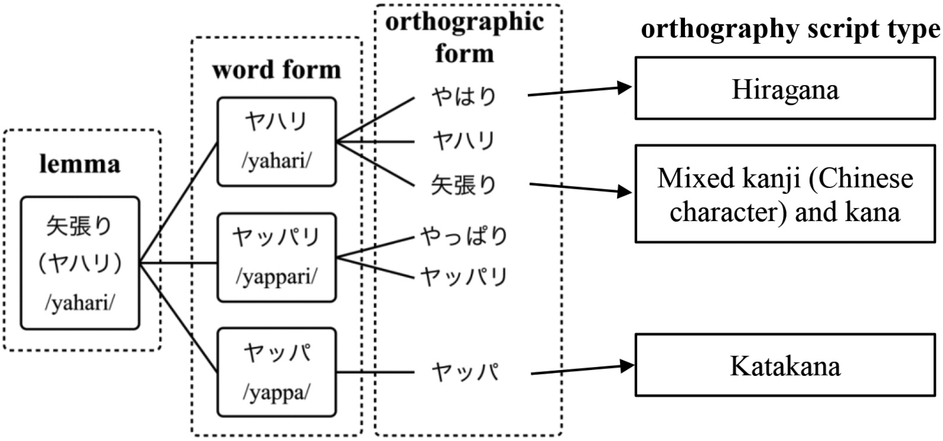

Figure 1. Word and orthographic forms of the adverb 矢張り *yahari* 'just as I thought' within the layered structure of UniDic.

take into consideration official orthography as well, which is not guaranteed to match our data-driven results.

In this research, we use the adverb list extracted from UniDic, which includes 7,432 orthographic forms of adverbs. However, due to the classification API not distinguishing between words having different pronunciation but sharing orthographic and lemma forms, 29 entries are removed, leaving 7,403 entries. Among these, we further exclude 878 orthographic forms which could not be found in any corpus, and 3,606 forms of onomatopoeic words, leaving the final number of adverbs used in the evaluation at 2,919. We then apply the register misuse identification method explained below to classify each into acceptable, unacceptable or unknown classes. In order to evaluate these predictions, we requested a Japanese language education expert separately evaluate the list.

3.3 Comparison between positive and negative corpus sets

Figure 2 shows the differences within the top 30 most frequent adverbs in the whole corpus set, the positive corpus set and the negative corpus set. The values in the figure represent PPM (parts-per-million), which corresponds to how many times an adverb occurred within a million-word long span of text. Based on the figure and statistics from the whole dataset, we can make several observations on three levels: adverb variety, magnitude of use, and adverb preference.

Firstly, the variety of adverbs employed differs greatly between the positive and negative sets: 1,023 used in the positive set compared with 2,253 used in the negative corpus set and 2,887 in the whole set. Just 73 of the most frequent adverbs in the positive corpus set account for more than 90% of all adverbs occurrences when compared with 174 adverbs for the negative set and 215 for the whole set. Secondly, overall adverb use is 4.6 times more frequent in the negative corpus set than in the positive corpus set. While the whole corpus set displays an overall higher adverb use that is 1.9 times higher than the negative corpus set, the negative corpus set has a more skewed distribution of high frequency adverbs within the top 30. Thirdly, among the top 30 adverbs in all sets, the positive corpus set features the most distinct variety and ordering of adverbs when compared with the negative and whole sets, which follow a similar pattern. When looking at the overall overlap of adverbs, all but 40 adverbs from the positive set are contained within the negative set. Among these low-frequent adverbs are a few like 別して *besshite* 'especially' that are appropriate for the academic register. However, most others are onomatopoeic adverbs which had likely originated within natural language processing research papers dealing with various aspects of onomatopoeia.

When comparing between the top 30 adverbs across the three corpus sets, we found that 18 out of the top 30 adverbs were common to the whole and positive corpus sets. Adverbs missing from the positive set include ones commonly used in informal speech, while those particular to the positive set include formal spoken adverbs that find their use in lectures and meetings such as *tatoeba* 'for example', *mottomo* 'the most', and

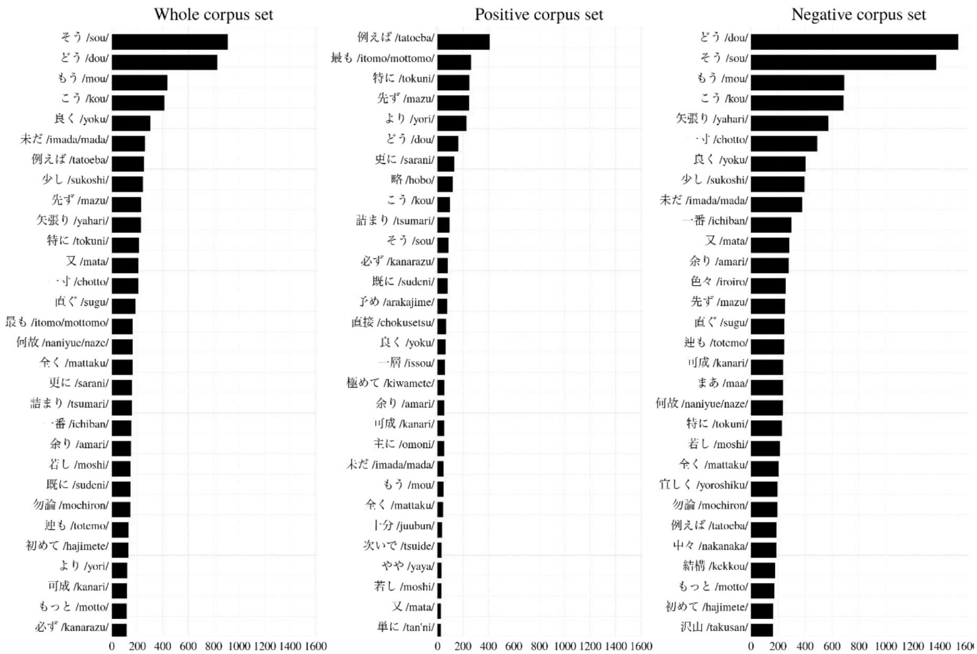

Figure 2. PPM of the top 30 adverbs in all corpora and within the positive and negative corpus sets.

tokuni 'especially'. As for the negative corpus set, 23 adverbs were shared with the whole corpus set, with the rest including adverbs such as *iroiro* 'many', *totemo* 'very', *yoroshiku* 'please', *nakanaka* 'hardly', *kekkō* 'much', *zutto* 'always, more', all of which are commonly used in everyday speech. Additionally, there were 17 adverbs not common to the positive and negative sets, including *ichiban* 'first place', *motto* 'more', *chotto* 'a bit', *mochiron* 'obviously', *yahari* 'as I thought', all of which were commonly used within learner writing and are a common source of signaling the wrong register.

