Computer Assisted Language Instruction Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

Leveraging a Large Learner Corpus for Automatic Suggestion of Collocations for Learners of Japanese as a Second Language

Lis Pereira, Erlyn Manguilimotan, and Yuji Matsumoto

Abstract

One of the challenges of learning Japanese as a Second Language (JSL) is finding the appropriate word for a particular usage. To address this challenge, we developed a collocational aid designed to suggest more appropriate collocations in Japanese. In particular, we address the problem of generating and ranking noun and verb candidates for correcting potential collocation errors in the learners' text. Given a noun–verb construction as input, our system generates possible noun or verb correction candidates based on noun and verb corrections extracted from a large Japanese learner corpus. We use this corpus to investigate the learner's tendency to commit collocation errors, and to produce a smaller and more realistic set of candidates. After combining nouns or verbs with the generated candidates to form noun–verb pairs, the system uses the Weighted Dice coefficient as the association measure to filter out inappropriate noun–verb pairs and rank the proper collocations. We report the detailed evaluation and results on learner data. In addition, we show that our system statistically outperforms existing approaches to collocation error correction. Finally, we report a preliminary user study with JSL learners.

Keywords: Collocation; Automatic Error Correction; Language Learner; Japanese

Affiliation

Nara Institute of Science and Technology.
email: lis-k@is.naist.jp (corresponding author)

CALICO JOURNAL VOL 33.3 2016 311–333

doi : 10.1558/cj.v33i3.26444

1. Introduction

One of the challenges of learning a second language is finding the appropriate word for a particular usage. Learners of a second language do not yet have the extensive experience of native speakers to know which words are often combined to make natural expressions. The accurate use of words that commonly occur together, or simply collocations, is crucial for clear and effective communication similar to that of a native speaker. Lewis (2000) argues that "increasing the learners' collocational competence is the way to improve their language as a whole" (p. 14). In a separate study, Hill (2000, p. 62) explains that a student who uses collocations competently will be far more competent in communication than a student who does not.

The literature defines collocations in several ways. Lea and Runcie define collocations as combinations of words in a language to produce natural-sounding speech and writing (2002, p. vii). Smadja describes collocations as recurrent combinations of words that co-occur with higher possibility than random chance, and correspond to some arbitrary word usages (1993, p. 143). Regarding the semantic compositionality, collocations can be distinguished from idioms, which have meanings that are more opaque (Seretan, 2011, p. 23). Despite being less fixed compared to idioms, collocations would be regarded as less appropriate when one of the components is replaced by another word (Chang, Chang, Chen, & Liou, 2008; Shei, & Pain, 2000; Leacock, Chodorow, Gamon, & Tetreault, 2010, p. 65). In summary, for the purposes of our present study, collocations are word combinations that:

1. are arbitrary and recurrent;
2. co-occur more often than expected by chance; and
3. would be regarded as less appropriate when one of the components is replaced by another word.

Studies confirm that the correct use of collocations is challenging, even for advanced second language learners (Liu, 2002; Nesselhauf, 2003; Wible, Kuo, Tsao, Liu, & Lin, 2003). Unfortunately, the number of tools designed to target language learner collocation errors is limited. Most spell checkers and grammar checkers can help correct errors made by native speakers, but offer no assistance for non-native errors. Futagi, Deane, Chodorow, and Tetreault (2008) note that common aids for second language learners, namely, dictionaries and thesauri, are often of limited value when the learner does not know the appropriate collocation and must sort through a list of synonyms to find one that is contextually appropriate. Yi, Gao, and Dolan (2008) observe that language learners often use search engines to check if a phrase is commonly used by observing the number of hits returned. However, search engines are not designed to offer better alternative phrases than the learner's phrase (Park, Lank, Poupart, & Terry, 2008).

Concordancers seem to be an alternative to search engines, but they retrieve too much information because they usually allow only single-word queries. Too much information might distract and confuse the user (Chen, Huang, Huang, Chang, & Liou, 2014). Thus, a computer program that automatically identifies potential collocation errors and suggests corrections would be a more appropriate resource for second language learners.

Collocation error correction involves substitutions from potentially large sets of open-class words, i.e., nouns, adjectives, verbs, and adverbs (Leacock et al., 2014, p.70). To have good accuracy, systems for correcting collocation errors should have a strategy in restricting the number of correction candidates. Typical methods assume that the set of candidate corrections consists of all the words with similar meaning to the writer's word choice (see Futagi et al., 2008; Liu, Wible, & Tsao, 2009; and Park et al., 2008 for examples). These methods assume that the main cause of collocation errors is the confusion of sense relations (when learners misunderstand the semantic scope of a word). However, these approaches might fail to generate the correction for errors that involve other factors such as overgeneralization, shortage of collocation knowledge, and L1 interference. After restricting the number of candidates, another issue that needs to be addressed is how to rank those candidates before suggesting them as corrections to the user.

In this article, we describe the development of a collocational aid that targets potential noun–verb collocation errors made by learners of Japanese as a Second Language (JSL). Using a combination of a large learner corpus and statistical techniques, the tool has two unique features. First, it generates corrections by using noun and verb corrections extracted from a large, annotated Japanese language learner corpus. Because this corpus contains typical grammatical mistakes made by second language learners, our hypothesis is that the system can explore the learners' tendency to commit collocation errors, and produce smaller and more realistic sets of correction candidates. Second, it uses the Weighted Dice coefficient (Kitamura & Matsumoto, 1997) as a statistical association measure for ranking the collocation correction candidates. In our previous study (Pereira, 2013), we showed evidence of the effectiveness of this measure for ranking the collocation correction candidates in our task. We present an outline of the tool and evaluate its performance on learner data. Furthermore, we compare its performance with existing approaches to collocation error correction. Finally, we conduct a preliminary evaluation with JSL learners.

