NAIST-IS-DD1361016

Doctoral Dissertation

Collocation Writing Assistant for Learners of Japanese as a Second Language

Lis Weiji Kanashiro Pereira

March 14, 2016

Department of Information Science
Graduate School of Information Science
Nara Institute of Science and Technology

A Doctoral Dissertation
submitted to Graduate School of Information Science,
Nara Institute of Science and Technology
in partial fulfillment of the requirements for the degree of
Doctor of ENGINEERING

Lis Weiji Kanashiro Pereira

Thesis Committee:

Professor Yuji Matsumoto	(Supervisor)
Professor Kenichi Matsumoto	(Co-supervisor)
Associate Professor Masashi Shimbo	(Co-supervisor)
Assistant Professor Hiroyuki Shindo	(Co-supervisor)

Collocation Writing Assistant for Learners of Japanese as a Second Language[*]

Lis Weiji Kanashiro Pereira

Abstract

Conventional word combinations, or collocations, have been long recognized as important in helping language learners to communicate more efficiently and to sound more like a native speaker. However, studies confirmed that collocations are challenging, even for advanced second language learners. While native speakers already have a large number of collocations available in their mental lexicon, learners often struggle to find the right combination of words.

The goal of this thesis is to prove the feasibility of using natural language processing techniques to develop a writing system to suggest more appropriate collocations in Japanese. In particular, we address the problem of generating and ranking candidates for correcting potential collocation errors in the learners' text. The system generates possible correction candidates based on corrections extracted from a large Japanese learner corpus. This corpus is used to investigate the learner's tendency to commit collocation errors and to produce a smaller and more realistic set of candidates. In addition, the system uses the Weighted Dice coefficient as the association measure to filter out inappropriate candidate pairs and rank the proper collocations.

We carried out experiments focusing on noun-verb constructions, which are one of the major types of collocation problems. 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 describe how to utilize this method to develop a writing assistant where learners can apply the given collocation suggestions to revise their composition.

[*] Doctoral Dissertation, Department of Information Science, Graduate School of Information Science, Nara Institute of Science and Technology, NAIST-IS-DD1361016, March 14, 2016.

Keywords:

Collocation, language learners, natural language processing, automated grammatical error correction, computer assisted language learning.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This chapter sets the topic of the dissertation. The background and the motivation for building a collocation writing assistant for learners of Japanese as a second language (JSL) are described in Section 1.1, Section 1.2 and Section 1.3. The contributions of this thesis are summarized in Section 1.4 and its outline is given in Section 1.5.

1.1 Language Learners and Word Usage

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. For instance, consider a learner's sentence: 私は朝ご飯を食べて、靴を着ます (*watashi wa asa gohan wo tabete kutsu wo kimasu*, lit. 'I eat breakfast and dress my shoes'). Since non-native speakers might see this sentence as a grammatical one, they might have difficulty in judging whether this sentence sounds awkward or not. It can be even harder to notice that the correct appropriate usage should be 靴を履きます (*kutsu wo hakimasu*, 'put shoes on').

In addition to the difficulty in identifying word usage misuses, learners are also prone to put together words unidiomatically from their vocabulary inventory of individual items, when really a prefabricated chunk is needed to create a native result [5]. For example, learners have difficulty in combining a proper verb to go together with the noun 夢 (*yume*, 'dream') to make the combination 夢を見る

(*yume wo miru*, ‘have a dream’) and often end up saying 夢をする (*yume wo suru*, lit. ‘do a dream’). This consists with Ozaki’s argument [48] that collocations are more problematic when they are used in productive skills than in receptive skills. Even if learners know the meaning of a certain word, this does not enable and guarantee to know how to use the word correctly.

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 [33] argues that “increasing the learners’ collocational competence is the way to improve their language as a whole” (p. 14). In a separate study, Hill [19](p. 62) explains that a student who uses collocations competently will be far more competent in communication than a student who does not.

Studies confirm that the correct use of collocations is challenging, even for advanced second language learners ([37], [44], [70]). 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, such as syntactic errors, but offer no assistance for non-native errors. Futagi, Deane, Chodorow and Tetreault [15] 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 [71] and Varghese, Varde, Peng and Fitzpatrick [68] observe that language learners often use search engines to check if a phrase is commonly used by observing the number of results returned. However, search engines are not designed to offer alternative phrases that are more commonly used than the learner’s phrase [49]. 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 [5]. Thus, a computer program that automatically identifies potential collocation errors and suggests corrections would be a more appropriate resource for second language learners.

1.2 What *Collocation* Means in this Work

The literature defines collocations in several ways. However, the precise definition of collocation is still unclear and remains controversial. Lea and Runcie, for example, define collocations as combinations of words in a language to produce natural-sounding speech and writing ([30], vii). Smadja describes collocations in statistical terms as recurrent combinations of words that co-occur with higher possibility than random chance and correspond to some arbitrary word usages ([63], p. 143).

The linguistically-motivated definition of Cowie [8] adopts a more restrictive view of collocations. Cowie defines a collocation as the "the co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern". In other words, the participating words of a collocation must be related syntactically ([58], p. 12).

Regarding the semantic compositionality, collocations "fall somewhere along a continuum between free word combinations and idioms" ([25], p.509). Collocations have a seemingly more limited meaning than the literal interpretation [24], but they can be distinguished from idioms, which have meanings that are more opaque ([58], 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 ([4], [59], [31], p. 65). For example, the collocation 音楽を聴く (*ongaku wo kiku*, 'to listen to music') would be regarded as less appropriate if the word 聴く (*kiku*, 'to listen') is substituted by the word 聞く (*kiku*, 'to listen'), even though 聴く and 聞く have similar meaning [1]. 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;
3. have components linked by a syntactic relation (e.g. object-verb); and
4. would be regarded as less appropriate when one of the components is replaced by another word.

[1] The difference in meaning of the two characters is the same as the difference between 'to hear' and 'to listen to' in English (e.g. 'listen to music' is preferred over 'hear music').

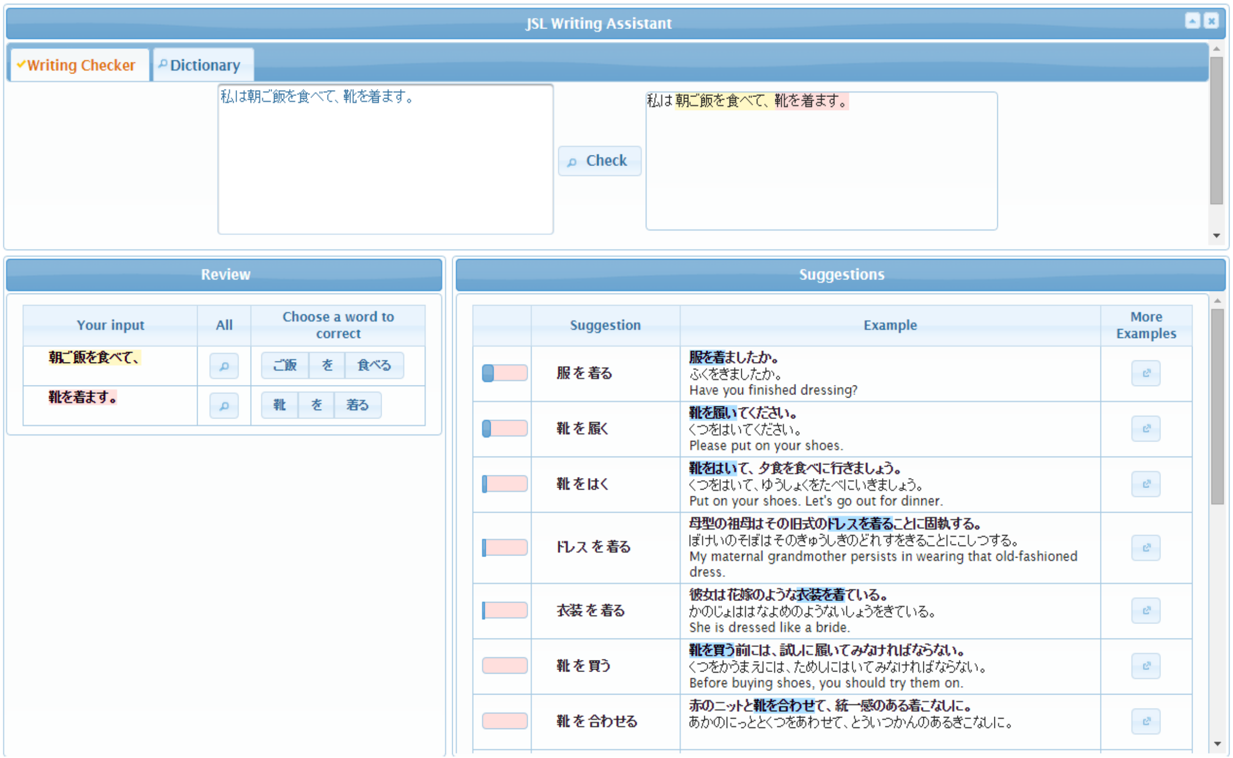
Some examples of Japanese collocations are shown in Table 1.1.

Collocation	Meaning	Literal Meaning
油を引く	put some oil in	pull oil
塩をする	salt	do salt
計画を立てる	make plans	put up a plan
夢を見る	dream; have a dream	see a dream
保険に入る	buy insurance	go into insurance
声をかける	call out (to a person)	hang voice
速度を落とす	slow down	drop the speed
アイデアが浮かぶ	an idea occurs to one; an idea occurs to mind	an idea floats
速度が速い	one's speed is fast	speed is fast
物価が安い	prices are low	price is cheap
成績が伸びる	improve one's grade	lengthen one's grades

Table 1.1: Examples of Japanese collocations. Source: Shoji [62].

1.3 Motivation

Collocation error correction involves substitutions from potentially large sets of open class words; i.e., nouns, adjectives, verbs and adverbs ([31], p.68). 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 [15], [36] and [49] 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.



私は朝ご飯を食べて、靴を着ます。

私は朝ご飯を食べて、靴を着ます。

Your input	All	Choose a word to correct
朝ご飯を食べて、	⌕	ご飯 | を | 食べる
靴を着ます。	⌕	靴 | を | 着る

	Suggestion	Example	More Examples
	服 を 着る	服を着ましたか。 ふくをきましたか。 Have you finished dressing?	
	靴 を 履く	靴を履いてください。 くつをはいてください。 Please put on your shoes.	
	靴 を はく	靴をはいて、夕食を食べに行きましょう。 くつをはいて、ゆうしょくをたべにいきましょう。 Put on your shoes. Let's go out for dinner.	
	ドレス を 着る	母型の祖母はその旧式のドレスを着ることに固執する。 ぼけいのそぼはそのきゅうしきのどれ すをきることにこしつする。 My maternal grandmother persists in wearing that old-fashioned dress.	
	衣装 を 着る	彼女は花嫁のような衣装を着ている。 かのじょははなよめのようないしょうをきている。 She is dressed like a bride.	
	靴 を 買う	靴を買う前には、試しに履いてみなければならない。 くつをかうまえには、ためしにはいてみなければならない。 Before buying shoes, you should try them on.	
	靴 を 合わせる	赤のニットと靴を合わせて、統一感のある着こなしに。 あかのにっととくつをあわせて、とういつかんのあるきこなしに。	

Figure 1.1: The interface of the JSL Writing Assistant system showing collocation suggestions to the wrong collocation 靴を着ます (*kutsu wo kimasu*, lit. 'to dress shoes').

This study generally aims at developing a system that targets potential collocation errors made by Japanese as a second language (JSL) learners (shown in Figure 1.1). This system accepts sentences as the input and suggests better collocations to the user. We propose a method to correct potential collocation errors made by JSL learners by a combination of a large learner corpus and statistical association measures. According to a 2013 report from the Japan Foundation [21], there are almost four million people learning Japanese outside Japan, and it is hoped that this automatic suggestion tool for collocations will help JSL learners improve their collocational knowledge.

1.4 Contribution

The contributions of this thesis are summarized as follows:

- We propose a writing assistant system that targets potential collocation errors made by learners of Japanese as a second language.
- The proposed tool has two unique features. First, it generates correction candidates by using 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 [27] as a statistical association measure for ranking the collocation correction candidates. This measure achieved the best performance for ranking the collocation correction candidates in our task.
- In an extensive empirical evaluation conducted, we show that our proposed method for suggesting collocations outperforms existing methods based on the semantic relation of words.

1.5 Organization of the Thesis

This dissertation is organized as follows:

This chapter gives the general introduction for the research. The motivation was discussed along with the contributions that this thesis wishes to achieve. Chapter 2 provides an overview of the field of automated grammatical error correction and a review of related studies that proposed systems specially developed for the collocation suggestion problem and for collocation learning in general. Chapter 3 presents our proposed method on how to automate collocation suggestion and describes the detailed experiments and evaluation results. In Chapter 4, we demonstrate how this proposed method can be used to build practical applications for JSL learners. Finally, we present the conclusions and the future work for our research in Chapter 5.

Chapter 2

Background

This chapter provides an overview of the automated grammatical error correction and collocation research fields. It also describes several systems specially developed for the collocation suggestion problem and for collocation learning in general.

2.1 Automated Grammatical Error Correction

Automated grammatical error correction falls under the broader category of Computer Assisted Language Learning (CALL) [31]. The field of CALL involves the use of a computer as supplementary material in language learning and teaching. Examples of CALL applications are flashcards, podcasts and chat-rooms. On the other hand, automated grammatical error correction refers to the specific task of detecting and correcting grammatical errors present in a text written by a second language earner. In this field, *grammatical errors* refer to written errors made by language learners. These errors can be categorized as being *syntactic errors* or *usage errors* [60].

Syntactic errors involve violations of structural syntactic rules that are clearly defined in any prescriptive grammar manual. These errors include ill-formed verb-phrases, violations of subject-verb and determiner-noun agreement and errors in pronoun case and often cause problems for automatic parsers. On the other hand, usage errors such as collocation errors can result from inaccurate memorization or complex interactions between syntactic features, lexical features, discourse factors, and, in some cases, world knowledge. With usage errors, usually there is no clear-cut syntactic rule that has been violated. Therefore, grammars generally cannot handle questions such as which noun is the best choice in a given context. Usage errors are rarely committed by native speakers, however, these are the most frequent errors of language learners. Most commercially available grammar checkers, such as Microsoft Word, are designed to be used by native speakers and thus focus on syntactic, not usage, errors.