3.4 Classification of adverbs with respect to suitability in the academic register

Nutmeg shows learners whether input words are acceptable or unacceptable in the academic register. Hodošček (2011) and Hodošček & Nishina (2011) proposed and described details on the basic idea of using the chi-square test on corpus data to identify expressions salient to a particular register. As the method described is the same one used in this research, we will only briefly explain the classification procedure by using two examples. The first example is *mottomo* or *itomo* 'the most', which is classified as an acceptable adverb for the academic register. The second is *chotto* 'a bit', which is classified as unacceptable. Table 3 shows the statistical data of the lemma and the orthographic forms of *mottomo* and *chotto* among adverbs from the whole corpus set, the positive corpus set and the negative corpus set. The system determines whether a word is acceptable or unacceptable for the academic resister by calculating how far its frequency within the target register deviates from the frequency distribution within all corpora by using the chi-square (χ^2) test. A word will be classified as acceptable if both the frequency of the positive set is significantly *higher* than that of all corpora and that of the negative corpus set is significantly *lower* than that of all corpora.

Table 3: Comparing classification results between the positive and negative corpus sets.

Lemma	Orth. Form	System Verdict	Frequency Whole	Frequency Positive	Frequency Negative	PPM Whole	PPM Positive	PPM Negative
最も mottomo/ itomo	最も	AC	17,492	7,189	1,274	121.83	**241.67**	**41.23**
	もっとも	UK	5,449	618	228	37.95	20.78	7.38
一寸 chotto	ちょっと	UA	27,677	193	13,975	192.77	**6.49**	**452.24**
	チョット	UK	277	2	237	1.93	0.07	7.67

Note: AC=acceptable, UA=unacceptable, UK=unknown; chi-square test: P(chi-square value > 17.275; α = 0.1)

Conversely, it will be classified as unacceptable if both the frequency of the positive set is significantly *lower* than that of all corpora and that of the negative corpus set is significantly *higher* than that of all corpora. In the examples, the lemma *mottomo* (最も) has two different orthographic forms: 最も, which is written in kanji, andもっとも, which is written in hiragana. The kanji form is classified as acceptable for the academic register, but the hiragana form is classified as unknown because no significant difference was observed. Similarly, the lemma *chotto* (一寸) also has two different orthographic forms: *chotto* (ちょっと) written in hiragana is classified as unacceptable, while *chotto* (チョット) written in katakana (the counterpart of the hiragana syllabic character pair) is classified as unknown, due to the χ^2 test not finding a significant difference between the opposing corpus sets.

3.5 Results of the differences in system and L2 expert judgments on the adverb list

Table 1 shows the number of adverbs classified as either unacceptable, acceptable or unknown by both the language expert and the system. For the purposes of this evaluation, classifications of the class unknown by the language expert were treated as insufficient grounds for identifying an adverb's use as unacceptable. We therefore treat these adverbs as acceptable for the purposes of the evaluation.

Table 1: Differences in system and L2 expert judgments on subset of UniDic adverb list for chi-square significance cutoff 0.1.

	Unacceptable	**Acceptable**	**Unknown**	**Total**
Expert	445	196	2,278	2,919
System	74	5	2,840	2,919

The system achieved a precision of 0.670, recall of 0.029, and F1 score of 0.055 when evaluated against expert's classifications. While the precision was shown to be better than a random baseline, the recall was very low, as the system identified only 74 adverbs as unacceptable, while the expert identified 445. Additionally, the system only identified 5 adverbs as particularly salient to the academic register, which also differs greatly to the 196 adverbs classified as acceptable by the expert. One reason for this is that there are few adverbs that are truly particular to the academic register, but many are acceptable simultaneously in both the academic register and other registers not represented in the positive corpus set. It should be noted, however, that the above performance numbers treat acceptable and unknown evaluations as the same—after all, the purpose of the system is to identify incorrect use, not point out if an expression is particularly well chosen. Finally, as the number of acceptable adverbs are quite low in either evaluation, we are able to presume that the overall use of adverbs in academic fields is quite limited.

4 Analysis of adverbs with erroneous classification results

In order to better understand the differences in judgement between the system and L2 expert, this section takes a detailed look at the adverbs where the judgments between the system and expert differed. As shown in Table 4, the two differ in several aspects. In order to analyze these differences, we examine the following two items in detail:

1) Complex lemma structure with several orthographic variations
2) Treatment of high frequency adverbs including KOSODO compounds

Table 4: Classification of different orthographic forms within frequently occurring adverbs.