2. Existing Systems for Automated Collocation Error Correction

Several researchers have proposed useful English corpus-based tools for correcting collocation errors (Futagi et al., 2008; Liu et al., 2009; Park et al., 2008; Chang

et al., 2008; Wible et al., 2003; Dahlmeier & Ng, 2011). In a user study, Park et al. (2008) observed positive reactions from users when using their system. In another study, Liou et al. (2006) showed that the miscollocation aid proposed by Chang et al. (2008) can help learners improve their knowledge in collocations.

Collocation error correction is commonly performed by computing the differences in distribution between collocations and their noncollocational counterparts. The general approach consists of two steps:

1. **Candidate generation**: In this step, a set of alternative words to the learner's word choice is generated. This set is called the *confusion set*. Collocation candidates are then generated by substituting the learners' word choice with each word in the confusion set.
2. **Candidate ranking**: In this step, association measures are used to measure the association strength between the words in each candidate. The words within a collocation are expected to have higher association strength. These measures rely on the frequency information of word occurrence and co-occurrence in a corpus (Seretan, 2011, p. 31) and are commonly used to identify collocations (Seretan, 2011, p. 34).

To reduce the number of candidates in the confusion set, most existing works emphasize that collocation errors involve semantically related words in resources such as dictionaries or thesauri. Futagi et al. (2008) generated synonyms for each candidate string using WordNet and *Roget's Thesaurus*, and then used the rank ratio measure to rank them. Liu et al. (2009) also used WordNet to generate synonym candidates, but used Pointwise Mutual Information to rank the candidates. Similarly, Park et al. (2008) used WordNet to generate synonym candidates, but used co-occurrence frequency to rank the candidates. Chang et al. (2008), in contrast, emphasized L1 interference as the main cause of collocation errors. They used bilingual dictionaries to derive collocation candidates and used the log-likelihood measure to rank them. Wible et al. (2003) used a small, manually constructed list of verb–noun collocation errors and their corrections from a learner corpus. The drawback of these approaches is that they rely on resources of limited coverage, such as dictionaries, thesauri, or manually constructed databases to generate the candidates. Other studies have tried to offer better coverage by automatically deriving paraphrases from parallel corpora (Dahlmeier & Ng, 2011). However, similar to the approach used by Chang et al. (2008), it is necessary to identify the learner's first language and to have bilingual dictionaries and parallel corpora for every first language to extend the resulting system. Another drawback is that most of these systems rely only on well-formed English resources (except Wible et al., 2003) and do not actually take into account the learners' tendencies toward collocation errors.

3. System Design

In this section, we describe the design and function of our system. The tool focuses on suggestions for potential collocation errors in Japanese noun–verb constructions. In Japanese, these constructions have a case marker between the noun and the verb, which indicates the grammatical relations (e.g., subject, object, dative) of the complement noun phrase to the verb. We worked on three construction types listed in Table 1.

Table 1: The Three Construction Types in This Study

Construction type	Representation	Case particle	Grammatical function
Object-verb	noun *wo* verb (noun-を-verb)	*wo* (を)	Object
Subject-verb	noun *ga* verb (noun-が-verb)	*ga* (が)	Subject
Dative-verb	noun *ni* verb (noun-に-verb)	*ni* (に)	Dative (object/location)

The general architecture of the system is shown in Figure 1 with the example *夢をする (*yume wo suru*, lit. 'to do a dream') given as input. The system follows the general approach in the existing literature with the whole correction process consisting of the candidate generation and candidate ranking steps. Each step is detailed in the following sections.

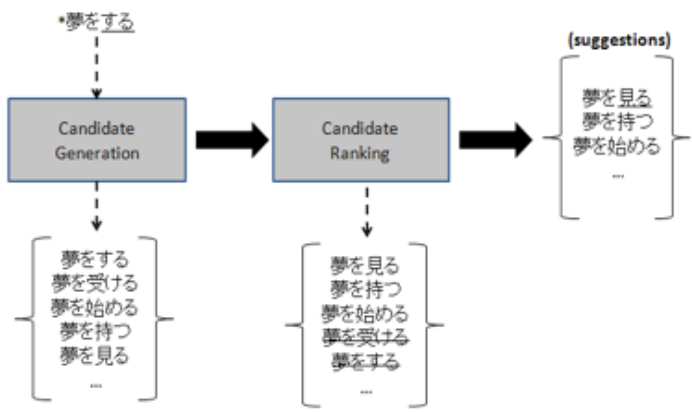

Figure 1: The processing stages of an erroneous collocation *夢をする (*yume wo suru*, lit. 'to do a dream'). The system generates alternative candidates for the verb する (*suru*, 'to do') and pairs them with the noun 夢 (*yume*, 'dream'). After filtering out the inappropriate pairs and ranking the remaining ones, the system offers collocation suggestions to the user. In the example, the verb 見る (*miru*, 'to see') is the appropriate verb to form the correct expression, 夢を見る (*yume wo miru*, 'to dream').

3.1. Candidate Generation

Given the noun–verb collocation input by a learner, the system first checks if it exists in the reference corpora. If not, the input is validated as a potential collocation error. Next, the system tries to find the more appropriate verb or noun collocates. The current system does not detect which component (noun or verb) is wrong in a noun–verb construction. Therefore, the learner must choose which component should be corrected by the system. For instance, if the learner types *夢をする (*yume wo suru*, lit. 'to do a dream'), the system flags this input as a potential collocation error. When the learner chooses to correct the verb, the system will search for the collocates of する (*suru*, 'to do') that are found in the confusion set. Similarly, if the learner types *光をつける (*hikari wo tsukeru*, lit. 'to attach light') and chooses to correct the noun, the system will search for the counterpart collocates of 光 (*hikari*, 'light') in the confusion set. The confusion set is constructed by using the Lang-8[1] learner corpus. This corpus was created by crawling the revision log of a language learning social networking service (SNS), Lang-8.[2] It contains journal entries written by language learners with different nationalities, which were manually corrected by native speakers. Hence, it contains typical grammatical mistakes made by second language learners. The biggest benefit of using such data is that we can obtain large-scale pairs of learners' sentences and corrections made by native speakers of Japanese. Although most Lang-8 members are not language experts, native speakers are generally good at telling what naturally sounds right and authentic to them (Cho, 2013). Lang-8 provides information about the L1 of the user for most of the sentences in our data set. However, we did not use this information in our experiments. The learners of Japanese in the data are distributed across 71 different nationalities. The top L1 of users in our experiments are listed in Table 2. Lang-8 does not provide information about the proficiency level of the users.