In the last few years, there has been a surge of interest in applying natural language processing (NLP) techniques for detecting and correcting non-native grammatical errors. Most works have focused on the correction of the usage of closed class words such as articles and prepositions in English ([16], [55], [66]). Closed classes are those that have relatively fixed membership. For example, prepositions are closed class words because there is a fixed set of them in English; new prepositions are rarely coined [23]. In Japanese, case particles are examples of closed class words. By contrast, open class words (i.e. noun, verb, adjective and adverb) are much larger in number and are continually coined or borrowed from other languages (e.g. foreign loan words (*gairaigo*) in Japanese).

The general approach for correcting non-native errors compares the writer's word choice to alternative choices and if one or more alternatives provide a much better fit to the context, then the writer's word is flagged as an error and the alternatives are suggested as corrections. In the task of correcting Japanese particles, for example, usually around 10 particles are considered as alternatives (see [65] and [46] for examples).

The vast majority of work in automatic grammatical error correction has been on the prediction of closed class words in text for two main reasons:

1. Errors involving closed class words are frequent in second language writing. In English, for example, definite and indefinite articles and prepositions account for 20%-50% of all grammar and usage errors [31]. In Japanese, similar phenomena occurs with case particles, wich account for 25% of the total errors [47].

2. They are simpler to correct compared to errors involving open class words. For example, in the case of English article errors, there are just two alternatives to consider as possible corrections. Similarly, for prepositions in

English and case particles in Japanese, the alternatives constitute a relatively small and fixed set.

Regarding to the work on the prediction of open class words, there has been research on collocation suggestion. The next section describes works on collocation research to gain deeper insights into the studies that are more closely related to ours.

2.2 Collocation Research

2.2.1 Collocation Extraction

Collocations are important for a number of applications, such as natural language generation (to make sure that the output sounds natural [39], p. 142), computational lexicography (to automatically identify the impotant collocations to be listed in a dictionary [39], p. 142), parsing (to solve attachment ambiguities by giving preference to analyses in which they occur ([58], p. 3)), corpus linguistic research, machine translation and so on. Given its importance, much of the NLP research on collocations have focused on collocation extraction methods.

A simple method to extract collocations is counting the frequency of the candidate phrases. Justeson and Katz [24], for example, used n-gram frequencies to identify technical terms in a text. To improve the results, they applied a part-of-speech filter to obtain only those patterns that are likely to be ‘phrases’ ([39], p. 143), such as adjective-noun and noun-noun phrases. Despite its simplicity, this method yielded surprisingly accurate results. However, only co-occurrency frequency is not enough to quantify the strength of collocativity [13] since high frequency can be accidental ([39], p.152). For example, if the two constituent words of a frequent bigram are frequently occurring words, then we expect the two words to co-occur a lot just by chance, even if they do not form a collocation ([39], p. 152).

Instead of the frequency-based method, several mathematical and statistical approaches to collocation use *association measures* to measure the strength of association between words [31] (see [58], [39], [50] and [13] for examples). These

	y	¬y	Total
x	freq(x,y)=a	freq(x,¬y)=b	a+b
¬x	freq(¬x,y)=c	freq(¬x,¬y)=d	c+d
Total	a+c	b+d	N = a+b+c+d

Table 2.1: Contingency table.

measures are mathematical formulas that consider both the co-occurrence frequency and the marginal frequencies of the individual words. The words within a collocation are expected to have higher association strength. Detailed descriptions of several association measures can be found in [39], [50] and [13]. Pecina [50], for example, presents the formulas for 84 such measures. We present in this section four associations measures: the pointwise mutual information, the log-likelihood ratio, the Dice coefficient and the Weighted Dice coefficient. Their formulas are defined in terms of a contingency table (Table 2.1). This table represents the unigram and bigram frequencies of two words x and y in a corpus. The labels ¬x and ¬y represent the absence of x and y, respectively and their marginal totals are equal to the total frequency (N) of all bigrams in the corpus minus the unigram frequency of each word. The table has four cells, representing the items containing both x and y (a), the items containing x but not y (b), the items containing y but not x (c) and the items containing neither of the two words (d).

Pointwise mutual information. Pointwise Mutual Information (PMI) [7] is an association measure related to Information Theory concepts. It measures how often two words co-occur in a corpus, compared with what we would expect if they occurred independently. PMI is defined as:

$$\log_2 \frac{aN}{(a+b)(a+c)} \tag{2.1}$$

Log-likelihood ratio. The log-likelihood ratio (LL) [12] compares two hypotheses to determine which hypothesis is more likely to occur than the other. The first hypothesis proposes that two terms occur independently from each other, while the second hypothesis proposes that the occurrence of one of the terms is dependent on the occurrence of the other term. The log-likelihood ratio is defined as follows:

$$\begin{aligned}
&a \log a + b \log b + c \log c + d \log d - \\
&(a+b)\log(a+b) - (a+c)\log(a+c) - \\
&(b+d)\log(b+d) - (c+d)\log(c+d) + \\
&\qquad (a+b+c+d)\log(a+b+c+d)
\end{aligned} \tag{2.2}$$

Dice coefficient The Dice coefficient is a classical measure used in information retrieval to calculate the similarity between two sets. It was used by Smadja [63] to extract collocations from text corpora. The Dice coefficient can be interpreted as a measure of predictability [13], based on the harmonic mean average of the proportion of instances of x that co-occur with y and the proportion of instances of y that co-occur with x [13]. It is defined as follows.

$$\frac{2a}{2a+b+c} \tag{2.3}$$

Weighted Dice coefficient The Weighted Dice coefficient [27] is a modification of the Dice coefficient. It corresponds to the Dice coefficient formula weighted by the logarithm of the co-occurrence frequency ($\log_2 a$). It is defined as follows:

$$\frac{2a}{2a+b+c} . \log_2 a \tag{2.4}$$

Criteria for the Application of Association Measures

The collocation extraction research has applied various association measures in different settings, without clearly defined criteria for choosing one particular association measure rather than another ([58], p. 42). However, there exist theoretical reasons that make some measures more appropriate for collocation extraction than others, in a given setting.

PMI, for instance, although it is the most popular association measure, in many cases it is not a good measure of what an interesting correspondence between two events is ([39], p. 168). PMI is a good measure of independence, but a bad one of dependence, because for dependence the score depends on the frequency of the individual words. Other things being equal, bigrams composed of low-frequency words will receive a higher score than bigrams composed of high-frequency words ([39], p. 182). That is the opposite of what a good measure for extracting collocations is expected to do since higher frequency means more evidence ([39], p. 182). On the other hand, by using a frequency threshold on large data, PMI was shown to lead to competitive results [51]. Despite all controversies, PMI remains popular.

Contrary to PMI, the log-likelihood ratio is argued to be appropriate to both rare and common phenomena, and to both large and small text samples [12]. It has a more clear interpretation than PMI: it tells how strongly the occurrence of a word x depends on the occurrence of a word y, which is a good evidence for an interesting collocation. The log-likelihood ratio is generally considered as the most appropriate measure for collocation extraction ([58], p. 43).

The Dice coefficient focuses on cases of very strong association rather than the comparison with independence, as it does not assume a well-defined value in the case of independence. It has been pointed out that it can identify pairs with a particularly high degree of lexical cohesion [64] and has emerged as a more competive measure in several works ([14], [54], [1], [27]).

The Weighted Dice coefficient [27] is a less known measure. However, it has been shown to be an improvement over the Dice coefficient ([27]). In the Dice coefficient, regardless of the frequency of the occurrence, the maximum value is 1 when the pair always occurs, while the Weighted Dice coefficient takes the absolute number of co-occurrences into consideration, assigning higher scores to higher frequency pairs.

Despite the existing theoretical reasons, the suitability of an association depends on various factors of the settings where the experiment takes place (e.g., language or domain) [13]. A common strategy is to compare the individual merits of association measures ([58], p. 43) to select the most appropriate one in a specific task.

2.2.2 Collocation Error Correction

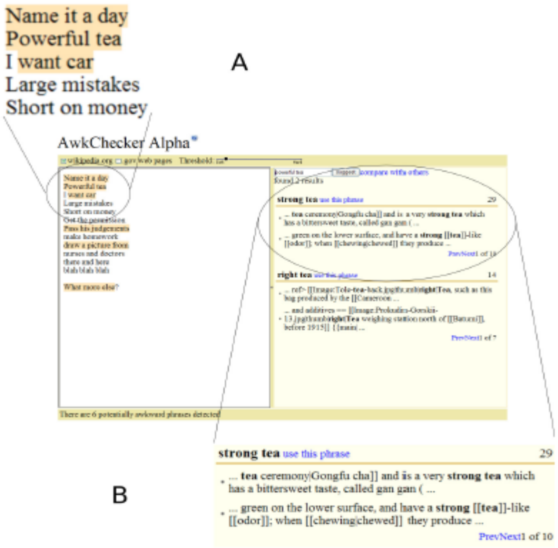
Collocation error correction is commonly performed by computing the differences in distribution between collocations and their non-collocational counterparts. Usually targeting syntactic constructions of a particular type or types, such as verb+noun, the general existing approach consists of two steps:

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 (e.g. pointwise mutual information) 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. A simpler approach to rank candidates is to use co-occurrence frequency [24]. Well-formed corpora are used as reference corpora when applying these measures, since they can provide evidence for constructions that are common in a language. Therefore, these corpora can be used to extract correct collocations and filter out learners' incorrect usages.

Several researchers have proposed useful corpus-based tools for correcting collocation errors ([15], [36], [49], [4], [70], [10], [59], [10], [45]). In a user study, Park et al. [49] observed positive reactions from users when using their system. In another study, Liou et al. [35] showed that the miscollocation aid proposed by Chang et al. [4] can help learners improve their knowledge in collocations. Some of the existing systems for automated collocation error correction are described in the next section. All these systems were proposed for English, except the work of Ostling and Knutsson [45]. Most of these works focus on verb-noun collocations, which are usually the largest in number and the most difficult for learners ([37], [44]).

Existing Collocation Error Correction Systems



Name it a day
Powerful tea
I want car
Large mistakes
Short on money
Get the permission
Pass his judgements
make homework
draw a picture from
nurses and doctors
there and here
blah blah blah

What more else?

There are 6 potentially awkward phrases detected

strong tea use this phrase 29

- ... tea ceremony|Gongfu cha]] and is a very **strong tea** which has a bittersweet taste, called gan gan (...
- ... green on the lower surface, and have a **strong** [[**tea**]]-like [[odor]]; when [[chewing|chewed]] they produce ...

PrevNext1 of 10

right tea use this phrase 14

- ... ref> [[Image:Tole-tea-back.jpg|thumb|**right**|Tea, such as this bag produced by the [[Cameroon ...
- ... and additives == [[Image:Prokudin-Gorskii-13.jpg|thumb|**right**|Tea weighing station north of [[Batumi]], before 1915]] {{main| ...

PrevNext1 of 7

Figure 2.1: The interface of *AwkChecker*. A) shows the flagged phrases in the composition window; and B) shows one of the suggested alternative phrases for the erroneous collocation “powerful tea”.

AwkChecker Park et al. [49] developed *AwkChecker*, an end-user tool geared towards helping non-native speakers detect and correct collocation errors in their writing. As a user writes, the system automatically flags collocation errors and suggests replacement expressions that correspond more closely to consensus usage (Figure 2.1). These suggestions include example usage to help users choose the best candidate. They used WordNet to generate synonym candidates and co-occurrence frequency obtained from Wikipedia corpus and “.gov” webpages to rank the candidates. In the user study conducted, the system was tested by five non-native speakers. Three of the participants were given an essay written by a non-native speaker to edit, while the other two edited their own content. Additionally, semi-structured interviews on the design of the interface, the features and the usefulness of AwkChecker were conducted. The tool received positive feedback from the users. However, it is unclear how accurate the system work, since no evaluation of the system performance was reported.

Educational Testing Service' System Futagi et al. [49] proposed a prototype for detecting collocation errors. Their system targeted seven kinds of collocations (Table 2.2). They generated synonyms for each candidate string using WordNet and Roget's Thesaurus. To rank the candidates, they used the rank ratio measure [11] applied to a corpus of more than one billion words consisting of literary and scientific journal articles and of elementary to post-graduate level texts. The system was evaluated on 1260 collocation strings extracted from TOEFL essays. The F-score (weighted average of the precision and recall) for acceptable collocations was 0.91, but for error collocations it was only 0.34.

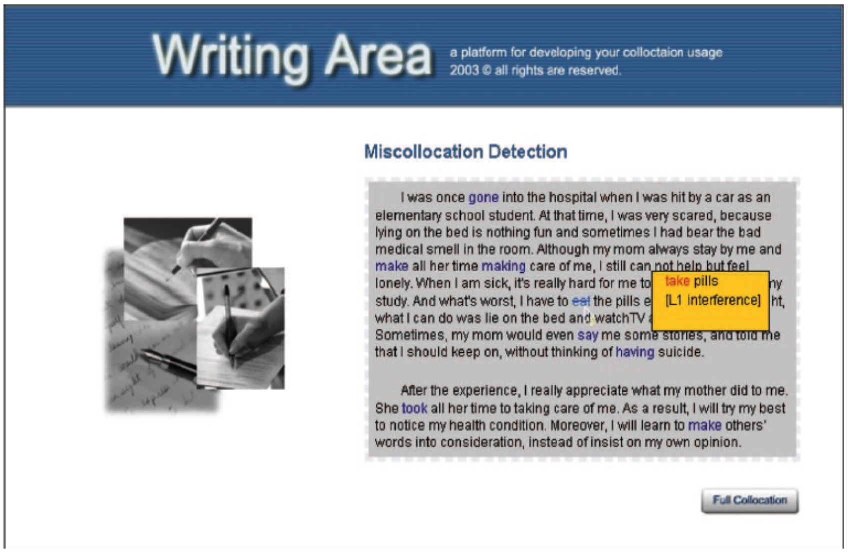
Ostling and Knutsson (2009) The work of Ostling and Knutsson [45] is one of the few works that focus on languages other than English. Targeting verb-noun and adjective-noun constructions, they proposed a tool to help the user find collocationally acceptable phrases in written Swedish. They generated synonyms for each candidate string using a synonym dictionary and a word similarity based on random indexing [56]. To rank the candidates, they used a combination of the synonym dictionary and the random indexing synonymity measures with the pointwise mutual information [7] and log-likelihood ratio [12]. The system was evaluated on a compiled list of 60 non-native like collocations (30 verb-noun and 30 adjective-noun collocations). The tool was able to automatically find an acceptable collocation 57% of the times.