Adverbs	System	L2 Expert	PPM Whole	PPM Positive	PPM Negative
例えば（例えば）*tatoeba*	UK	AC	14.4	330.6	146.4
例えば（たとえば）*tatoeba*	UK	AC	10.7	62.5	48.7
先ず（まず）*mazu*	UA	AC	22.4	237.6	154
先ず（先ず）*mazu*	UA	AC	0.5	2.5	7.9
特に（特に）*tokuni*	UK	AC	16.5	228.5	228.3
特に（とくに）*tokuni*	UK	AC	4.7	12.3	8.1
どう*dō*	UA	AC	43.7	154.6	1557.9
更に（さらに）*sarani*	UA	AC	14.3	99.6	98.8
更に（更に）*sarani*	AC	AC	1.5	27.4	18.7
こう（こう）*kō*	UA	AC	40.7	93.8	213.7
詰まり（つまり）*tsumari*	UA	AC	15.7	90.6	92.1
詰まり（詰まり）*tsumari*	UA	AC	0	0.1	0.6
必ず（必ず）*kanarazu*	UA	AC	10.5	71.3	122.3
必ず（かならず）*kanarazu*	UK	AC	0.8	0.9	3
そう（そう）*sō*	UA	AC	87.6	68.2	1368.3
良く（よく）*yoku*	UA	AC	28.3	55.8	385.8
良く（良く）*yoku*	UA	AC	0.9	3.4	30.1
最も*mottomo*	AC	AC	121.8	50.0	41.9
もっとも *mottomo*	UK	AC	37.9	4.3	7.7
可成（かなり）*kanari*	UA	AC	11.7	49.5	244
可成（可成）*kanaru*	AC	AC	0	0.4	0
より（より）*yori*	AC	UA	0.1	46.9	60.0
もう（もう）*mō*	UA	UA	43.3	44.4	712.7
予め（あらかじめ）*arakajime*	AC	AC	27.6	55.0	11.8

Adverbs	System	L2 Expert	PPM Whole	PPM Positive	PPM Negative
予め（予め）*arakajime*	AC	AC	6.3	4.0	3.7
（一層）一層 *issō*	AC	AC	25.2	11.5	20.5
（一層）一そう *issō*	UA	UK	0.2	0.0	0.1
矢張り（やはり）*yahari*	UA	UA	13.7	10.8	336.1
矢張り（矢張り）*yahari*	UK	UA	0.1	0.5	0.6
（主に）主に *omoni*	AC	AC	28.5	10.5	22.1
（主に）おもに *omoni*	UK	AC	3.4	0.1	0.8
（次いで）次いで *tsuide*	AC	AC	11.5	6.3	2.4
（次いで）ついで *tsuide*	AC	AC	2.7	0.3	0.3
（依然）依然 *izen*	AC	AC	14.9	5.3	7.7
（総じて）総じて *sōjite*	AC	AC	3.2	1.3	1.8
（総じて）そうじて *sōjite*	UK	UK	0.0	0.0	0.0
（概して）概して *gaishite*	AC	AC	2.7	0.6	0.9

Note: AC=acceptable, UA=unacceptable, UK=unknown

4.1 Complex lemma structure with several orthographic variations

As mentioned in Section 3, the lexical data used in the system is based on a subset of UniDic, namely the lemma and orthographic base forms of morphemes extracted using MeCab, which is an open source Japanese morphological analysis engine. We analyzed *yoku* and *yahari*, which are adverbs that both have a number of orthographic forms and were classified as 'unacceptable'. In order to explain the reasons behind this classification result, it is necessary to examine them from two viewpoints: lemma and orthographic form.

As mentioned in Section 3, the Japanese orthographic system has not yet been standardized. The Kōbunsho Manual is regarded as a sort of standard for writing Japanese official documents. The manual generally recommends hiragana notation for describing adverbs. As such, we assume that adverbs in the positive corpus set (White papers and Law documents, specifically) tend to conform to these standard guidelines. Hence, we will refer to the manual when analyzing the lemmas of 良く *yoku* and 矢張り *yahari* in our data.

4.1.1 良く yoku 'well'

The frequency distribution of the lemma 良く *yoku* is 300.5 PPM in the whole corpus, 62.9 PPM in the positive corpus and 404.0 PPM in the negative corpus. As the frequency in the negative corpus is significantly higher than in the whole corpus, and the frequency in the positive corpus significantly lower than in the whole corpus, it is classified as unacceptable.

However, when considering the frequency of the lemma 良く *yoku* within the positive corpus set, we find it ranks 15th most frequent and cannot thus be considered low. Usage of *yoku* may be considered unacceptable within the academic register depending on the semantic context it is used in. On conferring the Digital Daijisen Japanese dictionary (Matsumura et al. 1998) and other dictionaries, we assume that *yoku* has six different meanings:

1. Frequently, in quantity. Synonyms: しばしば *shibashiba* 'frequently', しきりに *shikirini* 'often'
2. Adequately, enough. Synonyms: 十分に *jūbun-ni* 'adequately', 徹底して *tettei-shite* 'thoroughly'
3. [Subjective use] With high ability. Synonyms: 上手く *umaku* 'well'
4. Highly, widely. Synonyms: 極めて *kiwamete* 'extremely', 非常に *hijōni* 'very', 高度に *kōdo-ni* 'to a high degree'
5. Completely. Synonyms*:* 十分に *jūbun-ni* 'thoroughly'
6. [Subjective use] Favorably. Synonyms: 好意をもって *kōi wo motte* 'with goodwill'

As meanings 1, 2, 4, and 5 of *yoku* can be used in objective contexts, and express the meaning of a high frequency, their use is permissible in academic documents, although paraphrasing them with other expressions is still preferable. At the current stage, the system cannot clearly distinguish between these meanings. Therefore, without a way of automatically disambiguating the exact sense used within the text, the system can only point to the objective uses of *yoku* as found in the positive corpus set, and these can serve as examples for the learner to reflect upon. For example, given the sentence この計画はよく考えられている *Kono kēkaku wa yoku kangaerarete iru* 'This plan is well thought out', it is possible to suggest the following alternative: この計画は十分に考えられている *Kono kēkaku wa jūbun ni kangaerarete iru* 'This plan is sufficiently thought out'.

As for the orthographic frequency, 良く *yoku* has a PPM of 0.9, while よく *yoku* has a PPM of 28.3 in the whole corpus. The figures are 3.4 PPM and 55.8 PPM respectively in the positive corpus set, and 30.1 PPM and 385.8 PPM respectively in the negative corpus set. *yoku* is classified as unacceptable in all cases. The hiragana orthographic form よく *yoku* is used 92.7% of the time in the negative corpus set, and 96.7% of the time in the whole corpus set. On the other hand よく *yoku* is only used 62.1% of the time in the positive corpus set. For comparison, the Kōbunsho Manual mandates the use of よく *yoku*.