For generating candidates, we used one year's worth of data (from 2010), which contained 1,288,934 pairs of learner's sentences and their corrections. We extracted all of the possible noun and verb corrections for each of the noun–verb constructions in the corpus[3] (Table 3).

Table 4 shows some of the extracted examples. The confusion set of the verb する (*suru*, 'to do') includes verbs such as 受ける (*ukeru*, 'to accept'), which does not necessarily have a similar meaning to する (*suru*,'to do'). The confusion set means that in the corpus, する (*suru*, 'to do') was corrected to either one of those verbs. For example, when the learner writes the verb する (*suru*,'to do') in a noun–verb construction, he or she might actually mean to write one of the other verbs in the confusion set, such as 受ける (*ukeru*, 'to accept'), 始める (*hajimeru*, 'to begin'), or 見る (*miru*, 'to see').

Even if the learner's input is not flagged as a potential error, it will undergo the correction process because better collocations might exist. In case the

learner types only a noun or only a verb, the system will suggest collocations containing words that strongly collocate with this input.

Table 2: Top L1s in Lang-8 data

L1	Percentage
English	30.2
Unknown	27.2
Simplified Chinese	16.0
Traditional Chinese	12.5
Korean	2.1
Russian	1.4
Cantonese	1.1
Spanish	1.0
German	0.8
French	0.8
Brazilian Portuguese	0.8
Vietnamese	0.6
Indonesian	0.6
Italian	0.6
Thai	0.6

Note. Unknown represents the percentage of sentences where the users did not inform their L1.

Table 3: Specification of the Data Used in the Candidate Generation Step

Data	Lang-8
Size	*1,288,934* pairs of learner's sentences and corrections (37.5M tokens)
# noun-*wo*-verb pairs	163,880
# noun-*ga*-verb pairs	63,312
# noun-*ni*-verb pairs	25,787
# unique nouns	38,999
# unique verbs	16,086

Table 4: Confusion Set for the Words する (*suru*, 'to do') and 光 (*hikari*, 'light')

	Input	**Confusion Set**				
Word	する	受ける	始める	見る	書く	言う
Reading	*suru*	*ukeru*	*hajimeru*	*miru*	*kaku*	*iu*
Meaning	'do'	'accept'	'begin'	'see'	'write'	'say'
Word	光	電気	物体	景色	明かり	周り
Reading	*hikari*	*denki*	*buttai*	*keshiki*	*akari*	*mawari*
Meaning	'light'	'electricity'	'object'	'view'	'light'	'surroundings'

3.2. Candidate Ranking

After the system has obtained a list of candidate nouns or verbs, it will generate candidate pairs by substituting each of the candidate nouns/verbs from the confusion set into the input phrase. The candidate pairs are then ranked by using the Weighted Dice coefficient (Kitamura & Matsumoto, 1997) applied to a large reference corpus. The reference corpus we used is the Balanced Corpus of Contemporary Written Japanese or BCCWJ (Maekawa et al., 2014). The portions of BCCWJ used in our experiments included magazine, newspaper, textbook, and blog data. In addition, we included 1,288,934 sentences from the corrected sentences of Lang-8 (year 2010 data) to the reference corpus. The data is described in Table 5.

An example of candidate ranking is as follows: Given the input phrase *夢をする (*yume wo suru*, 'lit. to do a dream'), the system generates candidate pairs such as 夢を受ける (*yume wo ukeru*, lit. 'to accept a dream'), 夢を始める (*yume wo hajimeru*, 'to begin a dream'), and 夢を見る (*yume wo miru*, 'to dream'). In this case 夢を見る (*yume wo miru*, 'to dream') would be the most appropriate correction. In the same way, given the input phrase *光をつける (*hikari wo tsukeru*, 'lit. to attach light'), the system generates candidate pairs such as 電気をつける (*denki wo tsukeru*, 'to turn the light on'), 物体をつける (*buttai wo tsukeru*, lit. 'to attach an object'), and 景色をつける (*keshiki wo tsukeru*, lit. 'to attach a view'). In this case, 気をつける (*denki wo tsukeru*, 'to turn the light on') would be the most appropriate correction.

Aside from the collocations, example uses for each phrase suggestion are displayed. The example shows the phrase within the context of a sentence. Showing phrases within a context can be crucial for users to determine which phrase is most appropriate (Park et al., 2008).

Table 5: Specification of the Data Used in the Candidate Ranking Step

Data	BCCWJ	Lang-8
Size	*871,184 sentences (54.81M tokens)*	*1,288,934 sentences (corrected sentences) (14M tokens)*
# noun-*wo*-verb pairs	194,036	163,880
# noun-*ga*-verb pairs	216,755	63,312
# noun-*ni*-verb pairs	300,362	25,787
# unique nouns	43,243	38,999
# unique verbs	18,212	16,086

4. Comparison with Other Methods for Generating Collocation Candidates

Our proposed method for generating collocation candidates is compared with other existing approaches. These approaches are the *thesaurus-based word similarity* and *distributional similarity*. Systems developed using these methods assume that most collocation errors are caused by confusion of sense relations.