Liu et al. (2009) Liu et al. [36] explored the notion of shared collocations. The idea is that collocations that can be clustered via overlapping collocates can be the source of collocation errors for language learners. For example, consider the verb-noun collocation error 'reach a purpose'. Their system generates a cluster by finding verbs that collocate with the noun 'purpose' and nouns that collocate with the verb 'reach'. The verbs are ranked by the number of collocates that they share with the error verb. The highest ranking verbs that also have 'purpose' as a collocate can then be offered as suggested corrections.

Their system used WordNet to generate synonym candidates and a probabilistic model that combined the notion of shared collocations, the ranking with Pointwise Mutual Information [7] and a semantic similarity measure derived from WordNet. The system was evaluated on 42 verb-noun miscollocations extracted from essays written by Taiwanese learners of English. The best model obtained a precision of 53%.

Writing Assistant Chang et al. [4], in contrast, emphasized L1 interference as the main cause of collocations errors. They used English-Chinese wordlists that were derived from bilingual dictionaries and from a word-aligned bilingual English-Chinese parallel corpus. For example, the Chinese word for 'eat' sometimes corresponds to the English 'eat' and sometimes to other English words, such as 'take' (as in 'take the pill'). As a result, 'eat' and 'take' were placed in the same 'synonym' set because both corresponded to Chinese 'eat' in at least some context. When 'eat the pill' was encountered in non-native text, the alternative phrase 'take the medicine' was generated as a candidate correction (as shown in Figure 2.2). This candidate phrase was then checked against the list of collocations that had been extracted by using the log-likelihood measure [12] applied to the British National Corpus (BNC) to filter out any that might be unacceptable. The system obtained an impressive overall performance of 98% precision and 91% recall. The quality of the suggested corrections were evaluated in terms of mean reciprocal rank (MRR) [69]. The system obtained a MRR value of 0.66, indicating that the correction was mostly found as the first or second suggestion. However, to extend their approach, bilingual dictionaries and parallel corpora would be needed for every L1. Unfortunately, these resources are scarce for many language pairs.

Dahlmeier and Ng (2011) Similar to the work of Chang et al. [4], Dahlmeier and Ng [10] also explored the idea that L1 interference is the main cause of collocations errors. However, they tried to offer better coverage by automatically deriving paraphrases from a Chinese-English parallel corpus to serve as collocation candidates. The paraphrase probability is computed for each candidate given the writer's collocation and for the writer's collocation given each candidate. They adapted an approach used in phrase-based statistical machine translation to use these probabilities to score and rank the candidates. The system obtained a precision of 38% and a mean reciprocal rank value of 0.57. Similar to the approach of Chang et al. [4], to extend the system, parallel corpora would be needed for every L1.



I was once gone into the hospital when I was hit by a car as an
elementary school student. At that time, I was very scared, because
lying on the bed is nothing fun and sometimes I had bear the bad
medical smell in the room. Although my mom always stay by me and
make all her time making care of me, I still can not help but feel
lonely. When I am sick, it's really hard for me to ny
study. And what's worst, I have to ~~eat~~ the pills e ht,
what I can do was lie on the bed and watchTV a
Sometimes, my mom would even say me some stories, and told me
that I should keep on, without thinking of having suicide.

After the experience, I really appreciate what my mother did to me.
She took all her time to taking care of me. As a result, I will try my best
to notice my health condition. Moreover, I will learn to make others'
words into consideration, instead of insist on my own opinion.

Figure 2.2: The interface of *Writing Assistant* showing an example of the pop-up prompt-box with the possible answers.

Collocation Tutor Shei and Pain [59] introduced the idea of using a learner corpus in the collocational aid they developed. The learner corpus consisted of English texts written by post-intermediate Chinese learners of English. The corpus was used to build an Error Library, which contained a collection of collocational errors made by learners. They also built a reference database which consisted of collocations extracted by applying the z-score to the British National Corpus (BNC). The learners' input are first checked against the reference database to see if they are a valid collocation. If they are not found in the reference database, they are looked up in the Error Library. If the structures are found in this library, they are marked as definetely anomalous. If the structures are still undecided, they use a synonym dictionary derived from Wordnet to see if any synonym would form a legitimate collocation with the other collocate(s). If an appropriate synonym is found, it is suggested to the user as an alternative. Finally, the entire structure may be replaced by more native-like collocations listed in what they call 'definition dictionary', which was created from paraphrases of collocations given by the learners. The performance of the system is not reported by the authors and, as pointed out by Futagi et al. [15], it is unclear to what extent their system depends on automated methods purely, as opposed to the manually constructed database of common errors that it incorporates.

Wible et al. (2003) Wible et al. [70] developed a system to detect collocation errors in verb-noun constructions. They extracted miscollocations from the IWiLL learners corpus, a collection of Taiwanese student essays. They manually constructed a list of nine nouns, each of which had its own list of corresponding miscollocated verbs which were inappropriately used with it in the corpus. The grammar checker automatically marked a collocation error when one of the nine nouns was found with a corresponding miscollocated verb. The system obtained a high precision rate of 95.5%. However, the limitation is that it would require and enormous annotation effort (which is also costly) to cover a wider variety of miscollocations.

Summary

To summarize the systems that have been proposed for collocation correction, we conclude with the following findings (also shown in Table 2.2):

1. 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. The drawback of these approaches is that these resources might have limited coverage to generate the candidates.

2. Although some works tried to offer better coverage by using resources such as parallel corpora, it is necessary to identify the learner's first language and to have parallel corpora for every first language to extend the resulting system. Unfortunately, parallel corpora is scarce for many language pairs.

3. Another drawback is that most of these systems rely only on well-formed English resources (except the works of Shei and Pain [59] and Wible et al. [70]) and do not actually take into account the learners' tendencies toward collocation errors.

4. Although the works of Shei and Pain [59] and Wible et al. [70] used learner corpora to identify collocation misuses in learners' writing, their systems relied on small learner corpus and on mannually constructed collocation error databases. These databases might also suffer from low coverage as the resources described in item 1.

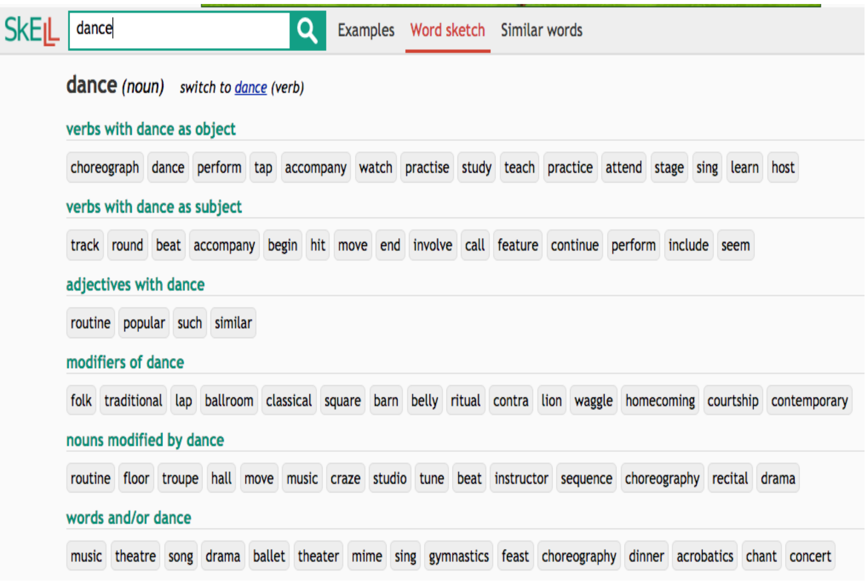
5. Finally, several of these works focus only on buiding their systems and do not provide extensive evaluation of system performance. Thus, it is unclear how accurate their systems work.

2.3 Other existing applications for collocation learning

Concordancers have also been proposed as tools to support collocation learning. Concordancers are computer programs that are used to retrieve target words/phrases in various text corpora [5]. They typically provide a word's grammatical or collocational behavior by displaying example sentences and had been primarily used by linguistic and literary researchers. The motivation to use concordancers as language learning tools came from the "data-driven learning" (DDL) approach [22] proposed by Johns [22]. Johns [22] suggested that "the language learner is also, essentially, a research worker whose learning needs to be driven by access to linguistic data" (p. 2). In the DDL approach, the concordancer is pointed as the most important computing tool.

Examples of concordancers that have been proposed as tools for collocation learning are: Word Sketch Engine [26] (Figure 2.3), TANGO [35] (Figure 2.4) and Natsume [20] (Figure 2.5). Word Sketch Engine and TANGO have been proposed for learning English, while Natsume has been proposed for learning Japanese.

However, as mentioned in Section 1.1, concordancers tend to retrieve too much information because they usually allow only single-word queries. Too much information might distract and confuse the user [5].



dance (noun) switch to dance (verb)

verbs with dance as object

choreograph dance perform tap accompany watch practise study teach practice attend stage sing learn host

verbs with dance as subject

track round beat accompany begin hit move end involve call feature continue perform include seem

adjectives with dance

routine popular such similar

modifiers of dance

folk traditional lap ballroom classical square barn belly ritual contra lion waggle homecoming courtship contemporary

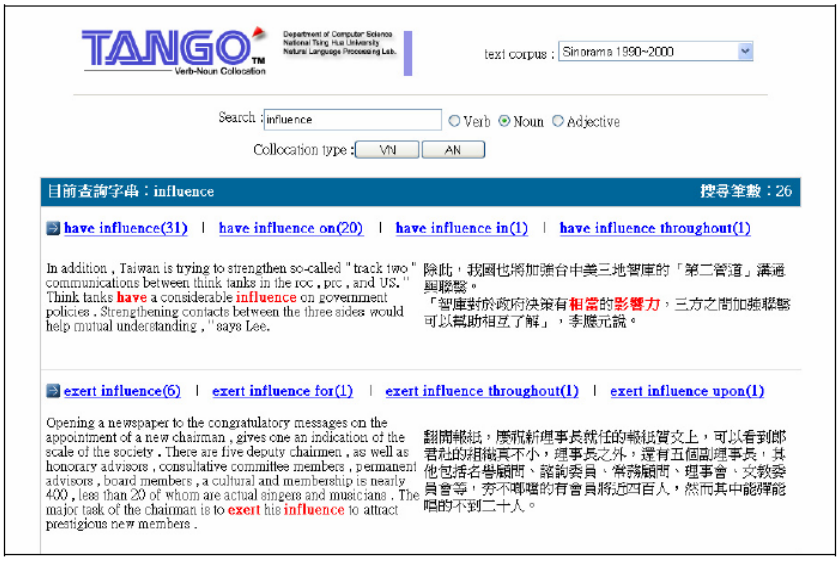
nouns modified by dance

routine floor troupe hall move music craze studio tune beat instructor sequence choreography recital drama

words and/or dance

music theatre song drama ballet theater mime sing gymnastics feast choreography dinner acrobatics chant concert

Figure 2.3: The interface of SKELL (Sketch Engine for Language Learning) system showing collocations for the noun ‘dance’.


TANGO™
Verb-Noun Collocation
Department of Computer Science
National Tsing Hua University
Natural Language Processing Lab.
text corpus : Sinorama 1990~2000
Search : influence
○ Verb ● Noun ○ Adjective
Collocation type : VN AN

目前查詢字串：influence 搜尋筆數：26

have influence(31) | have influence on(20) | have influence in(1) | have influence throughout(1)

In addition , Taiwan is trying to strengthen so-called " track two " communications between think tanks in the roc , prc , and US. " Think tanks have a considerable influence on government policies . Strengthening contacts between the three sides would help mutual understanding , " says Lee.

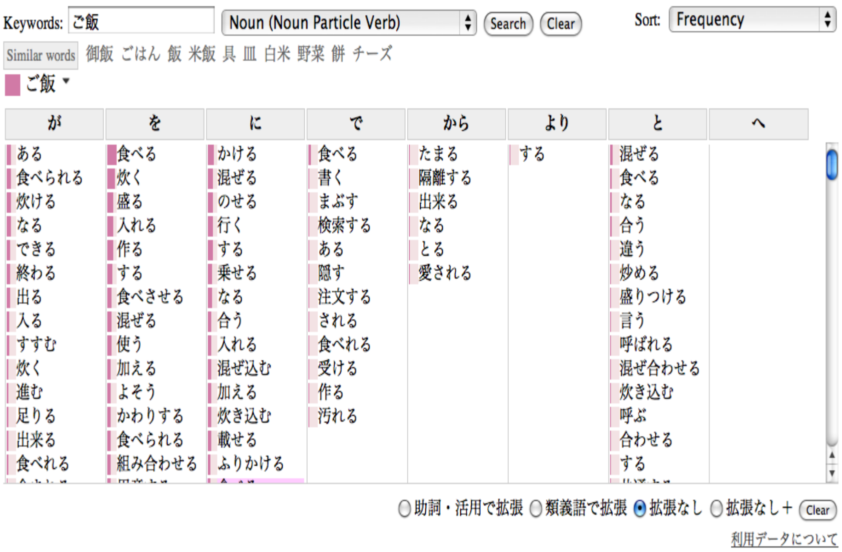
除此，我國也將加強台中美三地智庫的「第二管道」溝通與聯繫。
「智庫對於政府決策有相當的影響力，三方之間加強聯繫可以幫助相互了解」，李應元說。

exert influence(6) | exert influence for(1) | exert influence throughout(1) | exert influence upon(1)

Opening a newspaper to the congratulatory messages on the appointment of a new chairman , gives one an indication of the scale of the society . There are five deputy chairmen , as well as honorary advisors , consultative committee members , permanent advisors , board members , a cultural and membership is nearly 400 , less than 20 of whom are actual singers and musicians . The major task of the chairman is to exert his influence to attract prestigious new members .