These results show that academic documents use more mixed orthography even though the Kōbunsho Manual does not condone such use. For documents of corpora in which such standards do not apply, adverbs are also written using kanji (Chinese characters) and katakana.

On the other hand, an examination from the perspective of an expert in L2 Japanese language education classified these cases as acceptable within the academic register. It is therefore reasonable to assume that the choice of which orthographic form to use in academic writing depends on the intended meaning.

4.1.2 矢張り yahari 'as I thought'

As can be seen in Table 5, the lemma *yahari* can be written using 14 different orthographic forms. It should be noted that the last six all represent rare variations occurring less than ten times in the whole corpus set. In total, there are five word forms: *yahari* (やはり、矢張り), *yappari* (やっぱり, やつぱり, ヤッパリ), *yappa* (やっぱ), *yappashi* (やっぱし), and *yapa* (やぱ). As the system classifies at the orthographic form level, we are able to compare the results between different orthographic varieties of the same lemma.

The classification results for the lemma 矢張り give two orthographic forms (やはり, やっぱり) for which the verdict is unacceptable for the academic register, with the

Table 5. Orthographic form variations and their associated system classifications of the lemma 矢張り ordered according to their PPM in the whole corpus set.

Orthographic Base	System Verdict	Frequency Whole	Frequency Positive	Frequency Negative	PPM Whole	PPM Positive	PPM Negative
やはり	**UA**	**19,688**	**335**	**9,997**	**137.13**	**11.26**	**323.51**
やっぱり	**UA**	**11,130**	**48**	**6,357**	**77.52**	**1.61**	**205.72**
やっぱ	UK	1,502	10	1,210	10.46	0.34	39.16
矢張り	UK	107	16	19	0.75	0.54	0.61
やっぱし	UK	99	2	53	0.69	0.07	1.72
ヤッパリ	UK	49	13	32	0.34	0.44	1.04
やぱ	UK	36	-	35	0.25	0.00	1.13
やつぱり	UK	22	-	-	0.15	0.00	0.00
矢っ張り	UK	8	3	-	0.06	0.10	0.00
矢つ張り	UK	5	-	-	0.03	0.00	0.00
矢ッ張り	UK	4	-	-	0.03	0.00	0.00
やッぱり	UK	3	-	1	0.02	0.00	0.03
矢っ張	UK	1	-	-	0.01	0.00	0.00
矢つ張	UK	1	-	-	0.01	0.00	0.00
Total		32,655	427	17,704	227.45	14.35	572.92

Note: AC=acceptable, UA=unacceptable, UK=unknown rest classified as unknown. The lemma, as a whole, occurs at a rate of 572.92 PPM in the negative corpus set, 14.35 PPM in the positive corpus set and 227.45 PPM in the whole corpus. Thus, according to the system classification and large discrepancy between PPM rates, the adverb 矢張り is clearly not appropriate for use in the academic register.

The Kōbunsho Manual recommends the use of the hiragana やはり over the Chinese character (kanji) variant 矢張り. The orthographic variation やはり is used in 79.02% of the positive corpus set, 60.97 % of the whole corpus set, and 56.48% of the negative corpus set. While the Minutes of the Diet sub-corpus is a part of the negative corpus set, it is edited from transcribed speech data, a process which strictly follows the governmental guidelines and, as such, contains less orthographic variations than its sibling corpora of Yahoo! Q&A and Yahoo! Blogs.

In conclusion, we find that the hiragana variant of the orthographic form of the lemma *yahari* most commonly appears in the positive corpus set, which is also the form recommended by the Kōbunsho Manual. However, the system classified even this usage as unacceptable.

4.2 KOSOADO (こそあど) demonstrative words

The Japanese KOSOADO demonstratives have either *ko, so, a,* or *do* as the first syllable and are most commonly represented by the adverbs *kō*, *sō*, *ā*, and *dō*. These adverbs occur frequently in the whole corpus set and, with the exception of *ā*, also occur frequently in the positive corpus set. However, the system classifies them all as unacceptable for the academic register. Across the whole corpus set as well as the negative corpus set, *sō, dō* and *kō* are respectively the first, second and fourth most frequent adverbs. Even in the positive corpus set, *sō, dō* and *kō* are the eleventh, sixth, and ninth most frequent adverbs. The existence of *kono-yō-ni*, *sono-yō-ni*, and *dono-yō-ni*, formal counterparts to *kō*, *sō*, and *dō,* within the STJC is a possible reason for their relatively high rank. Finally, though less frequent than the rest, *ā* does appear in the negative corpus set, while its use within the positive corpus set was observed only within linguistic examples or language data in scientific articles and are otherwise absent from the main body of text. The inappropriate use of *ā* can also be found in the error annotations of the Natane learner corpus. The present system is able to advise learners that *ā* is unacceptable in academic documents.

4.2.1 こう kō

The lemma こう *kō* has no orthographic variation other than its hiragana form. As shown in Table 6, its frequency is much higher in the negative corpus set (687.7 PPM) than in the positive corpus set (97.8 PPM). As such, the lemma こう *kō* is classified as unacceptable for academic writing. However, if we look at the frequency of the compound adverbs, we find that the overall frequency in the positive corpus set is higher than in the negative corpus set. In order to uncover the reasons behind this shift in relative frequency between the positive and negative corpus sets, we analyze the usage of some of these compounds.

The compound adverb *kōshite* and compound noun modifier *kōshita* frequently appear in spoken language as well as written texts. These are paraphrased as *konoyōni* and *konoyōna* in the formal and academic texts as shown in the examples below. In addition, *kon'nani* and *kon'na* are casual expressions not found in the positive corpus set. Hence, it is possible to recommend the compound *konoyōni* for use in the academic register.

The following examples (2-4) show compound adverbs found in both the positive and negative corpus sets.