4.1. Thesaurus-Based Word Similarity

The basic approach of this method is to consider two words to be similar if they are near each other in the thesaurus hierarchy. In other words, a path within a predefined threshold length exists. For example, given the verb する (*suru*, 'to do'), we will obtain a list of candidate words such as さす (*sasu*, 'to make someone do') and し出す (*shidasu*, 'to begin to do'). In the same way, for the noun 光 (*hikari*, 'light'), we will obtain a list of candidate words such as きらめき (*kirameki*, 'glitter'), 閃光 (*senkou*, 'flash'), and 螢光 (*keikou*, 'fluorescence').

In our experiments, we used the Bunrui Goi Hyo Thesaurus, a Japanese thesaurus composed of 87,743 words that are classified into 32,636 unique semantic classes.

4.2. Distributional Similarity

Hand-built thesauri do not cover many words, phrases, and semantic connections, especially for verbs and adjectives, leading to low recall. Unlike thesaurus-based methods, distributional similarity models give better coverage. They can automatically extract synonyms and other relations from the corpora. Moreover, they can be used for automatic thesaurus generation for automatically populating or augmenting online thesauri (Jurafsky & Martin, 2009, p. 692). The basic idea of this method is that two words are considered similar if they have similar word contexts (Harris, 1954). In our work, context is defined by these grammatical dependency relations: object–verb, subject–verb, or dative–verb.

To compute similarity we use the Jenson-Shannon divergence (Lee, 1999). The corpora used are BCCWJ and the corrected sentences of Lang-8 (same data described in Table 5). For example, by using the Jenson-Shannon divergence, verbs similar to する (*suru*, 'to do') would be 終える (*oeru*, 'to finish'), 始める (*hajimeru*, 'to begin'), and 続ける (*tsuzukeru*, 'to continue') because they share similar nouns in their grammatical context. In the same way, nouns similar to 光 (*hikari*, 'light') would be 紫外線 (*segaisen*, 'ultraviolet rays'), 太陽 (*taiyou*, 'sun'), and 光沢 (*koutaku*, 'brilliance') because they share similar verbs in their grammatical context.

5. Selection of an Association Measure for Ranking Collocation Candidates

There are no defined criteria for choosing one particular association measure to apply in a specific task. The suitability of an association depends on various factors of the settings where the experiment takes place (e.g., language or domain). A common strategy is to compare the individual merits of association measures (Seretan, 2011, p. 43). In our previous study (Pereira, 2013), we evaluated four different association measuresfor ranking collocation candidates: Pointwise Mutual Information (Church & Hanks, 1990), log-likelihood ratio (Dunning, 1993), Dice coefficient (Smadja, McKeown, & Hatzivassiloglou, 1996), and Weighted Dice coefficient (Kitamura & Matsumoto, 1997). The Weighted Dice obtained the highest performance in our task, and we adopted it for ranking collocation candidates.

6. Evaluation on Learner Data

We conducted an automatic evaluation to test our system's performance on real examples of learner data and to compare it with the existing approaches. The evaluation was divided into two parts: a *verb suggestion task* and a *noun suggestion task*. For the verb suggestion task, our goal was to evaluate the performance of our system on learners' noun-*wo*-verb, noun-*ga*-verb, or noun-*ni*-verb constructions where the verb was misused. Likewise, for the noun suggestion task, our goal was to evaluate the performance of our system on learners' noun-*wo*-verb, noun-*ga*-verb, or noun-*ni*-verb constructions where the noun was misused. The system was evaluated on a test set constructed from Lang-8. The construction of this test set is detailed in the following section.

6.1. Test Set Construction

We used one year's worth of data (from 2011) from Lang-8 for constructing our test set. The data contained 2,246,059 pairs of learners' sentences and their corrections (26 million tokens) given by native speakers. For the verb suggestion task, we extracted all of the noun-*wo*-verb, noun-*ga*-verb, and noun-*ni*-verb pairs with incorrect verbs and their corresponding corrections. Similarly, for the noun suggestion task, we extracted all of the noun-*wo*-verb, noun-*ga*-verb, and noun-*ni*-verb pairs with incorrect nouns and their corrections.

Table 6: Statistics of the Extracted Pairs From Lang-8 (2011 Data)

	Total	f ≥5	5>f≥3	f =2	f =1
# noun-*wo*-verb pairs	60,916	1,197	3,092	7,636	48,991
# noun-*ga*-verb pairs	38,377	582	1,767	4,717	31,311
# noun-*ni*-verb pairs	28,055	329	1,217	3,349	23,160

Note:f stands for frequency.

Table 6 shows the statistics of the extracted pairs. These pairs of corrections are noun–verb expressions where native speakers had corrected either the noun or the verb. In the correction pair *夢をする→夢を見る, the verb する (*suru*) was corrected to the verb 見る (*miru*). We then sorted these corrections by their frequency *f* in the corpus. For instance, in the correction pair *夢をする→夢を見る, する (*suru*) was corrected to 見る (*miru*) 48 times (*f* = 48). Similarly, in the correction pair *光をつける→電気をつける, the noun 光 (*hikari*) was corrected to 電気 (*denki*) 19 times (*f* = 19). One problem of this selection criterion is that there are cases wherein the learner's construction sounds more acceptable than its correction. For example, cases such as 日記を書く (*nikki wo kaku*, 'to write diary') and its correction 日記を書ける (*nikki wo kakeru*, 'be able to write a diary'). 日記を書く (*nikki wo kaku*) sounds more correct than 日記を書ける (*nikki wo kakeru*). However, in the corpus it was corrected due to some contextual information. One example for that case is as follows:

Learner's sentence: 最近ちょっと忙しいから、**日記を書きません**.
Saikin chotto isogashii kara, ***nikki wo kakimasen****.*
I have been a bit busy lately, so **I don't write my diary.**

Sentence correction: 最近ちょっと忙しいから、**日記を書けません**.
Saikin chotto isogashii kara, ***nikki wo kakemasen****.*
I have been a bit busy lately, so **I can't write my diary**.