翻開報紙，慶祝新理事長就任的報紙賀文上，可以看到郎君社的組織真不小，理事長之外，還有五個副理事長，其他包括名譽顧問、諮詢委員、常務顧問、理事會、文教委員會等，夯不啷噹的有會員將近四百人，然而其中能彈能唱的不到二十人。

Figure 2.4: The interface of TANGO system showing collocations for the noun ‘influence’.



Keywords: ご飯
Noun (Noun Particle Verb)
Search
Clear
Sort: Frequency
Similar words
御飯 ごはん 飯 米飯 具 皿 白米 野菜 餅 チーズ
ご飯

が	を	に	で	から	より	と	へ
ある	食べる	かける	食べる	たまる	する	混ぜる	
食べられる	炊く	混ぜる	書く	隔離する		食べる	
炊ける	盛る	のせる	まぶす	出来る		なる	
なる	入れる	行く	検索する	なる		合う	
できる	作る	する	ある	とる		違う	
終わる	する	乗せる	隠す	愛される		炒める	
出る	食べさせる	なる	注文する			盛りつける	
入る	混ぜる	合う	される			言う	
すすむ	使う	入れる	食べれる			呼ばれる	
炊く	加える	混ぜ込む	受ける			混ぜ合わせる	
進む	よそう	加える	作る			炊き込む	
足りる	かわりする	炊き込む	汚れる			呼ぶ	
出来る	食べられる	載せる				合わせる	
食べれる	組み合わせる	ふりかける				する	

○ 助詞・活用で拡張 ○ 類義語で拡張 ● 拡張なし ○ 拡張なし＋ Clear

利用データについて

Figure 2.5: The interface of Natsume system showing collocations for the noun ご飯 (*gohan*, ‘rice’).

System	**Language**	**Type**	**Method to generate candidates**	**Method to rank candidates**
AwkChecker	English	Verb-Noun Adjective-Noun	synonyms from Wordnet	co-occurrence frequency
Educational Testing Service	English	Verb-Noun Adjective-Noun Noun-Noun Noun-Verb Noun of Noun Adverb-Adjective Verb-Adverb	synonyms from Wordnet and Roget's thesaurus	rank-ratio measure
Ostling and Knutsson (2009)	Swedish	Verb-Noun Adjective-Noun	synonym dictionary and random indexing	combination of the synonym dictionary and the random indexing synonymity measures with the pointwise mutual information and log-likelihood ratio
Liu et al. (2009)	English	Verb-Noun	synonyms from Wordnet	probabilistic model combining the notion of shared collocations, the ranking with pointwise mutual information and a semantic similarity derived from Wordnet
Writing Assistant	English	Verb-Noun	wordlists derived from bilingual dictionaries and from a word-aligned bilingual parallel corpus	log-likelihood ratio
Dahlmeier and Ng (2011)	English	not specified	paraphrases derived from bilingual parallel corpus	paraphrase probability
Collocation Tutor	English	Verb-Noun	synonyms from Wordnet	z-score
Wible et al. (2003)	English	Verb-Noun	mannually constructed list of nine nouns, each of which had its own list of corresponding miscollocated verbs and their corrections	no ranking method was used

Table 2.2: Summary of the programs dedicated for collocation correction.

2.4 This work

In this study, we describe the development of a writing aid for JSL learners. The system targets potential collocation errors in their writing. We propose to use corrections automatically extracted from a large Japanese learner corpus. The system is designed to better explore the learner's tendency to commit collocation errors compared to standard methods that generate candidates based on the semantic relation of words. In addition, we propose to use the Weighted Dice measure to rank the collocation candidates. In the next chapter, we provide an extensive evaluation of our system performance and report on it in detail, in the hope of providing insight into the complexity of the problem at hand and of finding effective solutions.

Chapter 3

Collocation Suggestion based on a Large Learner Corpus and Statistical Association Measures

In this study, we designed a system to cope with potential collocation errors produced by JSL learners and prompt them with more appropriate collocations for revision. In this chapter, we describe in detail the method proposed to suggest collocations, as well as the evaluation experiments conducted.

3.1 Introduction

It has been widely acknowledged that collocations are both important and problematic for language learners. Thus, a system that can detect learners collocation errors and suggest the most appropriate collocations as corrections can help improve their collocational knowledge.

We propose a system that automatically extracts noun-verb collocations from learners' texts, detects any potential misuses, and provides more appropriate collocations as suggestions. For example, when a learner produces an erroneous noun-verb collocation 夢をする (*yume wo suru*, lit. 'do a dream'), the system would highlight it and flag it as a potential collocation error. Then, the collocation undergoes a correction process to find more appropriate collocations. Finally, the system outputs stronger noun-verb collocations as suggestions, such as 夢を見る

(*yume wo miru*, ‘to dream’).

3.2 Collocation Suggestion Method

The main property of this system is the reliance on a large learner corpus to suggest collocations. The suggested collocations are offered based on correction pairs extracted from this copus. In addition, a collocation list built by appliying the Weighted Dice coefficient[27] to a large reference corpus is used as reference collocation database to provide correct collocation uses. More specifically, given a text written by a learner, the system does the following:

1. **Target Phrase Extraction**: the system extracts noun-verb phrases from the input text by using a dependency parser;

2. **Potential Collocation Error Detection**: for each phrase extracted, the system checks if it exists in the collocation database. If not, the phrase is flagged as a potential collocation error;

3. **Candidate Generation**: the user first chooses the component word (noun or verb) he/she wants to correct in each phrase. Then, the system creates a set of correction candidate phrases by substituting the chosen word with words that it was corrected to in the large learner corpus. The learner's original phrase is also included in the set;

4. **Candidate Ranking**: the system filters out all phrases that do not exist in the collocation database and measure the strength of association in each remained phrase. Finally, the higher-ranking phrases are suggested as corrections.

Even if the learner's noun-verb input phrase is not flagged as a potential error, it will undergo the correction process because better collocations might exist. Figure 3.1 shows a flow-chart representation of the processing stages. Each stage is detailed in the next sections.

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.

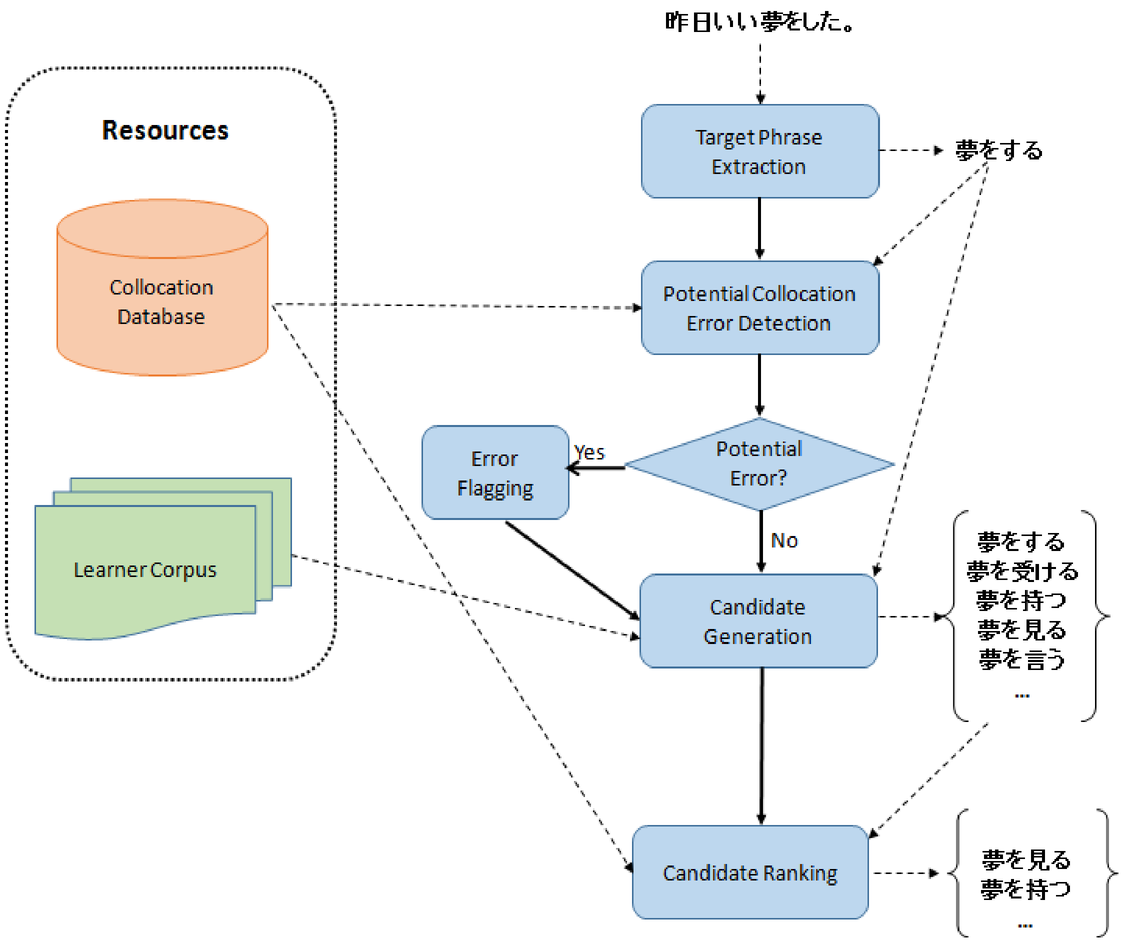


Figure 3.1: The processing stages of a learner's text.

3.2.1 Target phrase extraction

The first stage aims to identify phrases in the learner's text that match target syntactic patterns. The system focuses on noun-verb constructions and considers the lemma of the component words of the phrase. For instance, in the sentence example from Figure 3.1 昨日いい夢をした (*kinou ii yume wo shita*, lit. 'I did a good dream yesterday'), the system extracts the target noun-verb phrase *夢をする (*yume wo suru*, lit. 'do a dream'). The current version of the system does not have a module that specifically handles the kanji-kana conversion, considering the original form (kanji or katakana/hiragana) of the input string.

The noun-verb constructions in Japanese 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 noun-verb patterns listed in Table 3.1.

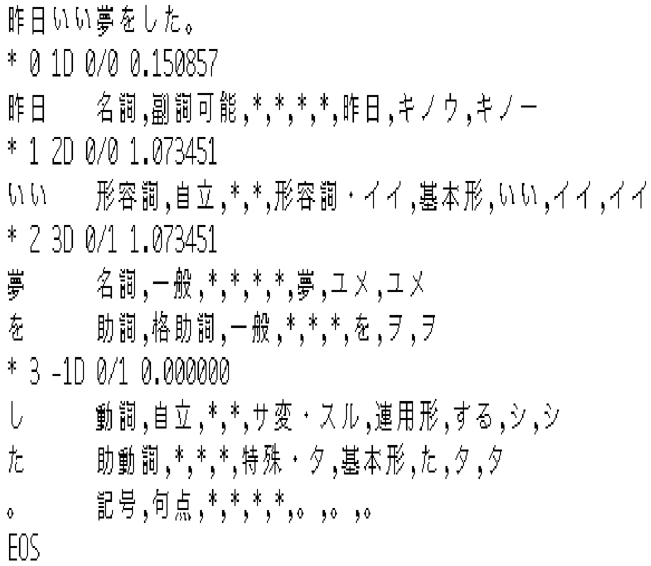
Construction Type	Representation	Case Particle	Grammatical Function
Object-verb	noun wo verb (noun-を-verb)	wo (を)	Object
Subject-verb	noun ga verb (noun-が-verb)	ga (が)	Subject
Dative-verb	noun ni verb (noun-に-verb)	wo (に)	Dative (object/location)

Table 3.1: The three noun-verb construction types in this study.

The patterns are extracted by using the Japanese dependency parser Cabocha[1]. By using a dependency parser, improvements can be gained in the quality of the extraction results: more collocation types and collocation instances can be retrieved from the source text and the noise of the extraction is, at the same time, reduced ([57]).

The Cabocha parser uses a Cascaded Chunking Model, which parses a sentence determiniscally focusing on whether a sentence segment modifies a segment

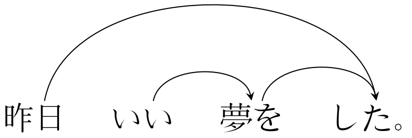
[1] http://taku910.github.io/cabocha/



```
昨日いい夢をした。
* 0 1D 0/0 0.150857
昨日    名詞,副詞可能,*,*,*,*,昨日,キノウ,キノー
* 1 2D 0/0 1.073451
いい    形容詞,自立,*,*,形容詞・イイ,基本形,いい,イイ,イイ
* 2 3D 0/1 1.073451
夢      名詞,一般,*,*,*,*,夢,ユメ,ユメ
を      助詞,格助詞,一般,*,*,*,を,ヲ,ヲ
* 3 -1D 0/1 0.000000
し      動詞,自立,*,*,サ変・スル,連用形,する,シ,シ
た      助動詞,*,*,*,特殊・タ,基本形,た,タ,タ
。      記号,句点,*,*,*,*,。,。,。
EOS
```

Figure 3.2: The output of Cabocha after parsing the sentence 昨日いい夢をした (*kinou ii yume wo shita*, lit. 'I did a good dream yesterday').

on its right hand side [28]. Each segment is called *bunsetsu*, which is the basic unit of Japanese expressions and consists of either one or more morphemes that can be separated into content and grammatical or functional portions. The learner's sentence 昨日いい夢をした (*kinou ii yume wo shita*, lit. 'I did a good dream yesterday') (Figure 3.1) is parsed as follows. The arrows point from dependent bunsetsu to their heads. Every sentence has a root, which doesn't depend on anything (e.g. a final verb).



The output of Cabocha is shown in Figure 3.2. With this information, we can extract the target noun-verb phrase *夢をする (*yume wo suru*, lit. 'do a dream') from the sentence.

3.2.2 Potential Collocation Error Detection

In this step, all noun-verb phrases extracted from the learner's input text are checked against a collocation database. The phrases that do not exist in the database are flagged as potential collocation errors. For example, the collocation *夢をする would be flagged as a potential collocation error, since it is an erroneous collocation and usually does not occur frequently in Japanese texts.