Ex. 2) たくさん問題をこなしているうちに，パターンが身につきます．こうして身についたパターンは，忘れることがなくなり，本当の学力につながりますよ． *Takusan mondai wo konashite iru uchi ni, patān ga mi ni tsukimasu. Kō-shi-te mi ni tsuita patān wa wasureru koto ga nakunari, hontō no gakuryoku ni tsunagarimasu yo.* 'You will never forget the patterns you have mastered this way, and this will lead to real learning.' (Yahoo! Q&A: OC12_05972)

Ex. 3) 今回のこうした不幸な事件を引き起こした大きな原因は、やはり外交上の問題があったと思うのです． *Konkai no kō-shi-ta fukō na jiken wo hikiokoshita ōkina genin wa, yahari gaikō-jō no mondai ga atta to omoun desu.* 'The main reason which caused such an unfortunate accident on this occasion is due to diplomatic problems.' (Minutes of the Diet: OM21_00010)

Ex. 4) 「今の世の中では，大学に進むのが当たり前だから」と答える親は極めて少ない。このように，親の側には，大学教育の役割について理想的なイメージがあるといえる。 *”Ima no yononaka dewa, daigaku ni susumu no ga atarimae dakara” to kotaeru oya wa kiwamete sukunai. Kono-yō-ni, oya no gawa ni wa, daigaku kyōiku no yakuwari ni tuite risō teki na imēji ga aru to ieru.* 'There are extremely few parents who would answer that "it is natural for their children to go to university in today's world". From this we can say that parents have an ideal image about the role of university education.' (White paper on public lifestyle: OW2X_00000)

Next, the *kōshite* in examples 5 and 6 is used as a direct deictic and not as a contextual demonstrative, making its use unsuitable for academic writing. *Kōshite* in example 5 indicates the way in which the speaker wants the food to be cut. Similarly, *kōshite* in example 6 indicates an ambiguous object, which cannot be determined from the context.

Ex. 5) 食べやすい大きさにこうしてちぎってください。 *Tabeyasui ōkisa ni kōshite chigitte kudasai.* 'Tear it into bite size pieces in this way, please.' (Nishida et al. (2003). Ryōri kyōji hatsuwa no kōzōkaiseki [Structural analysis of recipe instructions utterances]. *Proceedings of the 9th Annual Conference of the Association of Natural Language Processing*, 601-604.)

Ex. 6) 「もっとこうしてほしい」っていうのは彼に伝えた方がいいと思います。 *'Motto kōshite hoshi' tte iuno wa kare ni tsutaeta hou ga ii to omoimasu.* 'I think you should tell him "I want you to do it more in this way"' (Yahoo Q&A: OC09_06241)

We suggest that the deictic usage of *kō*—including in the compound adverbs as mentioned above—should be discouraged in academic writing. Consequently, we have to divide the usages of *kō*, including its compound variants, into those suitable for academic writing and those unsuitable based on these observations.

It is possible to say that *kōshita* and *kōshite* are acceptable because of their frequent use in the STJC corpus. We have to take into account both a word's current usage tendencies as well as its normative uses.

Table 6: Frequency of *ko* as part of compound expressions.

Adverb	Expression Type	Frequency Whole	Frequency Positive	Frequency Negative	PPM Whole	PPM Positive	PPM Negative
こう *kō*	SM	59,100	2,908	21,190	411.4	97.6	685.7
こうして *kōshite*	CM	6,407	260	590	44.6	8.7	19.1
こうした *kōshita*	CM	14,390	2,225	1,120	100.2	74.8	36.2
こういう *kōiu*	CM	18,788	41	12,096	130.8	1.8	391.4
こう言う *kōiu*	CM	390	3	139	2.7	0.1	4.5
こう云う *kōiu*	CM	34	0	7	0.2	0.0	0.2
このような *konoyōna*	CM	21,394	7,588	2,196	149.0	255.1	71.1
この様な *konoyōna*	CM	229	48	123	1.6	1.6	4.0
このように *konoyōni*	CM	11,406	3,308	1,406	79.4	111.2	45.5
この様に *konoyōni*	CM	58	19	20	0.4	0.6	0.6
このようにして *konoyōnishite*	CM	1,055	375	25	7.3	12.6	0.8
こうやって *kōyatte*	CM	784	2	268	5.5	0.07	8.7
こんな *kon'na*	SM	28,860	110	10,304	200.9	3.7	333.4

Note: SM=single morpheme, CM=compound morphemes

On the other hand, *kōyatte* and *kon'na* are scarcely found in the positive corpus set. Hence, we will add the former two compound words into the list of acceptable adverbs, but exclude the latter two compound words.

4.2.2 そう sō

The lemma *sō* has the highest frequency within all corpora. Additionally, it is also frequent in both the positive and negative corpus sets. However, our system classifies そう *sō* as unacceptable, even though its frequency is as high as that of こう *kō*. Next, comparing the compound words of *sō* and *kō*, we find that *kō* tends to occur more frequently in the positive corpus set, and *sō* in the negative corpus set. As can be seen from Table 7, the PPM value of *sō* is relatively higher for all the items in the negative corpus set.

Table 7: Frequency of *sō* showing the conjugated compound adverbs *sō-iu* and *sō-itta* used within the positive corpus set.