For our application, there was a need to filter out such contextually induced corrections because we were only considering the noun, particle, and verb that the learner wrote. To solve this problem, we included in the test set the top high frequency (*f*≥5) pairs (670 in total, approximately 200 samples for each of the three construction types) and asked a professional Japanese annotator to manually validate them. Each correction pair was checked by the annotator to determine whether or not it was a collocation error and whether or not the correction was appropriate. Only the correction pairs judged as collocation errors and with appropriate corrections were included in the test set. Regarding the corrections, the professional annotator and the annotators in Lang-8 agreed in 99% of the cases. Table 7 summarizes the test set obtained after annotation.[4] This test set was used for evaluation in our experiments.

Table 7: Test Set Obtained After Manual Annotation

	# correction pairs
Verb suggestion	**317**
Noun suggestion	**213**

6.2. Evaluation Metrics

We compared the collocation candidates generated and ranked by the system with the human correction assigned in the Lang-8 data. A match was counted as a true positive (tp). A false negative (fn) occurred when the system could not offer any suggestion. The metrics we used for the evaluation were precision, recall, and the mean reciprocal rank (MRR).

We reported precision at rank *k*, which corresponds to how often the correction was ranked in the top *k* suggestions. For instance, precision at rank 1 (P@1) computed how often the correction was ranked in first place by the system, and precision at rank 5 (P@5) computed how often the correction was ranked within the top five suggestions by the system.

The recall measures how often the system could offer the correction in the collocation suggestion list. In other words, it computed how often the correction was found anywhere in the collocation suggestion list. The collocation suggestion list had the size of the threshold we stipulated (270). Recall was computed using the following formula:

$$\frac{tp}{tp + fn} \qquad (1)$$

Because the system returned a ranked list of suggestions, it makes sense to award partial credit for cases wherein the system made a correct suggestion but did not rank it first. To address this, we used the MRR, a standard metric used for evaluating ranked retrieval systems (Voorhees, 1999). The MRR values range from 0 to 1, with 1 being the best possible value. This metric was used to assess whether or not the suggestion list contained the correction and how far up it was in the list. MRR was calculated as follows:

$$MRR = \frac{1}{N}\sum_{i=1}^{N}\frac{1}{rank(i)} \qquad (2)$$

where N is the size of the test set. If the system did not return the correction for a test instance, we set $1/rank(i)$ to zero.

6.3. Results

To explore the tendency of the results, we first evaluated the performance of our system on the verb-suggestion and noun-suggestion tasks in object–verb constructions. As Table 5 and Table 6 show, this is the most common noun–verb construction type in the learners' writing. Table 8 shows the

results for three models made by combining different candidate generation methods with the Weighted Dice coefficient. In this table, *Thesaurus+WD* refers to the model that used a thesaurus for generating candidates and the Weighted Dice coefficient (WD) for computing the association strength and ranking the candidates. The *DS+WD* model used distributional similarity for generating candidates. *CS Lang-8+WD*, our proposed model, generated candidates using the confusion set (CS) based on correction pairs from Lang-8.

Table 8 reports the precision, recall, and MRR for verb and noun suggestion tasks for all models. The model that used a thesaurus (*Thesaurus+WD*) achieved the highest precision rate among the other models. However, it had the lowest recall. This model could make good suggestions for cases wherein the learner's word choice and the correction suggested by the Lang-8 data had similar meaning (i.e., words are near each other in the thesaurus hierarchy). Some examples are shown in Table 9. However, for cases wherein the learner's word choice and the correction suggested by the Lang-8 data did not have a similar meaning, the *Thesaurus+WD* model could not generate the correct candidate in the candidate generation stage. Consequently, it could not suggest a correction resulting in a low recall. The recall improved greatly with the model that used distributional similarity (*DS+WD*). This means that the correct candidate could be generated for many cases. However, the precision rate decreased because the correction obtained a low rank in the collocation suggestion list. *CS Lang-8+WD*, our proposed model, achieved the highest MRR and values. In most test-set instances, this model suggested the correction in first or second place as indicated by the MRR values. By using a large learner corpus to generate the correction candidates, the system included more collocation choices that learners tend to choose. Because a wide range of factors cause such errors, it is difficult to capture all the error patterns using either thesaurus-based methods or distributional similarity methods.

Table 10 shows examples in which the model that used a thesaurus could not suggest any correction because the learners' word choice and the correction suggested in the Lang-8 data did not have similar meanings. Alternatively, the other models suggested the correction among the ten best-ranked candidates. We can also see that our proposed model obtained higher precision. It generated the correction with higher rank compared to the model that used distributional similarity. Using a two-tailed t-test with a confidence interval of 99%, we measured the statistical significance. We found that for both verb- and noun-suggestion tasks our *CS Lang-8+WD* model performed significantly better than the other two models.

Table 8: The Precision, Recall, and MRR of Different Models Applied to Object–Verb Constructions

	Verb suggestion				Noun suggestion			
System	*P@1*	*P@5*	*Recall*	*MRR*	*P@1*	*P@5*	*Recall*	*MRR*
Thesaurus+WD	**0.94**	**1.00**	0.11	0.11	**0.84**	**1.00**	0.24	0.22
DS+WD	0.54	0.80	0.73	0.49	0.38	0.67	0.47	0.23
CS Lang8+WD	0.63	0.89	**0.95**	**0.72**	0.63	0.97	**0.86**	**0.66**

Note: WD stands for Weighted Dice, DS stands for Distributional Similarity, and CS Lang8 stands for confusion set from the Lang-8 corpus.