The collocation database was created by extracting all noun-verb pairs that occurred 3 or more times in the reference corpus. In theory, any pair of words that co-occur at least twice in a corpus is a potential collocation. However, to reduce the enormous amounts of data that have to be processed, it is common to apply frequency thresholds. Following the work of Evert [13], common threshold values for collocation extraction are 3, 5 and 10. The association strength in each pair was then measured by applying the Weighted Dice coefficient [27] to a reference corpus. The reference corpus used is the Balanced Corpus of Contemporary Written Japanese or BCCWJ [38], since it covers a variety of text categories. The portions of BCCWJ used in our experiments included magazine, newspaper, textbook and blog data. In addition, in order to include the learners' vocabulary, we used 1,288,934 sentences from the corrected sentences of Lang-8 (year 2010 data) to the reference corpus. The data is described in Table 3.2.

3.2.3 Candidate Generation

In this stage, the system generates a set of collocation correction candidates. The current system does not detect which component (noun or verb) is wrong in a noun-verb construction. Therefore, the learner must first choose which component should be corrected by the system. The system will then construct alternative phrases by substituting the chosen word (noun or verb) with words that are found in its confusion set. The confusion set of a word contains all words that it was corrected to in the learner corpus. For instance, the confusion set of the verb する (*suru*, 'to do') include verbs such as 受ける (*ukeru*, 'to accept'), 始める(*hajimeru*, 'to begin') and 見る (*miru*, 'to see'). Therefore, to correct the verb in the phrase *夢をする (*yume wo shita* , lit. 'to do a dream'), the system generates a set containing candidate pairs such as 夢を受ける (*yume wo ukeru*,

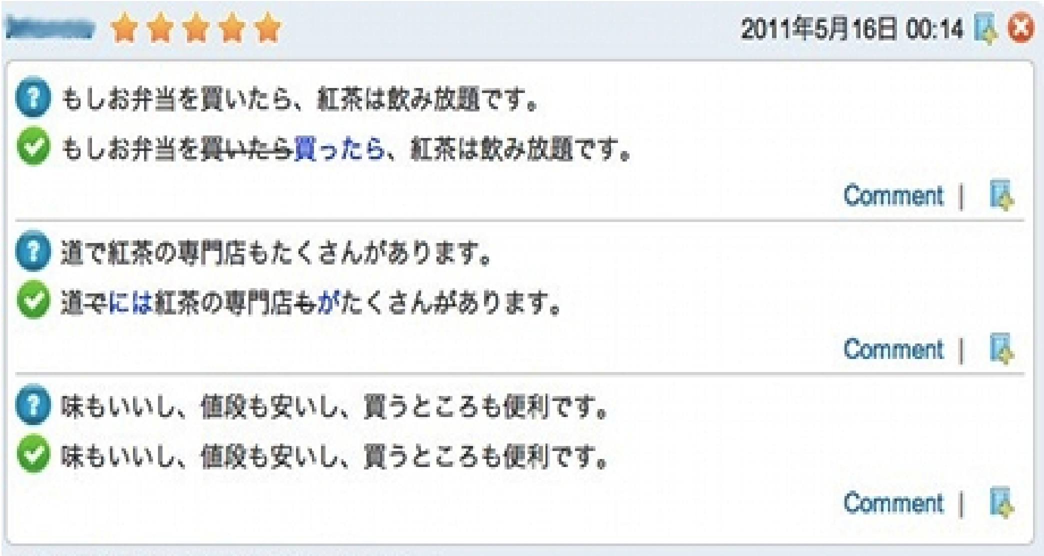
Data	BCCWJ	Lang-8
Size	871,184 sentences (54.81M tokens)	1,288,934 sentences (corrected sentences) (14M tokens)
# noun-wo-verb pairs	194,036	163,880
# noun-ga-verb pairs	216,755	63,312
# noun-ni-verb pairs	300,362	25,787
# unique nouns	43,243	38,999
# unique verbs	18,212	16,086

Table 3.2: Specification of the data used in the candidate ranking step.
Note: All noun-verb pairs were extracted by using the Japanese dependency parser Cabocha.

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’). The learner’s original phrase is also included in this set.

The confusion set is constructed by using the Lang-8[2] learner corpus. A learner corpus is a computerized collection of texts produced by language learners. This corpus was created by crawling the revision log of a language learning social networking service (SNS), Lang-8[3] [43]. 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 [6]. 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] http://cl.naist.jp/nldata/lang-8/
[3] http://lang-8.com/



[redacted] ★★★★★ 2011年5月16日 00:14

? もしお弁当を買いたら、紅茶は飲み放題です。

✓ もしお弁当を~~買いたら~~買ったら、紅茶は飲み放題です。

? 道で紅茶の専門店もたくさんがあります。

✓ 道~~で~~には紅茶の専門店~~も~~がたくさんがあります。

? 味もいいし、値段も安いし、買うところも便利です。

✓ 味もいいし、値段も安いし、買うところも便利です。

Figure 3.3: The interface of Lang-8 website.

3.3. Lang-8 does not provide information about the proficiency level of the users.

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 (Table 3.4).

Table 3.5 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').

3.2.4 Candidate Ranking

In this stage, the collocation database (described in Section 3.2.2) again is used to rank the candidate pairs. The candidate pairs that do not exist in the database are regarded as improper collocations and are left out. For each remained candidate

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%

Table 3.3: Top L1s in Lang-8 data.
Note. Unknown represents the percentage of sentences where the users did not inform their L1.

pair, the database returns a score which reflects its association strength . Finally, the system returns to the user the candidate pairs ranked by the association score.

All measures presented in Section 2.2 have been tested for ranking candidates in our task. Among these, we retained the Weighted Dice coefficient [27] as the default association measure in our system. This choice is both theoretically (as discussed in Section 2.2) and empirically motivated (as it will be discussed in Section 3.4.3).

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 3.4: Specification of the data used in the candidate generation step. *Note*: All noun-verb pairs were extracted by using the Japanese dependency parser Cabocha.

3.3 Other Methods for Generating Collocation Candidates

Most existing systems for collocation error correction assume that collocation errors are mainly caused by confusion of sense relations. Therefore, these systems generate collocation correction candidates by substituting the learner's word choice with words that have similar meaning. In other words, the confusion set of the learner's word choice contains words that have similar meaning. For example, for correcting the verb in the phrase *夢をする (*yume wo suru*, lit. 'to do a dream'), alternative phrases generated by these systems can be phrases such as 夢をやる (*yume wo yaru*, lit. 'to do a dream'), 夢を行う(*yume wo okonau*, lit. 'to perform a dream') and 夢を作る (*yume wo tsukuru*, lit. 'to make a dream'), since やる (*yaru*, 'to do'), 行う(*okonau*, 'to perform') and 作る (*tsukuru*, 'to make') have similar meaning. Common approaches used by these systems for creating the confusion set are the thesaurus-based word similarity and distributional similarity. Another method that follows the same intuition and that can be also applied is the distributed representation of words. Each of these methods are described in the following sections.

	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

Table 3.5: Confusion set for the words する (*suru*, ‘to do’) and光 (*hikari*, ‘light’).

3.3.1 Thesaurus-based Word Similarity

A thesaurus is a hierarchically organized lexical resource that groups words according to similarity. The basic approach of the thesaurus-based 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 pre-defined threshold length exists. For example, by applying this method to generate the confusion set of 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 generating the confusion set of 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’).

3.3.2 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 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 on-line thesauri ([23], 692). The basic idea of this method is that two words are considered similar if they have similar word contexts (Harris, 1954). Each word and its context are represented as co-occurrence vectors (Table 3.6 and Table 3.7).

	書く write	読む read	つける attach
日記を diary	15	11	8

Table 3.6: Context of a particular noun represented as a co-occurrence vector.

	ご飯を rice	ラーメンを ramen noodle	カレーを curry
食べる eat	164	53	39

Table 3.7: Context of a particular verb represented as a co-occurrence vector.

Context can be defined by a grammatical dependency relation (e.g. verb-object).

To compute similarity between two word vectors $\vec{v}$ and $\vec{w}$, a similarity metric is used. A popular metric is the Jensen-Shannon divergence [32] defined by the following formula:

$$JS(P||Q) = KL\left(P||\frac{P+Q}{2}\right) + KL\left(Q||\frac{P+Q}{2}\right) \tag{3.1}$$

P and Q are defined as:

$$P(x_i) = \frac{v_i}{\sum_{i=1}^{N} v_i} \tag{3.2}$$

$$Q(x_i) = \frac{w_i}{\sum_{i=1}^{N} w_i} \tag{3.3}$$

where v_i denotes the ith component of vector $\vec{v}$.

KL is the Kullback-Leibler divergence [29] and it is defined as:

$$KL(P||Q) = \sum_{x} P(x) \log \frac{P(x)}{Q(x)} \tag{3.4}$$

For example, by using the Jensen-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.

3.3.3 Distributed Representation of Words

As mentioned previously, distributional similarity models tend to give better coverage than thesaurus-based methods. However, there are two main limitations of this method: 1) the vectors are long since they have the same length as the size of the vocabulary and 2) the vectors are sparse: most elements are zero. Consequently, for words that are rare in the training data, the model parameters will be poorly estimated [67]. To overcome these limitations, several researchers have investigated unsupervised methods for inducing word representations over large unlabeled corpora. One approach is to learn a distributed representation. A distributed representation of a word (word embedding) is a vector of features which characterize the meaning of the word and are not mutually exclusive. These vectors are dense, low dimensional and real-valued. Each dimension of the vector represents a latent feature of the word that captures useful syntactic and semantic properties [67]. The advantage of this approach is that it allows the model to generalize well to sequences that are not in the set of training word sequences, but that are similar in terms of their features [2]. Word embeddings are typically induced using neural language models, which are language models based on neural networks as the underlying predictive model [67]. A language model is an algorithm that captures the salient statistical characteristics of the distribution of sequences of words in a natural language, typically allowing one to make probabilistic predictions of the next word given preceding ones[2]. On the other hand, a neural network is a computational model that provides a robust approach to approximating real-valued, discrete-valued and vector-valued target functions [42]. A neural network is composed of a number of nodes, or units, connected by links and it is usually arranged in layers. A numeric weight is associated with each link and learning usually takes place by updating these weights to best fit a training set.

Mikolov et al. [40] proposed the implementation of the continuous bag-of-words (CBOW) and Skip-gram algorithms for computing word embeddings. These implementations are based on a neural network architecture with the hidden layer replaced by a projection layer to reduce the computational cost. In our experiments, we use the CBOW model, which worked best in our task. In the CBOW model, the training objective is to combine the representations of surroundings words to predict the word in the middle (Figure 3.4). In other words, given a sequence of training words w_1, w_2, w_3...,w_T, the objective of the CBOW model is to maximize the log probability

$$\frac{1}{T}\sum_{t=1}^{T} \log p(w_t|w_{t-k}^{t+k}) \tag{3.5}$$

where T is the total number of words in the training corpus and w_{t-k}^{t+k} is the set of words in the windows of size k centered at word w_t (with w_t excluded). The probability $p(w_t|w_{t-k}^{t+k})$ is defined as

$$\frac{\exp\left(\vec{e'_{w_t}}^{\top} \cdot \sum_{-c\leq j\leq c, j\neq 0} \vec{e_{w_{t+j}}}\right)}{\sum_{w}\exp\left(\vec{e'_{w_t}}^{\top} \cdot \sum_{-c\leq j\leq c, j\neq 0} \vec{e_{w_{t+j}}}\right)} \tag{3.6}$$

where $\vec{e_{w_t}}$ and $\vec{e'_{w_t}}$ represent the input and output embeddings respectively, i.e., the assignments to the latent variables for word w. In order to reduce the complexity of of the computation, the techniques of hierarchical softmax and negative sampling are used. The model is trained using stochastic gradient descent. The gradient is computed using the backpropagation rule. This algorithm is used to tune the network parameters to best fit a training set of input-output pairs [42] by repeatedly calculating the mean square error of the output response to the sample input until the error value is minimized.

3.4 Evaluation on Learner Data

We conducted an automatic evaluation on real examples of learner data. The evaluation was divided into three parts: selection of an association measure,

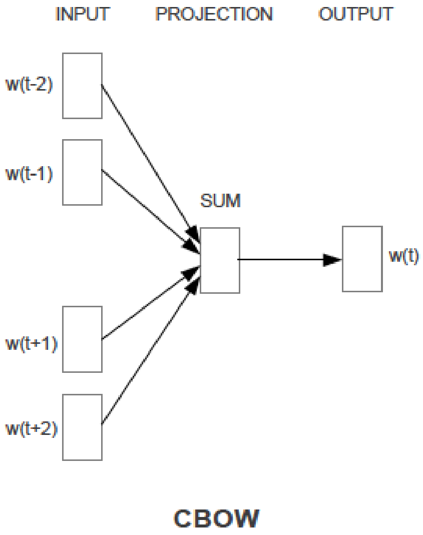


Figure 3.4: Graphical representation of the CBOW model. In this model, the distributed representation of context (or surrounding words) are combined to predict the word in the middle.

a verb suggestion task and a noun suggestion task. For the selection of the association measure part, the aim was to find the most appropriate measure for ranking the collocation candidates. The verb suggestion task and noun suggestion task evaluated the performance of our system against existing approaches. For the verb suggestion task, the 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, the 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. All experiments were conducted on a test set constructed from Lang-8. The construction of this test set is detailed in the following section.

3.4.1 Test Set Construction

Before this work began, no standard test set was available. In order to evaluate our own experiments, we were compelled to develop an appropriate test set on our own.

	Total	f $\geq$ 5	5 > f $\geq$ 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

Table 3.8: Statistics of the extracted pairs from Lang-8 (2011 data).
Note: f stands for frequency. All noun-verb pairs were extracted by using the Japanese dependency parser Cabocha.

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 (26M 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 3.8 shows the statistics of the extracted pairs. These pairs of corrections are nounverb 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 ∗光を付ける→ 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:

	# correction pairs
Verb Suggestion	317
Noun Suggestion	213

Table 3.9: Test Set obtained after manual annotation.

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 3 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 3.9 summarizes the test set obtained after annotation. This test set was used for evaluation in our experiments.