Adverb	**Expres-sion Type**	**Frequency Whole**	**Frequency Positive**	**Frequency Negative**	**PPM Whole**	**PPM Positive**	**PPM Negative**
そう *sō*	SM	130,824	42,449	2,521	910.6	84.7	1,373.7
そうして *sō-shi-te*	CM	2,898	604	21	20.2	0.7	19.5
そうした *sō-shi-ta*	CM	8,182	1,186	302	57.0	10.2	38.4
そういう *sō-iu*	CM	32,907	17,176	79	229.0	2.7	555.8
そう言う *sō-iu*	CM	1,244	293	5	8.7	0.2	9.5
そう云う *sō-iu*	CM	76	9	1	0.5	0.0	0.3
そのような *sono-yō-na*	CM	7,847	1,637	1,433	54.6	48.2	53.0
その様な *sono-yō-na*	CM	117	92	8	0.8	0.3	3.0
そのように *sono-yō-ni*	CM	1,915	607	78	13.3	2.6	19.6
その様に *sono-yō-ni*	CM	22	15	0	0.2	0.0	0.5
そのようにして *sono-yō-ni-shi-te*	CM	180	24	7	1.3	0.2	0.8
そうやって *sō-ya-tte*	CM	948	199	4	6.6	0.1	6.4
そんな *son'na*	SM	45,427	13,689	152	316.2	5.1	443.0

Note: SM=single morpheme, CM=compound morphemes

Although these idiomatic patterns are found in academic texts, they are relatively less frequent than words in the *kō* group. Words in the *sō* group are noted for their use in anaphoric expressions such as *'A ga B de aru bāi, ippō, A ga sō de nai bāi'* (In case A is B, and, on the other hand, in case A is not so).

Ex. 7) ペアが含まれるなら真、そうでないなら偽. *Pea ga fukumareru nara shin, sōdenainara gi.* 'If the pair is present, it is true, and if it is not so, then it is false.' (STJC: Murawaki, Y. & Kurohashi, S. (2007). Jōhō bunseki no tame no jutsugo kōzō wo mochiita dōteki ontorojī kōchiku [Construction of a dynamic ontology for information analysis using predicate structure]. In *Proceedings of the 13th Conference of the Association of Natural Language Processing* (pp. 867-870)).

Sōdenainara in this example paraphrases the previous expression *pea ga fukumareru,* which is its general function. On the other hand, substitution with *sonoyōdenainara* is unacceptable for reasons of syntax, although this may be substituted with the compound word *sonoyōni* which is more academic and formal than *sō* as a single morpheme. For example, it is possible to rewrite the expression *sō kaishaku dekiru* 'it is possible to interpret in that way' into the expression *sonoyōni kaishaku dekiru* in academic discourse. Hence, we are able to say that expressions such as *sō, sōitta* and *sōiu* are rather uncommon in academic discourse. The following examples extracted from the positive corpus set (examples 8 and 10) and the negative corpus set (example 9) illustrate these general observations.

Ex. 8) 最後の第九グループは，脂肪族化合物でアミノ基を有する場合の挙動を探ったものであるが，末端にある場合と，そうでない場合で多少反応性が異なり，場合によっては阻害性も発現する傾向がある．*Saigo no daikyū grūpu wa, shibōzoku kagōbutsu de aminoki wo yūsuru bāi no kyodō wo sagutta mono de aru ga, mattan ni aru bāi to, sō de nai bāi de, tashō hannōsē ga kotonari, bāi ni yotte wa sogaisē mo hatsugen suru keikō ga aru.* 'The last ninth group is an exploration of the behavior of possessing an amino group with an aliphatic compound. The reactivity is different in the occasion in the end and the occasion which is not so. The obstruction also tends to be manifested by a case.' (STJC: Watanabe O., & Nagai K.. (2000). Effect of Additive Reagents on the Reactivity of Lacquer Tree Paint. *Journal of the Chemical Society of Japan*, (3), 211-216.)

Ex. 9) 「おはようメール」がたまに届いたりしてました。ですが、最近はそういったメールが入ってきません。*Ohayō mēru ga tamani todoitari shite imashita. Desu ga, saikin wa sō-itta mēru ga haitte kimasen.* 'I had been occasionally receiving "good morning mails". But I have not received such mails recently.' (Yahoo Q&A: OC09_06528)

Ex. 10）上記のように極めて短期の需給見通し等の場合にはそのようなおそれがあるとみられる。*Jōki no yō ni kiwamete tan'ki no jikyū mitōshi nado no bāi ni-wa sono yōna osore ga aru to mirareru.* 'It seems risky in case of such an extremely_short-term supply and demand outlook as described above.' (Anti-trust white paper: OW3X_00120)

Having observed the corpora, usage of *so* in compound adverbs and in adjectival expressions such as *sōshite, sōshita, sōitta, sonoyōni, son'nani,* and *son'na* is extremely frequent in the negative corpus set compared to the positive corpus set. Therefore, we have to admit that anaphoric usage of *sō* is permitted in academic discourse. Even though the adverb *sō* is classified as unacceptable according to the system classification, the human evaluator classified the anaphoric usage of *sō* with compound words such as *so de areba* 'if it is so' and *so de nakereba* 'if it is not so' as acceptable. Consequently, we need to re-examine the system's focus on processing morphemes in isolation; expanding the unit size and taking into account the compound expressions is a promising avenue for increasing the accuracy of the system.

4.2.3 どう *dō*

The basic usage of the lemma *dō* is as the interrogative word of a sentence. It is ranked as the second most frequent in the whole corpus set, the sixth in the positive corpus set and the first in the negative corpus set (see Figure 2). With respect to PPM values, however, *dō* is most frequent in the negative corpus set; its frequency in the positive corpus set is significantly smaller than the norm to mark it as inappropriate for the academic register. The reason for this is clear if we analyze words co-occurring with *dō*: the frequency of the compound word *donoyōni* in the positive corpus set is higher than in the negative corpus set.

As shown in example 11, some adverbial *dō* appear as a part of constructions where they are followed by a verb, *ka* and a closing phrase such as *dō miru ka* or *dō kangaeru ka*. As shown in Table 8, the frequency of *ka dō ka* is highest in the positive corpus set, which covers approximately 67% of all instances of *dō*. However, the system classified it as unacceptable for the academic register. Instead of *ka dō ka, ka ina ka* is often used in academic fields, and the system has classified it as acceptable for the academic register. From this, we must admit *ka dō ka* as an alternative choice for learners, particularly since *ka dō ka* is relatively frequent in academic documents. We still recommend using *ka ina ka* as the first choice.