Table 9: Rank of Correct Verb Given by the Model That Used a Thesaurus for Generating the Correction Candidates

	Misused noun+verb			Correction			Rank of correction
Japanese	薬	を	*食べる	薬	を	飲む	1
Reading	*kusuri*	*wo*	*taberu*	*kusuri*	*wo*	*nomu*	
Meaning	medicine		eat	medicine		drink	
Japanese	家	を	*出かける	家	を	出る	1
Reading	*ie*	*wo*	*dekakeru*	*ie*	*wo*	*deru*	
Meaning	house		go out	house		leave	
Japanese	日本語	を	*独学する	日本語	を	勉強する	1
Reading	*Nihongo*	*wo*	*dokugaku suru*	*Nihongo*	*wo*	*benkyō suru*	
Meaning	Japanese		self-taught	Japanese		study	

A similar phenomenon occurred for the noun-suggestion task. Table 11 shows some examples of the ranking for the corrections assigned by all three models.

We applied our *CS Lang-8+WD* model to subject–verb (noun-*ga*-verb) and dative–verb constructions as well. Table 12 and Table 13 summarize the results for verb and noun suggestions. For both subject–verb and dative–verb constructions, the system obtained high recall and MRR values.

Table 10: Rank of Correct Verb Given by the Models That Use Distributional Similarity (*DS+WD*) and Confusion Set Derived From Lang-8 (*CS Lang-8+WD*)

Misused noun+verb				Correction				Rank of correction		
								Thesaurus+ WD	***DS + WD***	***CS Lang8 + WD***
Japanese	スピーチ	を	*言う	スピーチ	を	する		–	9	4
Reading	*supiichi*	*wo*	*iu*	*supiichi*	*wo*	*suru*				
Meaning	speech		tell	speech		do				
Japanese	試験	を	*参加する	試験	を	受ける		–	1	1
Reading	*shiken*	*wo*	*sanka suru*	*shiken*	*wo*	*ukeru*				
Meaning	exam		attend	exam		accept				
Japanese	夢	を	*する	夢	を	見る		–	4	1
Reading	*yume*	*wo*	*suru*	*yume*	*wo*	*miru*				
Meaning	dream		do	dream		see				

Note: The "–" indicates cases where the system was not able to generate the correct candidate.

Table 11: Rank of Correct Noun Given by the Models That Used aThesaurus (*Thesaurus+WD*), Distributional Similarity (*DS+WD*), and Confusion Set Derived from Lang-8 (*CS Lang-8+WD*)

	Misusednoun+verb			**Correction**			**Rank of correction**		
							Thesaurus* + *WD	***DS* + *WD***	***CS Lang8* + *WD***
Japanese	*新聞	を	聞く	ニュース	を	聞く			
Reading	*shinbun*	*wo*	*kiku*	*nyūsu*	*wo*	*kiku*	1	–	1
Meaning	newspaper		listen	news		listen			
Japanese	*光	を	つける	電気	を	つける			
Reading	*hikari*	*wo*	*tsukeru*	*denki*	*wo*	*tsukeru*	–	-	1
Meaning	light		attach	electricity		attach			
Japanese	*自身	を	持つ	自信	を	持つ			
Reading	*jishin*	*wo*	*motsu*	*jishin*	*wo*	*motsu*	–	2	1
Meaning	own		carry	confidence		carry			

Note: The "–" indicates cases where the system was not able to generate the correct candidate.

Table 12: The Precision, Recall, and MRR of the Confusion Set from Lang-8 and Weighted Dice Measure Combinations Applied to Subject–Verb Constructions

	Verb suggestion				Noun suggestion			
Model	*P@1*	*P@5*	*Recall*	*MRR*	*P@1*	*P@5*	*Recall*	*MRR*
CS Lang8+WD	0.63	0.94	1.00	0.77	0.54	0.73	0.80	0.55

Note: CS Lang8 stands for confusion set from the Lang-8 corpus and WD stands for Weighted Dice.

Table 13: The Precision, Recall, and MRR of the Confusion Set from Lang-8 and Weighted Dice Measure Combinations Applied to Dative–Verb Constructions

	Verb suggestion				**Noun suggestion**			
Model	*P@1*	*P@5*	*Recall*	*MRR*	*P@1*	*P@5*	*Recall*	*MRR*
CS Lang-8+WD	0.29	0.52	1.00	0.65	0.34	0.61	0.59	0.33

Note: CS Lang8 stands for confusion set from the Lang-8 corpus and WD stands for Weighted Dice.

6.4. System Limitations

The limitations of the system can be categorized into two main types (Table 14):

1. For some cases, our system failed to generate the adequate collocation candidate if the learner's word choice and its correction were not observed in the learner corpus. For instance, there is no occurrence in the learner corpus where the noun 成熟 (*seijuku*, 'maturity') was corrected to the noun 大人 (*otona*, 'adult'). Therefore, the system cannot generate 大人 (*otona*) as a correction candidate. Additional learner-annotated corpora might help solve this problem. Alternatively, one can have a weighted combination of the confusion sets generated from the three methods we evaluated: (a) thesaurus-based method, (b) distributional similarity, and (c) confusion set generated from the learner corpus.
3. Even if the adequate collocate candidate can be generated, there are cases wherein the system fails to offer correct suggestions because the correct candidates paired with nouns/verbs cannot be found in the reference corpora we used for ranking the candidates. Incorporating larger corpora from different domains might help overcome this limitation.

7. Preliminary User Study of the System

We conducted a preliminary evaluation with JSL learners to gather their feedback on using our system. The results gave us insights about the usefulness of the system and about the possible interesting evaluations that should be carried out in the future.

Table 14: Example of Cases Where the System Fails to Offer the Correct Collocation

Type		Misused noun+verb				Correction			
No correction observed in the learner corpus	Japanese	*成熟	に	なる		大人	に	なる	
	Reading	*seijuku*	*ni*	*naru*		*otona*	*ni*	*naru*	
	Meaning	maturity		become		adult		become	
No occurrence of the correct collocation in the reference corpus	Japanese	問題	に	*会う		問題	に	出会う	
	Reading	*mondai*	ni	au		*mondai*	ni	deau	
	Meaning	problem		encounter		problem		encounter	

7.1. Participants

In this study, ten JSL learners, all graduate students from the same institution as the authors were invited to participate. Participants' ages ranged from 24 to 33 years, and the average age was 27.5. Among the respondents, two were female and eight were male, and they had different language backgrounds (Chinese, Indonesian, Tagalog, Swahili, Spanish, and Basque). Regarding

their proficiency level, three were beginners, three were intermediate, and four were advanced learners, based on the Japanese-language proficiency test[5] certificate level they previously obtained. All participants were regular computer users.