3.4.2 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 5 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), which corresponds to the maximum value of the confusion set size when using Lang- 8 data for generating the confusion set. Recall was computed using the following formula:

$$\frac{tp}{fn + tp} \tag{3.7}$$

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 [69]. 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:

$$MRR = \frac{1}{N}\sum_{i=1}^{N} = \frac{1}{rank(i)} \tag{3.8}$$

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

3.4.3 Evaluation on Selection of an Association Measure

We evaluated the performance of PMI, log-likelihood ratio, Dice coefficient and Weighted Dice coefficient in the verb suggestion task. We first built four systems which implemented the candidate generation and candidate ranking stages as follows:

Data	Mainichi Shimbun
Size	996,219 sentences (M tokens)
# noun-wo-verb pairs	224,185
# unique nouns	16,781
# unique verbs	37,300

Table 3.10: Specification of the Mainichi Shimbun data used in the selection of an association measure experiment.
Note: All noun-verb pairs were extracted by using the Japanese dependency parser Cabocha.

Association Measure	P@1	P@5	Recall	MRR
PMI	0.28	0.60	0.93	0.39
Log-likelihood ratio	0.17	0.63	0.93	0.33
Dice coefficient	0.39	0.67	0.93	0.52
Weighted Dice coefficient	0.42	0.69	0.93	0.56

Table 3.11: The precision and recall rates of models of association measures computed using BCCWJ corpus.

- Candidate Generation: For each test instance, the confusion set of the verb was defined as all the nouns that collocate with it at least three times (f $\geq$ 3) in the BCCWJ Corpus (same data as described in Table 3.2).

- Candidate Ranking: Each system applied one of the four association measures (PMI, log-likelihood ratio, Dice measure and Weighted Dice) to compute the association strength between the words in each candidate using a collocation database constructed from the BCCWJ Corpus (same data as described in Table 3.2).

To further validade the results, we built another four systems by replacing the BCCWJ corpus with another corpus, Mainichi Shimbun [61]. This corpus consists of 996, 219 sentences from Japanese newspaper articles.

Association Measure	P@1	P@5	Recall	MRR
PMI	0.22	0.56	0.82	0.28
Log-likelihood	0.20	0.54	0.82	0.24
Dice coefficient	0.30	0.62	0.82	0.38
Weighted Dice coefficient	0.29	0.65	0.82	0.40

Table 3.12: The precision and recall rates of models of association measures computed using Mainichi Shimbun corpus.

Results

The results are shown in Table 3.11 and Table 3.12. As we can see in both tables, the Weighted Dice coefficient yielded the highest precision values for most cases and was therefore chosen for constructing the other models. In general, PMI and log-likelihood ratio performed similarly, assigning high scores to pairs that had a lower co-occurrence frequency than the correction assigned in the learner corpus, while Dice and Weighted Dice assigned lower scores to such pairs. In other words, PMI and the log-likelihood obtained lower performance because they assigned higher scores to expressions that are not commonly used by learners. The Weighted Dice coefficient performed better than the Dice coefficient, assigning even higher scores to higher frequency pairs, i.e. they assigned higher scores to expressions that are more commonly used by learners. Table 3.13 shows some examples of the correct collocation as ranked by each association measure.

3.4.4 Evaluation on Verb Suggestion and Noun Suggestion tasks

In order to compare the performance of our system with existing approaches, we built three baseline systems by combining each of the other existing methods or generating collocation candidates (thesaurus-based word similarity method, distributional similarity and distributed representation of words), with the Weighted Dice coefficient, for ranking the candidates. Henceforth, Thesaurus+WD refers to the model that used a thesaurus for generating candidates and the Weighted

	Misused Noun+Verb			Correction			Rank of correction			
							PMI	LL	Dice	WD
Japanese Reading Meaning	教育 kyouiku education	を wo	*もらう morau get	教育 kyouiku education	を wo	受ける ukeru receive	7	12	5	1
Japanese Reading Meaning	汗 ase sweat	を	*出る deru come out	汗 ase sweat	を	かく kaku perspire	2	4	1	1
Japanese Reading Meaning	言葉 kotoba word	を	言う iu say	言葉 kotoba word	を	使う tsukau use	30	48	6	3

Table 3.13: The correct verb as ranked by each association measure.
Note. The strength of association was calculated using the BCCWJ corpus.

Dice coefficient (WD) for computing the association strength and ranking the candidates. The DS+WD model used distributional similarity (DS) for generating candidates. The DR+WD model used distributed representation (DR) of words for generating candidates. CS Lang-8+WD, our proposed model, generated candidates using the confusion set (CS) based on correction pairs from Lang-8.

All models ranked the candidates pairs by their Weighted Dice coefficient score. The score is obtained from the collocation database we built (described in Section 3.2.2). Therefore, the models implemented by each system differ only in the candidate generation stage. The candidate generation stage in each model was implemented as follows:

- Thesaurus+WD: This model generated the confusion set by using the Bunrui Goi Hyo Thesaurus, a Japanese thesaurus composed of 87,743 words that are classified into 32,636 unique semantic classes. The thesaurus-based word similarity method selects words in the same sub-tree as candidate words. By sub-tree, we mean the tree with distance 2 from the leaf node (the learner's written word) doing a pre-order tree traversal, which gives an average number of 28 candidates in the confusion set.

- DS+WD: This model generated the confusion set by applying the Jensen-Shannon divergence to a corpora. The corpora used are BCCWJ and the corrected sentences of Lang-8 (same data described in Table 3.2). When constructing the vectors, context is defined by the object-verb, subject-verb or dative-verb grammatical dependency relation. We also tested other distributional similarity measures: Cosine Similarity, Dice measure [9], and Kullback-Leibler divergence (KL divergence) [29]. However, the Jenson-Shannon divergence obtained the best performance. The detailed evaluation that we conducted is reported in Appendix A.

- DR+WD: This model generated the confusion set by applying the CBOW model [40] to a corpora. The corpora used are BCCWJ and the corrected sentences of Lang-8 (same data described in Table 3.2). To implemente the CBOW model, we used the *word2vec*[4] package. The configuration used in our experiment was with the context windows of size 8 and with vector dimensionality of 1000, which worked best in our setting. Moreover, we used the negative subsampling with threshold at $t = 1e^{-4}$.

- CS Lang-8+WD: Our proposed model generated candidates using the confusion set (CS) by using correction pairs from Lang-8 (detailed in Section 3.2.3).

To explore the tendency of the results, we first evaluated on object-verb constructions. As Table 3.2 and Table 3.8 show, this is the most common noun-verb construction type in the learners' writing.

Results

Table 3.14 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

[4] https://code.google.com/p/word2vec/

are shown in Table 3.15. However, for cases wherein the learner's word choice and the correction suggested by the Lang-8 data did not have a similar meaning, 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 models that used distributional similarity (DS+WD) and distributed representations (DR+WD). This means that the correct candidate could be generated for many cases. However, the precision rate decreased with the distributional similarity because the correction obtained a low rank in the collocation suggestion list. On the other hand, with the distributed representation model, the precision was high, but the recall was lower than the obtained with the distributional similarity model for the verb suggestion task. 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 3.22 shows examples in which the model that used a thesaurus and distributed representations 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 10 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 three models.

A similar phenomenon occurred for the noun suggestion task. Table 3.23 shows some examples of the ranking for the corrections assigned by all four models.

We applied our CS Lang-8+WD model to subject-verb (noun ga verb) and dative-verb (noun ni verb) constructions as well. Table 3.16 and Table 3.17

	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
DR+WD	0.78	0.97	0.43	0.37	0.60	0.86	0.66	0.47
CS Lang8 +WD	0.63	0.89	0.95	0.72	0.63	0.97	0.86	0.66

Table 3.14: The precision, recall and MRR of different models applied to object-verb constructions.
Note: WD stands for Weighted Dice, DS stands for Distributional Similarity, DR stands for Distributed Representation and CS Lang8 stands for confusion set from the Lang-8 corpus.

	Misused Noun+Verb			Correction			Rank of correction
Japanese	話	を	*聴く	話	を	聞く	1
Reading	hanashi	wo	kiku	hanashi	wo	kiku	
Meaning	talk		listen	talk		listen	
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	薬	を	*食べる	薬	を	飲む	1
Reading	kusuri	wo	taberu	kusuri	wo	nomu	
Meaning	medicine		eat	medicine		drink	

Table 3.15: Rank of correct verb given by the model that used a thesaurus for generating the correction candidates.

summarize the results for verb and noun suggestions. For both subject-verb and dative-verb constructions, the system obtained high recall and MRR values.

3.4.5 Evaluation on Different Collocation Types

We started our research by investigating the suggestion on noun-verb collocations as it has been argued to be usually the most difficult for learners to master. However, collocation usages cover a wider range of word usages, such as noun-adjective and adverb-verb. Different types of collocational usages are also worthwhile to be evaluated to show the feasibility of our system. In the experiment on adjective-noun collocations, we applied our proposed model (CS Lang-8+WD) described in Section 3.4.4. The specification of the data used in the candidate generation and candidate ranking steps is given in Table 3.18 and Table 3.19. We also applied the baseline models (described in Section 3.4.4) for comparison.

The test set was built by applying the same method described in 3.4.1. The specification of the test set is given in Table 3.20.

The results obtained by each model are described in Table 3.21. Compared with noun-verb collocations, the total number of noun-adjective pairs is far smaller (Table 3.18 and Table 3.19) and so the number of features extracted. Therefore, the result might not be as satisfactory as the results obtained in the evaluation on noun-verb constructions; however, it shows that the proposed method still copes with this task better than the baseline methods. Using a two-tailed t-test with a confidence interval of 99%, we measured the statistical significance. Our proposed model performed significantly better than the other three models.

Table 3.24 and Table 3.25 give some examples of the ranking of the correction given by our proposed method and baseline methods.

3.4.6 System Limitations

The experiment results show that our proposed method does provide the satisfactory suggestion performance. However, we observed that some limitations did lie in our system and it is noteworthy for further discussion. The limitations of the system can be categorized into two main types (Table 3.26):

	Verb Suggestion				Noun Suggestion			
System	*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

Table 3.16: The precision, recall and MRR of the Confusion Set from Lang-8 and Weighted Dice measure combinations applied to subject-verb constructions. Note: CS Lang8 stands for confusion set from the Lang-8 corpus and WD stands for Weighted Dice.

	Verb Suggestion				Noun Suggestion			
System	*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

Table 3.17: The precision, recall and MRR of the Confusion Set from Lang-8 and Weighted Dice measure combinations applied to dative-verb constructions. Note: CS Lang8 stands for confusion set from the Lang-8 corpus and WD stands for Weighted Dice.

Data	Lang-8
Size	1,288,934 pairs of learner's sentences and corrections (37.5M tokens)
# noun-ga-adjective pairs	23,227
# unique nouns	6,416
# unique adjectives	1,138

Table 3.18: Specification of the data used in the candidate generation step.
Note: All pairs were extracted by using the Japanese dependency parser Cabocha.

Data	BCCWJ	Lang-8
Size	871,184 sentences (54.81M tokens)	1,288,934 sentences (corrected sentences) (14M tokens)
# noun-ga-adjective pairs	58,337	23,227
# unique nouns	15,397	6,416
# unique verbs	2,203	1,138

Table 3.19: Specification of the data used in the candidate ranking step.
Note: All pairs were extracted by using the Japanese dependency parser Cabocha.

	# correction pairs
Adjective Suggestion	58
Noun Suggestion	57

Table 3.20: Test Set for noun-adjetive constructions.

	Adjective Suggestion				**Noun Suggestion**			
System	*P@1*	*P@5*	*Recall*	*MRR*	*P@1*	*P@5*	*Recall*	*MRR*
Thesaurus+WD	0.15	0.16	0.09	0.08	0.46	1.00	0.23	0.14
DS+WD	0.30	0.30	0.18	0.21	0.18	0.64	0.19	0.06
DR+WD	0.52	1.00	0.62	0.39	0.27	0.73	0.72	0.33
CS Lang8 +WD	0.67	0.98	0.78	0.61	0.39	0.72	0.82	0.44

Table 3.21: The precision, recall and MRR of different models applied to noun-adjective constructions.
Note: WD stands for Weighted Dice, DS stands for Distributional Similarity, DR stands for Distributed Representation and CS Lang8 stands for confusion set from the Lang-8 corpus.

	Misused Noun+Verb			Correction			Rank of correction			
							Thesaurus + WD	DS + WD	DR + 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	嘘	を	*作る	嘘	を	つく	-	1	-	1
Reading	uso	wo	tsukuru	uso	wo	tsuku				
Meaning	lie		make	lie		attach				
Japanese	夢	を	*する	夢	を	見る	-	4	-	1
Reading	yume	wo	suru	yume	wo	miru				
Meaning	dream		do	dream		see				
Japanese	汗	を	*流す	汗	を	かく	-	1	-	1
Reading	ase	wo	nagasu	ase	wo	kaku				
Meaning	sweat		pour	sweat		perspire				
Japanese	知識	を	*取る	知識	を	得る	-	1	1	1
Reading	tishiki	wo	toru	tishiki	wo	eru				
Meaning	knowledge		take	knowledge		obtain				

Table 3.22: Rank of correct verb given by the models that used a thesaurus (Thesaurus+WD), distributional similarity (DS+WD), distributed representation (DR+WD) and confusion set derived from Lang-8 (CS Lang-8+WD).

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

Table 3.23: Rank of correct noun given by the models that used a thesaurus (Thesaurus+WD), distributional similarity (DS+WD), distributed representation (DR+WD) and confusion set derived from Lang-8 (CS Lang-8+WD).