Ex. 11）「どう思うか教えて下さい。」 *Dō omou ka oshiete kudasai* 'Tell me what you think about it.' (Yahoo! Q&A: OC09_13396)

Ex. 12）「この意見に対してしてどう思います？」*Kono iken ni taishite dō omoimasu?* 'What do you think about this opinion?' (Yahoo! Q&A: OC09_14216)

Ex. 13) 業界ごとの市場規模を調べるにはどういう手段がありますか？*Gyōkai goto no shijō-kibo wo shiraberu ni wa dōiu shudan ga arimasu ka?* 'What means are availwable for researching the market size of each industry?' (Yahoo Q&A: OC03_02066)

Ex. 14) 明日初めてロンドンに行くのですがどういった服装でいけばいいですか？ *Ashita hajimete rondon ni iku no desu ga dōitta fukusō de ikeba ii desu ka?* 'I will visit London for the first time tomorrow, so what kind of clothes should I wear?' (Yahoo Q&A: OC13_02305)

Examples 11 and 12 illustrate the usage of *dō* in conversations. Example 13 illustrates the usage of the expression *dōiu*. These expressions also appear in the positive corpus set although they are not very frequent.

Table 8: Frequency of *dō* as a single morpheme and as part of compounds.

Adverb	**Expression Type**	**Frequency Whole**	**Frequency Positive**	**Frequency Negative**	**PPM Whole**	**PPM Positive**	**PPM Negative**
どう *dō*	SM	118,995	47,508	4,822	828.3	162.1	1,537.4
どうして*dōshite*	CM	17,304	5,405	127	120.4	4.3	174.9
どうした*dōshita*	CM	6,985	2,861	53	48.6	1.8	92.6
どういう *dōiu*	CM	11,158	5,488	153	77.7	5.1	177.6
どう言う *dōiu*	CM	112	86	0	0.8	0.0	2.8
どう云う *dōiu*	CM	31	2	0	0.2	0.0	0.0
どのような *donoyōna*	CM	9,680	2,052	2,723	67.4	91.5	66.4
どの様な *donoyōna*	CM	146	114	21	1.0	0.7	3.7
どのように *donoyōni*	CM	8,545	2,253	1,820	59.5	61.2	72.9
どの様に *donoyōni*	CM	161	130	13	1.12	0.4	4.2
どのようにして *donoyōnishite*	CM	867	183	91	6.0	3.1	5.9
どうやって *dōyatte*	CM	3,163	1,459	35	22.0	1.2	47.2
どんな *don'na*	CM	22,791	7,787	374	158.6	12.6	252.0
かどうか *ka dō ka*	CM	16,122	4,565	3,220	112.2	108.2	147.7
か否か *ka ina ka*	CM	2,728	168	1,411	19.0	47.4	5.4

Note: SM=single morpheme, CM=compound morphemes

Moreover, *dōitta* is used with the same meaning as that used in example 14. On the other hand, *donoyōni* is more formal and is the preferred substitution for *dō* in written academic Japanese. Lastly, with regards to *donoyōni*, we found that it is used more frequently in the positive rather than the negative corpus set.

In summation, we surveyed adverbs that include KOSODO demonstratives, and compared their respective frequencies in the positive and negative corpus sets. The results show that *KO* group adverbs are used more in the positive corpus. The frequency of *SO* group adverbs is comparatively lower in the positive corpus, although instances of anaphora usage seem to be permitted. *DO* group adverbs are relatively infrequent, except as the noun modifier *donoyōna* that was observed in the positive corpus set. Consequently, we must be careful when classifying cases of *sō* and *dō* usages; in most cases, *kō* is more acceptable. Also, we need to be especially aware of their compound usages, which are not immediately clear from the short-unit word morphological annotation of corpora.

5 Discussion and conclusion

This paper analyzed the distributional trends of adverbs within the corpora used for the automatic classification of register misuse with the goal of improving the classification rate of adverbs in the academic writing of L2 Japanese language learners. Register misuse was identified by comparing distributional trends between corpora representing the target register of academic writing and the opposing register of informal and spoken corpora against the backdrop of all the corpora combined. The results of classifying all adverbs extracted from the UniDic dictionary were compared against the classifications given by an L2 academic Japanese teaching expert. We summarize our results as follows:

I. Our original and general algorithm for classifying register misuse was found to work for the specific case of adverbs. By using the adverb list of UniDic, the system was able to find occurrences of 6,935 adverbs and classify each into adverbs into either acceptable, unacceptable or unknown groups. In total, it classified 121 adverbs as acceptable and 2,712 as unacceptable for use in academic writing.

II. We were able to clarify the tendencies of orthographic usage differences in each genre by taking into account the relationship between lemmas and their orthographic forms using UniDic. From this investigation, we also found that the existing orthographic standards in Japan are not comprehensive or widespread enough in their use. However, by basing our recommendations on the distributional tendencies of lemma within the positive corpus set, we were able to recommend the use of hiragana for most adverbs, with exceptions such as 最も *mottomo* and 極めて *kiwamete*, which are written using a mixture of kanji and hiragana.

III. We found some expressions that were classified as unacceptable but seem to be useful for academic writing when approached as compounds. Expressions such as *sō de nakereba* and *konoyōni* that contain demonstrative adverbs from the KOSODO word group are observed in academic writing. These kinds of compound adverbs should be either whitelisted or deferred to the classification dealing with longer word units at a deeper linguistic level.

On the other hand, we found the following problems with our classification approach:

IV. While the classification was based on orthographic forms, we also examined words from a lemma-centric viewpoint. From the perspective of learner writing in a setting without an official style guide, it is important to convey the fact that the same lemma may contain different orthographic forms, some acceptable and some unacceptable for use in the academic register. While the variation that exists within words that have multiple forms is often used to convey different nuances, especially within the more creative literary writing found in the Books sub-corpora of the BCCWJ, as the goals of the academic genre are to disseminate information in a standardized and clear way, this variation is undesirable and consequently, rarely employed in academic writing.