7.2. Procedure

A collocation test was designed to examine whether or not the tool could help JSL learners find proper Japanese collocations. This included 12 Japanese sentences from the Lang-8 learner corpus and from another small annotated Japanese-learner corpus, NAIST Goyo Corpus (Oyama, Komachi, & Matsumoto, 2013). The sentences and their corrections were further validated by a professional Japanese teacher. Each sentence contained one noun–verb collocation error made by JSL learners. The participants were asked to use the system to identify and correct the errors.

After performing the task, a survey questionnaire was also administered to better understand the learners' impressions of the tool. The questionnaire contained 43 questions answerable by a 7-point Likert-scale (with 7 labeled "strongly agree" and 1 labeled "strongly disagree"). The second part of the questionnaire contained seven open-ended questions. Our survey questionnaire inquired on the difficulty of Japanese collocations, the usefulness of the system and the quality of the retrieved data.

7.3. Results on the Collocation Test and Survey Questionnaire

The participants successfully found corrections for an average of 8.9 (SD = 1.6) out of 12 cases. The average time participants took to complete the task was 29 (SD = 16) minutes. The average score of beginner and intermediate learners was 9.6 (SD = 0.5). They scored higher than advanced learners, who obtained an average score of 8.2 (SD = 2.0). Analyzing the log files of their interactions with the system, we observed that intermediate and beginner learners used the system 40% more (on average) than the advanced learners. We noticed that two advanced learners tried to answer the questions without using the system when they felt confident about the answer, whereas the beginners and intermediate learners used the system for all sentences and obtained higher scores. The participants had difficulty in correcting two particular long sentences in the test. The noun–verb collocations in the sentences alone were not incorrect, but they were not appropriate in the context they appeared. They had difficulty in finding sentence examples close to the meanings of these sentences in the test. Although we need to evaluate this tool with a larger number of users, we observed that it was effective in helping the learners choose the proper collocations.

In the questionnaire administered, all participants acknowledged their difficulty in using Japanese collocations appropriately and stated that the software aids they used did not provide enough information about the meaning of Japanese phrases nor help in correcting errors in Japanese expressions. Their attitude toward the usefulness of the system was mostly positive, and they thought it was useful to help choose the proper way to use Japanese expressions. Regarding the quality of the retrieved data, the participants expressed satisfaction with the retrieved collocations, with an average score of 6.5 (SD = 0.7). They also expressed satisfaction with the ranking of the collocations presented, with an average score of 5.8 (SD = 0.6). Additionally, they reported that the sentence examples further helped them understand in which context an expression should be used. However, some participants expressed dissatisfaction with the complexity of some example sentences: some of the sentences were too long and difficult to understand.

In the second part of the questionnaire, some participants stated that the system could be helpful when learning new words and when one does not know which word combinations to use. They also suggested that the tool could be useful for teachers too when giving feedback to their students about the common errors they make and when providing alternative ways of expressing the same idea.

8. Conclusions

In this article, we presented a collocational aid system for learners of Japanese as a Second Language. Using noun and verb corrections extracted from a large annotated Japanese-learner corpus, our system can better explore the learners' tendency to commit collocation errors compared to standard methods that generate candidates based on the semantic relation of words.

Our system received positive feedback from JSL learners in a preliminary user study. The system can be used independently as a phrase dictionary, or it can be integrated into the writing component of some bigger CALL systems. For example, the system can be used by teachers as a way to obtain better understanding about learners' errors and help them provide better feedback to students.

One limitation of our experiments is the limited contextual information (only the noun, particle, and verb written by the learner). In the future, to verify our approach and to improve on our current results, we plan to consider a wider context size and other types of constructions (e.g., adjective–noun, adverb–verb, etc.). We also plan to conduct a more extensive evaluation with JSL learners to verify its usefulness in practical learning scenarios.

Notes

1. http://cl.naist.jp/nldata/lang-8/
2. http://lang-8.com/
3. All corpora used in our experiments were lemmatized, so we considered only the base form of the verb. All noun–verb pairs were extracted by using the Japanese dependency parser Cabocha (Kudo & Matsumoto, 2002).
4. Data is publicly available on: https://drive.google.com/file/d/0BysB3EnjLYH4VVFiZU01dGxXck0/edit?usp=sharing
5. http://www.jlpt.jp/e/index.html

About the Authors

Lis Pereira completed her PhD at Nara Institute of Science and Technology in 2016 working on how to address content word choice errors in L2 Japanese.

Erlyn Manguilimotan is a PhD candidate in the Graduate School of Information Science at Nara Institute of Science and Technology. She is working on part-of-speech and syntactic analysis of the Tagalog language.

Yuji Matsumoto is currently a Professor of Information Science at the Nara Institute of Science and Technology. He received his MSc and PhD degrees in information science from Kyoto University in 1979 and 1989 respectively. He joined the Machine Inference Section of the Electrotechnical Laboratory in 1979. He has been an academic visitor at the Imperial College of Science and Technology, a deputy chief of the First Laboratory at ICOT, and an associate professor at Kyoto University. His main research interests are natural language understanding and machine learning.