	Misused Noun+Adjective			Correction			Rank of correction			
							Thesaurus + WD	DS + WD	DR + WD	CS Lang8 + WD
Japanese	風	が	*大きい	風	が	強い				
Reading	kaze	ga	ōkī	kaze	ga	tsuyoi	-	-	1	1
Meaning	wind		big	wind		strong				
Japanese	気温	が	*寒い	気温	が	低い				
Reading	kion	ga	samui	kion	ga	hikui	-	-	-	1
Meaning	air temperature		cold	air temperature		low				
Japanese	値段	が	*低い	値段	が	安い				
Reading	nedan	ga	hikui	nedan	ga	yasui	-	-	1	1
Meaning	price		low	price		cheap				
Japanese	数	が	*大きい	数	が	多い				
Reading	kazu	ga	ookii	kazu	ga	ooi	-	1	2	1
Meaning	a number		big	a number		many				
Japanese	水	が	*涼しい	水	が	冷たい				
Reading	mizu	ga	suzushii	mizu	ga	tsumetai	1	1	1	1
Meaning	water		cool	water		cold				

Table 3.24: Rank of correct adjective given by the models that used a thesaurus (Thesaurus+WD), distributional similarity (DS+WD), distributed representation (DR+WD) and confusion set derived from Lang-8 (CS Lang-8+WD).

	Misused Noun+Adjective			Correction			Rank of correction			
							Thesaurus + WD	DS + WD	DR + WD	CS Lang8 + WD
Japanese	*太陽	が	強い	日差し	が	強い				
Reading	taiyō	ga	tsuyoi	taiyō	ga	tsuyoi	-	-	2	1
Meaning	sun		strong	sunlight		strong				
Japanese	*体	が	悪い	体調	が	悪い				
Reading	karada	ga	ōkī	taichō	ga	warui	-	-	-	2
Meaning	body		bad	physical condition		bad				
Japanese	*暇	が	ない	時間	が	ない				
Reading	hima	ga	nai	jikan	ga	nai	-	-	1	1
Meaning	free time		lack	time		lack				

Table 3.25: Rank of correct noun given by the models that used a thesaurus (Thesaurus+WD), distributional similarity (DS+WD), distributed representations (DR+WD) and confusion set derived from Lang-8 (CS Lang-8+WD).

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 four methods we evaluated: i) thesaurus-based method, ii) distributional similarity, ii) distributional representation and iv) confusion set generated from learner corpus.

2. 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.

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

Table 3.26: Example of cases where the system fails to offer the correct collocation.
Note: The "-" indicates cases where the system was not able to generate the correct candidate.

3.4.7 Summary of the chapter

In this chapter, we presented a method to correct potential collocation errors in JSL writing. Using corrections extracted from a large annotated Japanese learner corpus, the proposed 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.

Chapter 4

Applications

As we discussed earlier (in Chapter 1), collocations are important in helping language learners achieve native-like fluency. Previous works revealed that collocations are challenging, even for advanced language learners. Moreover, the number of tools designed to target language learner collocation errors is limited. As pointed out by Leacock et al. [31], an application that can detect a learners' collocation errors and suggest the most appropriate "ready-made units" as corrections is an important goal for natural language processing (p. 26). Therefore, we try to utilize the method described in Chapter 3 to propose two end-user applications for pedagogical use.

In the first application, we developed a prototype of a collocational aid, *Collocation Assistant*, aiming at providing writing hints for collocational usages. Given a noun-verb collocation input by a learner, the tool automatically flags possible collocation errors and suggests better collocations. Each suggestion includes several usage examples to help learners choose the best candidate.

For the second application, *JSL Writing Assistant*, we extended the Collocation Assistant, by allowing the user to input a full sentence/paragraph and by allowing the processing of different collocation types as well.

The detailed description of these two applications will be presented in Section 4.1 and Section 4.2. For developing both applications, the following open source frameworks were used:

- Apache Lucene[1]: an open source Java text search engine library.
- Primefaces[2]: an open source User Interface (UI) component library for JavaServer Faces (JSF) based applications. JSF technology is a server-side user interface component framework for Java technology-based web applications.

4.1 Collocation Assistant

The Collocation Assistant is a collocational aid prototype that focuses on providing Japanese collocation suggestions for potential collocation errors in Japanese noun-verb constructions. It allows the user to perform two types of search: single-word and multi-word query search.

- Single-word query search: This search allows the user to input a single word (noun or a verb) and triggers the system to work similar to a concordancer. Given a noun or verb input by the user, the system will suggest collocations containing words that strongly collocate with this input (Figure 4.1). The system makes use of a bilingual Japanese-English dictionary, Edict [3] and also provides the English translation of the input word.
- Multi-word query search: This search allows the user to input a noun-verb collocation and triggers the system to initiate the correction process described in Section 3. The multi-word query has the advantage of targeting the specific word sense the user desires and provides a more precise access to the desired collocation. 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 and a message is displayed to the user. Next, the system suggests more appropriate noun-verb collocations. For instance, if the learner types *夢をする (*yume wo suru* , lit. 'to do a dream'), the system flags a collocation error. When the user clicks on "same noun", the system displays better collocations with the same noun input by the user, such as 夢を見る (*yume wo miru*, 'to dream') and 夢を持つ (*yume*

[1] https://lucene.apache.org/

[2] http://www.primefaces.org/

wo motsu, 'to hold a dream'), as shown in Figure 4.2. Likewise, when the user clicks on "same verb", the system displays better collocations with the same verb input by the user. If the user clicks on "View all suggestions", all possible better collocations with the same noun or the same verb input by the user are displayed. Aside from the collocations, sentence examples for each phrase suggestion are displayed, showing the phrase in context with surrounding text. Showing phrases in context can be crucial in helping users determine which phrase is most appropriate [49]. Even if the learner's input is not flagged as an error, it will undergo the same correction process, since better collocations than the input might exist. In this case, the learner will check the ranked suggestions and sentence examples and choose the most appropriate expression. The current system does not detect which component (noun or verb) is wrong in a noun-verb construction. Therefore, the learner must specify which component would be corrected by the system.

In the interface, the suggested collocations are sorted by the association strength score (Weighted Dice coefficient score) of the collocation in the corpora. Every part of each collocation suggested is highlighted.

4.1.1 Resources used for providing sentence examples

We used several monolingual and bilingual resources for providing useful sentence examples to users. These resources are:

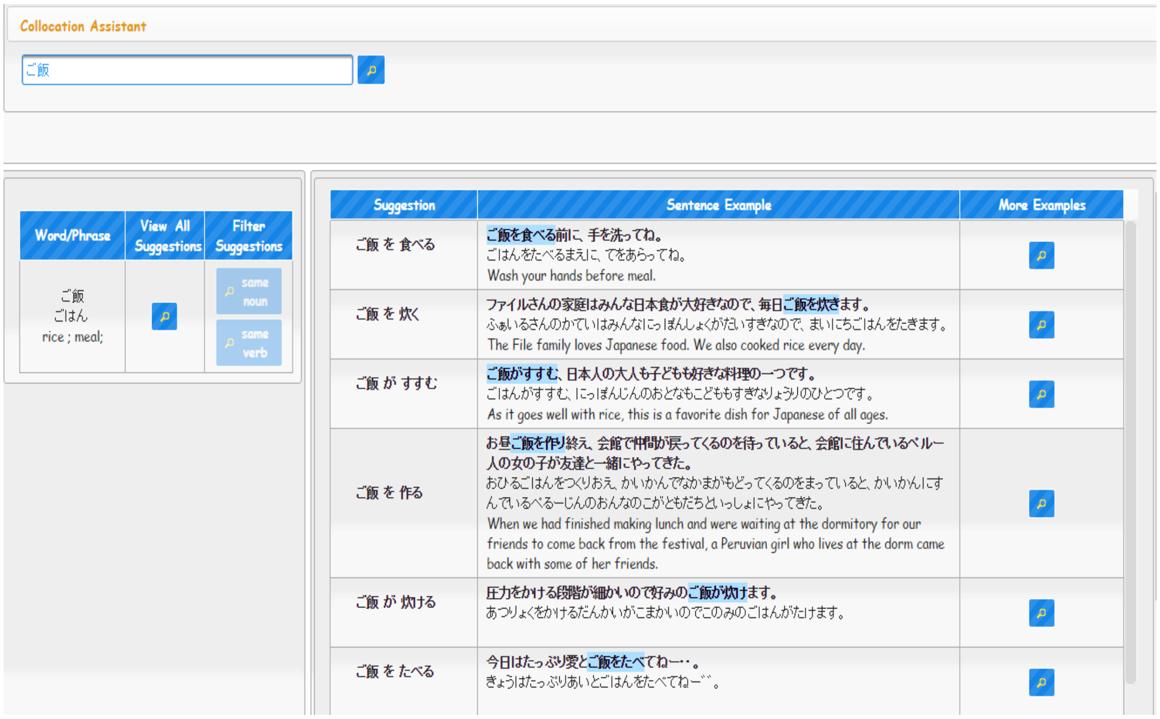
Bilingual resources. The bilingual resources we used consist of Japanese-English parallel corpora. These corpora are:

- Tatoeba Corpus[3], a free collaborative online database of example sentences geared towards foreign language learners. Its name comes from the Japanese term 例えば (*tatoeba*), meaning "for example". Tatoeba focuses on translation of complete sentences into several different languages. We used the Japanese-English sentences available in the website.

- Hiragana Times (HT) Corpus[4], a Japanese-English bilingual corpus of magazine articles of Hiragana Times, a bilingual magazine written in Japanese

[3] https://tatoeba.org/eng/

[4] http://www.hiraganatimes.com/



Word/Phrase	View All Suggestions	Filter Suggestions
ご飯 ごはん rice ; meal;	🔍	same noun same verb

Suggestion	Sentence Example	More Examples
ご飯 を 食べる	ご飯を食べる前に、手を洗ってね。 ごはんをたべるまえに、てをあらってね。 Wash your hands before meal.	🔍
ご飯 を 炊く	ファイルさんの家庭はみんな日本食が大好きなので、毎日ご飯を炊きます。 ふぁいるさんのかていはみんなにっぽんしょくがだいすきなので、まいにちごはんをたきます。 The File family loves Japanese food. We also cooked rice every day.	🔍
ご飯 が すすむ	ご飯がすすむ、日本人の大人も子どもも好きな料理の一つです。 ごはんがすすむ、にっぽんじんのおとなもこどももすきなりょうりのひとつです。 As it goes well with rice, this is a favorite dish for Japanese of all ages.	🔍
ご飯 を 作る	お昼ご飯を作り終え、会館で仲間が戻ってくるのを待っていると、会館に住んでいるペルー人の女の子が友達と一緒にやってきた。 おひるごはんをつくりおえ、かいかんでなかまがもどってくるのをまっていると、かいかんにすんでいるぺるーじんのおんなのこがともだちといっしょにやってきた。 When we had finished making lunch and were waiting at the dormitory for our friends to come back from the festival, a Peruvian girl who lives at the dorm came back with some of her friends.	🔍
ご飯 が 炊ける	圧力をかける段階が細かいので好みのご飯が炊けます。 あつりょくをかけるだんかいがこまかいのでこのみのごはんがたけます。	🔍
ご飯 を たべる	今日はたっぷり愛とご飯をたべてねー･･。 きょうはたっぷりあいとごはんをたべてねー˘˘。	🔍

Figure 4.1: The interface of the Collocation Assistant showing collocations for the noun ご飯 (*gohan*, ‘rice’).

and English to introduce Japan to non-Japanese, covering a wide range of topics (culture, society, history, politics, etc.).

- Kyoto Wikipedia (KW) Corpus[5], a corpus created by manually translating Japanese Wikipedia articles (related to Kyoto) into English.

Monolingual resource: the BCCWJ corpus [38] was used as Japanese monolingual resource for the noun-verb expressions where no bilingual examples were available.

The specification of each resource is given in Table 4.1.

	# *jp* sentences	# *en* sentences
Tatoeba	203,191	203,191
HT	117,492	117,492
KW	329,169	329,169
BCCWJ	871,184	-

Table 4.1: Data used as sentence examples.

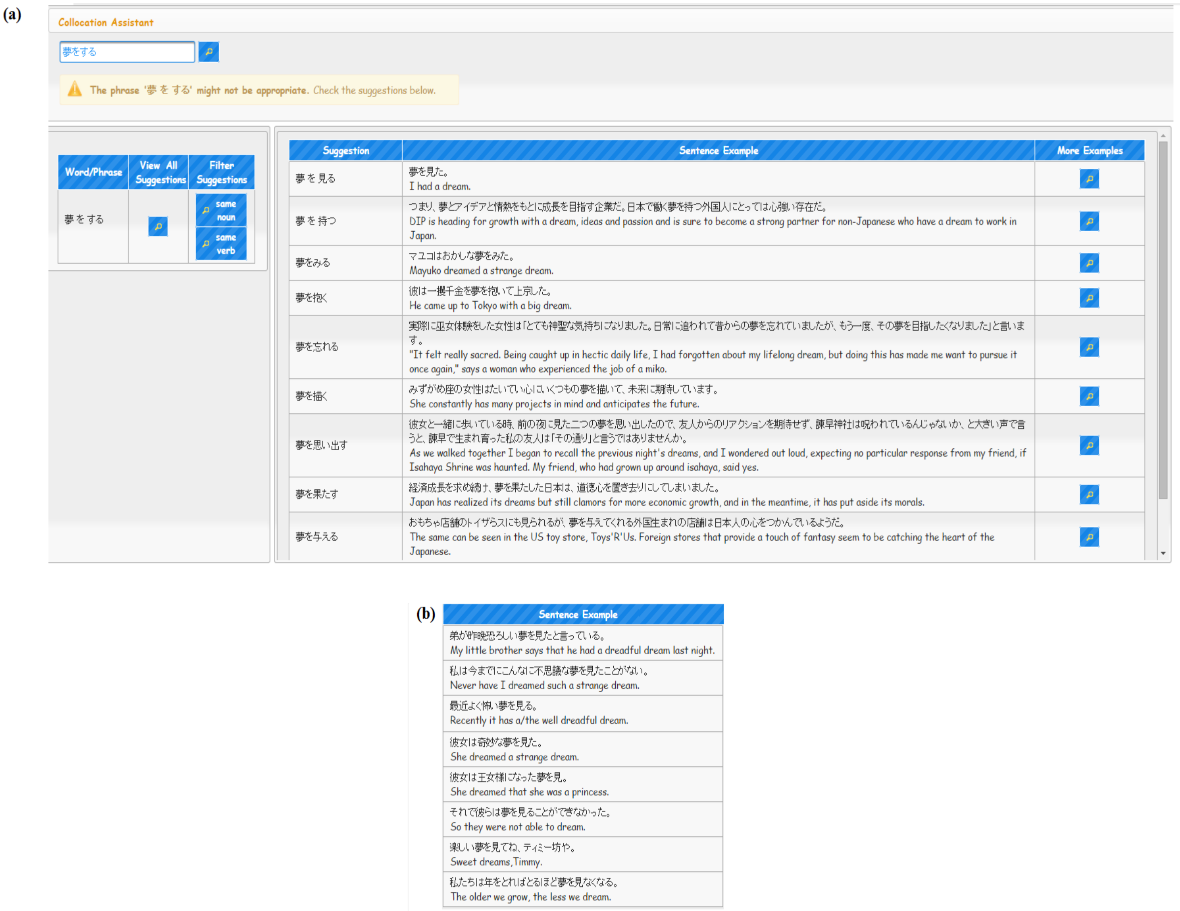
4.1.2 Preliminary User Study of the System

We conducted a preliminary evaluation with JSL learners to gather their feedback on using the Collocation Assistant 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.