V. There are some considerable problems when using the present data. Firstly, several academic papers, predominantly from natural language processing journals, include examples of conversational language that skewed the results for some adverbs within the positive corpus set. In addition to the identification and deletion of these parts, the addition of more data from scientific and technical fields not related to language should also help alleviate this problem. As the treatment of collocations is related to the study of multi-morpheme compounds, further linguistic investigation along these lines is needed. At the same time, the treatment of orthographic variation under the lemma promises to be an interesting research area. As the Japanese orthography phenomenon is quite complicated for learners to grasp, we plan to consider supporting learners by introducing a new method focused on assisting orthographic choice.

VI. Because the classification algorithm compares the relative frequencies between the positive and negative corpus sets, adverbs having a high frequency within the positive corpus set may still be classified as 'unacceptable', although their frequency is quite high. We also found differences between the classifications of the L2 language education expert and the system. Further consideration of the algorithms in the system is needed.

VII. The current system classifies a few extremely low frequency adverbs as acceptable. However, it is possible to prevent this if we set a minimum threshold value for classification with the end goal being to reduce the number of false positives (i.e. classifying correct expressions as incorrect). Also, decreasing the number of unknown classifications by lowering the significance threshold of the chi-square test could be used to improve the recall of the system. This will be left to further research.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Number JP15K01114 (C) (PI: Abekawa Takeshi; 2015-2017) for the 'Study on effective information displays for error detection and revising within a Japanese writing support system'.

Literature

Abekawa, T., Yagi, Y., Hodošček, B., & Nishina, K. (2015). Improvement of Error Feedback Method in Japanese Composition Support System "NUTMEG". In *Proceedings of the 6th International Conference on Computer Assisted Systems For Teaching & Learning Japanese (CASTELJ)* (pp. 115–118). Honolulu, Hawaii.

Halliday, M. & Hasan, R. (1976). *Cohesion in English*. English Language Series. Longman.

Hodoščeк, 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. & Nishina, K. (2011). 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. (2012). 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

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., … Den, Y. (2013). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 1–27. doi:10.1007/s10579-013-9261-0

Matsumura, A., Ikegami, A., Kaneda, H., Sugizaki, K., Suzuki, T., Nakajima, T., … Hida, Y. (1998). *Digital Daijisen (Japanese dictionary)* (2015th ed.). Shogakukan.

Ministry of Education, Culture, Sport, Science and Technology-Japan, Section for the Japanese Language (Ed.). (2014). *Kōyōbun no kakiarawashikata no kijun [Criteria for the writing of official documents]* (Revised Edition). Ōkurashō Insatsukyoku.

Mizumoto, T. & Komachi, M. (2012). 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.

Ng, H. T., Wu, S. M., Briscoe, T., Hadiwinoto, C., Susanto, R. H., & Bryant, C. (2014, May). The CoNLL-2014 Shared Task on Grammatical Error Correction. In *CoNLL Shared Task* (pp. 1-14).

Ogiso, T., Ogura, H., Koiso, H., Miyauchi, S., Watanabe, R., & Den, Y. (2010). Keitaiso kaiseki jisho no benchimāku tesuto: IPAdic, NAIST-jdic, UniDic no janrubetsu seido hikaku [Benchmark tests on a morphological dictionary: Between-genre comparison for IPAdic, NAIST-jdic and UniDic]. In *Proceedings of the 16th Conference of the Association of Natural Language Processing* (pp. 326–329).

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.

Watanabe, S. (2010). Analysis of the use of adverbs in essays written by undergraduate international students and Japanese students. *Kyoto Sangyo University (Ronshū), 41*, 77–92. Retrieved from http://ci.nii.ac.jp/naid/110007523044/

Yagi, Y., Hodošček, B., Abekawa, T., & Nishina, K. (2014a). Nihongo sakubun suikō shien shisutemu "Natsumegu" ni okeru gakushū hyōka jikken no bunseki [Analysis of Learner Evaluations of Japanese Composition Support System "Nutmeg"]. In *International Conference on Japanese Language Education 2014.* Sydney: International Conference on Japanese Language Education 2014.

Yagi, Y., Hodošček, B., Abekawa, T., & Nishina, K. (2014b). Problems Found in a Learner Evaluation Experiment Using Japanese Composition Supporting System "Nutmeg". In *Dai rokkai kōpasu nihongo wākushoppu yokōshū [Proceedings of the 6th Workshop on Japanese Corpus Linguistics]* (pp. 229–232). NINJAL.

Yagi, Y., Hodošček, B., Abekawa, T., Nishina, K., & Murota, M. (2014). Analysis of Learner Responses to Errors Identified Using a Composition Support System. In *Proceedings of the Research Report of the JSET Conference* (Vol. 14, 5, pp. 151–156). Japan Society for Education Technology (JSET).

Internet resources

Hinoki Project (Natsume, Nutmeg, Natane): https://hinoki-project.org/ (25.8.2019)

MeCab Japanese morphological analyzer: (25.8.2019) https://taku910.github.io/mecab/

要旨 (Abstract in Japanese)

「日本語作文支援システム「ナツメグ」を利用した作文に 見られる副詞用法の適切さの分析」

ホドシチェック・ボル（大阪大学）
仁科喜久子（東京工業大学）
八木豊（株式会社ピコラボ）
阿辺川武（国立情報学研究所）

現在公開中の日本語学習者のため作文支援システム「ナツメグ」（http://hinoki-project.org/nutmeg/）は、自動添削を可能にすることを最終目的としている。本稿ではアカデミック日本語における学習者作文における副詞をレジスターの視点から観察し、科学技術論文を含む大規模な日本語コーパスを用いて、論文で用いる副詞と用いられない副詞を計量的に分けることを試みた。コーパスからはUniDicで定義された副詞2,919項目を抽出し、各副詞が論文としてレジスターに適切か否かを日本語教育あるいは日本語学の専門家による判定と、システム判定を比較した結果、専門家が科学技術レジスターで適切とした多くの副詞群が、システムでは「不適切」となった。その原因のひとつは、UniDicの副詞が短単位であるため、複数の単位からなる、短単位の複合形が抽出できないためと分かった。今後、複合形を含む副詞辞書の整備が必要であるものの、レジスター判定で学習者の論文作成を支援する可能性があることが明らかとなった。