References

Chang, Y. C., Chang, J. S., Chen, H. J., & Liou, H. C. (2008). An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology. *Computer Assisted Language Learning, 21*(3), 283–299. Retrieved from: http://dx.doi.org/10.1080/09588220802090337

Chen, M.-H., Huang, C.-C., Huang, S.-T., Chang, J.S., & Liou, H.C. (2014). An automatic reference aid for improving EFL learners' formulaic expressions in productive language use. *IEEE Transactions on Learning Technologies*, *7*(1), 57–68. Retrieved from: http://dx.doi.org/10.1109/TLT.2013.34

Cho, Y. S. (2013). Software review: Lang-8. *CALICO Journal, 30*(2), 293–299. Retrieved from: http://dx.doi.org/10.11139/cj.30.2.293-299

Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. In *Proceedings of the 27th Annual Meeting on Association for Computational Linguistics* (pp. 76–83). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from: http://dx.doi.org/10.3115/981623.981633

Dahlmeier, D., & Ng, H. T. (2011). Correcting semantic collocation errors with L1-induced

paraphrases. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing* (pp. 107–117). Stroudsburg, PA: Association for Computational Linguistics.

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics, 19*(1), 61–74.

Futagi, Y., Deane, P., Chodorow, M., & Tetreault, J. (2008). A computational approach to detecting collocation errors in the writing of non-native speakers of English. *Computer Assisted Language Learning, 21*(4), 353–367. Retrieved from http://dx.doi.org/10.1080/09588220802343561

Harris. Z. (1954). Distributional structure. *Word, 10*(2–3), 146–162. Retrieved from http://dx.doi.org/10.1080/00437956.1954.11659520

Hill, J. (2000). Revising priorities: From grammatical failure to collocational success. In Michael Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 88–117). Hove: Language Teaching Publications.

Jurafsky, D., & Martin, J. H. (2009). *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* (2nd ed.). Upper Saddle River, NJ: Prentice Hall PTR.

Kitamura, M., & Matsumoto, Y. (1997). Automatic extraction of translation patterns in parallel corpora. *Information Processing Society of Japan Journal, 38*(4), 727–735.

Kudo, T., & Matsumoto, Y. (2002). Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning* (pp. 1–7). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from http://dx.doi.org/10.3115/1118853.1118869

Lea, D., & Runcie, M. (Eds.) (2002). *Oxford Collocations Dictionary for Students of English*. Oxford: Oxford University Press.

Leacock, C., Chodorow, M., Gamon, M., & Tetreault, J. (2010). *Automated Grammatical Error Detection For Language Learners* (Synthesis lectures on human language technologies 3(1), pp. 1–134). San Rafael, CA: Morgan & Claypool.

Lee, L. (1999). Measures of distributional similarity. In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics* (pp. 25–32). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from: http://dx.doi.org/10.3115/1034678.1034693

Lewis, M. (2000). There is nothing as practical as a good theory. In Michael Lewis (Ed.), *Teaching Collocation: Further Developments in the Lexical Approach* (pp. 10–27). Hove: Language Teaching Publications.

Liou, H., Chang, J., Chen, H., Lin, C., Liaw, M., Gao, Z., ... You, G. (2006). Corpora processing and computational scaffolding for a Web-based English learning environment: The CANDLE project. *CALICO Journal, 24*(1), 77–95.

Liu, A. L.-E.,Wible, D., & Tsao, N.-L. (2009). Automated suggestions for miscollocations. In *Proceedings of the Fourth Workshop on Innovative Use of NLP for Building Educational Applications* (pp. 47–50). Stroudsburg, PA: Association for Computational Linguistics. Retrieved from: http://dx.doi.org/10.3115/1609843.1609850

Liu, L. E. (2002). *A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learners' English* (Master's thesis). Tamkang University, Taipei.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T.,Ogura, H., Kashino, W., … Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation, 48*(2), 345–371. Retrieved from: http://dx.doi.org/10.1007/s10579-013-9261-0

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223–242. Retrieved from http://dx.doi.org/10.1093/applin/24.2.223

Oyama, H., Komachi, M., & Matsumoto, Y. (2013). Towards automatic error type classification of Japanese language learners' writings. In *Proceedings of the 27th Pacific Asia Conference on Language, Information, and Computation* (pp.163–172). Taipei, Taiwan.

Park, T., Lank, E., Poupart, P., & Terry, M. (2008). "Is the sky pure today?" AwkChecker: An assistive tool for detecting and correcting collocation errors. In *Proceedings of the 21th Annual Association for Computing Machinery Symposium on User Interface Software and Technology* (pp. 121–130). Monterey, CA, USA.

Pereira, L. (2013). *Collocation suggestion for Japanese second language learners* (Master's thesis). Nara Institute of Science and Technology, Ikoma, Japan.

Seretan, V. (2011). *Syntax-Based Collocation Extraction* (Text, speech and language technology series, 44). New York: Springer-Verlag. Retrieved from http://dx.doi.org/10.1007/978-94-007-0134-2_4

Shei, C.-C., & Pain, H. (2000). An ESL writer's collocational aid. *Computer Assisted Language Learning, 13*(2), 167–182. Retrieved from http://dx.doi.org/10.1076/0958-8221(200004)13:2;1-D;FT167

Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational Linguistics, 19*(1), 143–177.

Smadja, F., McKeown, K. R., & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational Linguistics, 22*(1),1–38.

Voorhees, E. M.(1999). The TREC-8 question answering track evaluation. In E. M. Voochees & D. K. Harman (Eds.), *Proceedings of the Text Retrieval Conference (TREC-8)* (pp. 83–105). NIST Special Publication 500-246.

Wible, D., Kuo, C., Tsao, N., Liu, A., & Lin, H. (2003). Bootstrapping in a language learning environment. *Journal of Computer-Assisted Learning, 19*(1), 90–102. Retrieved from http://dx.doi.org/10.1046/j.0266-4909.2002.00009.x

Yi, X., Gao, J., & Dolan, W. (2008). A web-based English proofing system for English as a Second Language users. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing* (pp. 619–624). Stroudsburg, PA: Association for Computational Linguistics.