Participants

In this study, 10 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, 2 were female and 8 were male and they had different language backgrounds (Chinese, Indonesian, Tagalog, Swahili, Spanish and Basque). Regarding their proficiency level,

[5] https://alaginrc.nict.go.jp/WikiCorpus/



Collocation Assistant

夢をする

The phrase '夢 を する' might not be appropriate. Check the suggestions below.

Word/Phrase	View All Suggestions	Filter Suggestions
夢 を する		same noun same verb

Suggestion	Sentence Example	More Examples
夢 を 見る	夢を見た。 I had a dream.	
夢 を 持つ	つまり、夢とアイデアと情熱をもとに成長を目指す企業だ。日本で働く夢を持つ外国人にとっては心強い存在だ。 DIP is heading for growth with a dream, ideas and passion and is sure to become a strong partner for non-Japanese who have a dream to work in Japan.	
夢をみる	マユコはおかしな夢をみた。 Mayuko dreamed a strange dream.	
夢を抱く	彼は一攫千金を夢を抱いて上京した。 He came up to Tokyo with a big dream.	
夢を忘れる	実際に巫女体験をした女性は「とても神聖な気持ちになりました。日常に追われて昔からの夢を忘れていましたが、もう一度、その夢を目指したくなりました」と言います。 "It felt really sacred. Being caught up in hectic daily life, I had forgotten about my lifelong dream, but doing this has made me want to pursue it once again," says a woman who experienced the job of a miko.	
夢を描く	みずがめ座の女性はたいてい心にいくつもの夢を描いて、未来に期待しています。 She constantly has many projects in mind and anticipates the future.	
夢を思い出す	彼女と一緒に歩いている時、前の夜に見た二つの夢を思い出したので、友人からのリアクションを期待せず、諫早神社は呪われているんじゃないか、と大きい声で言うと、諫早で生まれ育った私の友人は「その通り」と言うではありませんか。 As we walked together I began to recall the previous night's dreams, and I wondered out loud, expecting no particular response from my friend, if Isahaya Shrine was haunted. My friend, who had grown up around isahaya, said yes.	
夢を果たす	経済成長を求め続け、夢を果たした日本は、道徳心を置き去りにしてしまいました。 Japan has realized its dreams but still clamors for more economic growth, and in the meantime, it has put aside its morals.	
夢を与える	おもちゃ店舗のトイザらスにも見られるが、夢を与えてくれる外国生まれの店舗は日本人の心をつかんでいるようだ。 The same can be seen in the US toy store, Toys'R'Us. Foreign stores that provide a touch of fantasy seem to be catching the heart of the Japanese.	

Sentence Example
弟が昨晩恐ろしい夢を見たと言っている。 My little brother says that he had a dreadful dream last night.
私は今までにこんなに不思議な夢を見たことがない。 Never have I dreamed such a strange dream.
最近よく怖い夢を見る。 Recently it has a/the well dreadful dream.
彼女は奇妙な夢を見た。 She dreamed a strange dream.
彼女は王女様になった夢を見。 She dreamed that she was a princess.
それで彼らは夢を見ることができなかった。 So they were not able to dream.
楽しい夢を見てね、ティミー坊や。 Sweet dreams,Timmy.
私たちは年をとればとるほど夢を見なくなる。 The older we grow, the less we dream.

Figure 4.2: An example of collocation suggestions produced by the Collocation Assistant given the erroneous collocation *夢をする (*yume wo suru, lit. ‘to do a dream’*) as input. (a) Collocaiton suggestions are shown on the left and an example sentence for each suggestion is shown on the right. In the example, 夢を見る (*yume wo miru*, ‘to dream’) is the correct collocation. (b) Further examples for each suggestion are shown when the user clicks on “More examples”. In the example, further example sentences for the collocation 夢を見る are displayed.

three were beginners, three were intermediate and four were advanced learners, based on the Japanese-Language proficiency test [6] certificate level they previously obtained. All participants were regular computer users.

Procedure

A collocation test was designed to examine whether or not the tool could help JSL learners find proper Japanese collocations (Table 4.2). 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. Additionally, they were asked to use the tool to write a paragraph in Japanese. 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 7 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.

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 times (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

[6] http://www.jlpt.jp/

For each sentence, try to identify if there are mistakes in the in the noun-wo-verb, noun-ga-verb and noun-ni-verb expressions. Use the collocation assistant system if you need. For each mistake identified, please provide one possible correction. In the end, please rewrite the sentence with the corrections you made.

Question	Sentence
1	でも、毎日家に居て、*エアコンを開けて、これも最高だ。
2	アダルトスケーターの友達一人が来月二級の*テストをとることになった。
3	これから*宿題を書く。
4	私は自分の部屋に入って、*光をつけました。
5	マイクからやさしい*声が出てくる。
6	一番大切なことは*お祈りを願うことです。
7	先生の日の式では、まず、校長先生が学生たちに*スピーチを言う。
8	その時訪問した人は*あいさつをやる。
9	たとえば、公共の場所では喫煙室だけでたばこを吸って、ぜったいほかの人に*被害をかけないようにすることがいいです。
10	先に*スープを食べました。
11	来年私は*留学生をする。
12	また、キリスト教の人は教会に、仏教の人はお寺に行って、*新年を初めます。

Table 4.2: Collocation Test.

Items	Average	Standard Deviation
Difficulty of Japanese collocations		
• I often have doubts about the meaning of Japanese collocations.	6.4	0.7
• The tools I currently use provide help with Japanese collocations.	2.2	1.3
Design of the Collocation Assistant		
• The interface of the tool was easy to use.	6.3	0.8
• The interface of the tool was easy to understand.	6.1	0.7
Usefulness of the Collocation Assistant		
• The tool is helpful for Japanese second language learner students (beginner, intermediate and advanced learners).	5.9	1.2
• The tool is helpful in choosing the proper collocations.	6.6	0.5
Quality of the retrieved data.		
• The collocations suggested by the tool were useful.	6.5	0.7
• The way the collocations are arranged and presented was helpful.	5.8	0.6
• The sentence examples showed by the tool helped me further understand the meaning of an expression.	6.1	0.7

Table 4.3: Results of the student survey.

learners used the system for all sentences and obtained higher scores. The participants had difficulty in correcting two particular long sentences in the test (Table 4.2, questions 5 and 9). They had difficulty in finding sentence examples close to the meaning of the sentences in the test. They also found the tool useful when writing a paragraph in Japanese. 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 other software aids they have 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 tool is simple and intuitive to find the right expressions in Japanese. The rank of the expressions is very useful!”
“The tool is very useful to find examples of usage because when we learn a new word it is easier to memorize it if we know the context and if we see an example.”
“I think this tool can help learners to detect any errors in collocations and in this way it also will increase their writing level.”
“I think this tool is helpful for inexperienced Japanese teacher or non-native Japanese teachers, that might find it difficult giving feedback to their students about the errors they make and/or alternative ways of expressing the same idea.”
“The sample sentences for each expression can be useful, however, some sentences are very long and it is hard to understand the meaning of the expression quickly.”

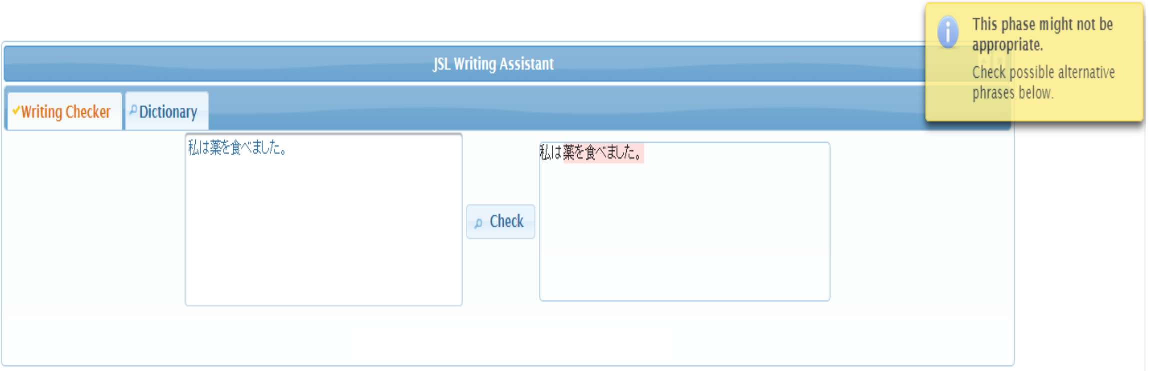
Table 4.4: Feedback from the second part of the student survey.

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. Some of the questions administered in the questionaire are shown in Table 4.3.

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. Examples of the feedback given by the participants are shown in 4.4.

4.2 JSL Writing Assistant

The *JSL Writing Assistant* is an extension of the Collocation Assistant. It allows the user to input a full sentence/paragraph and allows the processing of differ-



This phase might not be appropriate.
Check possible alternative phrases below.

私は薬を食べました。

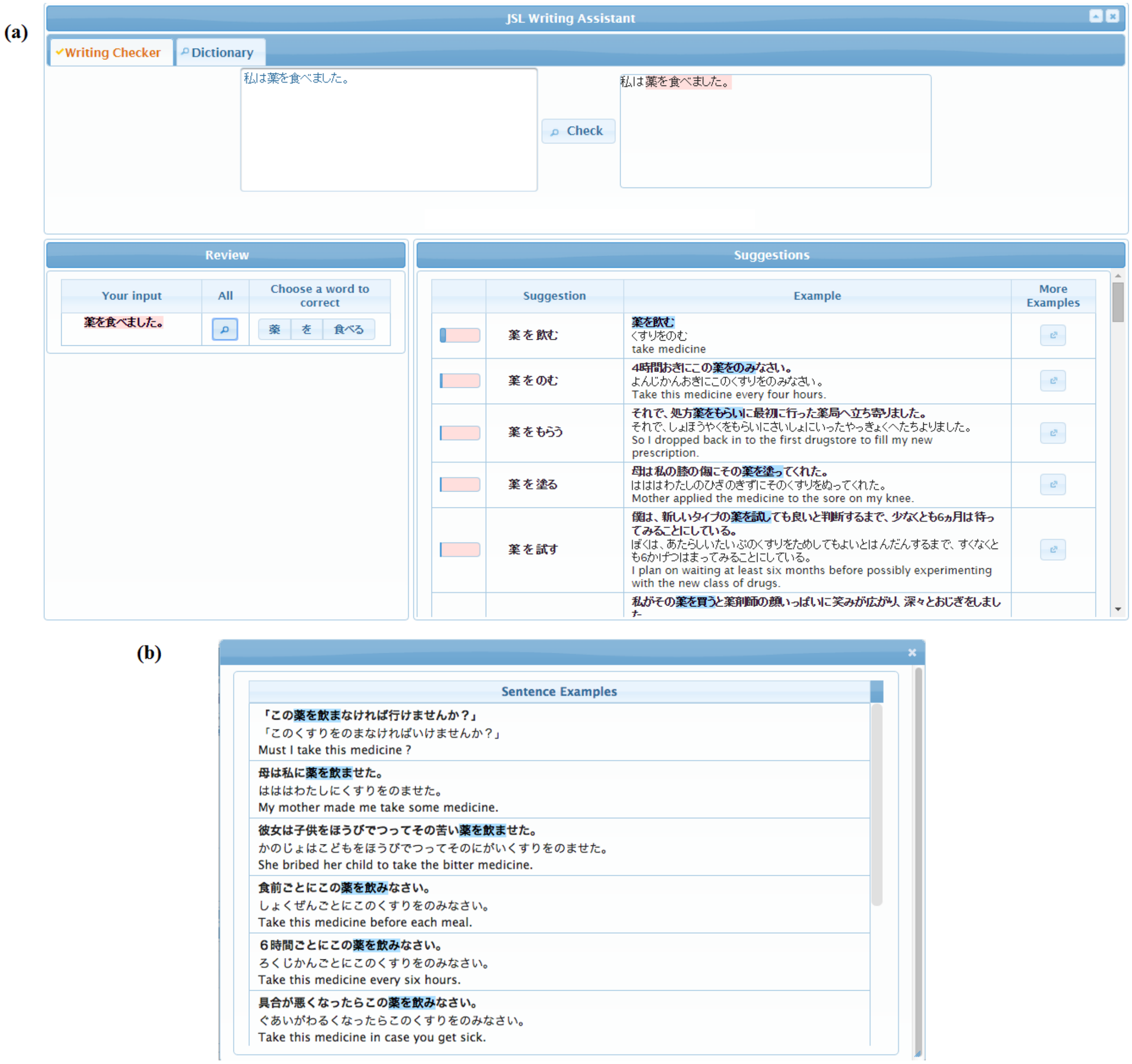
私は薬を食べました。

Figure 4.3: Interface of the JSL Writing Assistant showing in highlight potential collocation errors in the input sentence.

ent collocation types as well. The collocation patterns that the tool supports are: noun-verb, adjective-noun, noun-adjective and adverb-verb. The collocation suggestion module will be launched immediately after users submit their text to the system. The system will first parse the user's input and extract the target collocations. Next, it will initiate the correction process. The potential collocation errors in the users' input will be highlighted (Figure 4.3) and, after clicking on one of the them, the user will be prompted with the collocation suggestions along with sentence examples (Figure 4.4). With this interface, users can write and revise their texts[7].

In the interface, the suggested collocations are sorted by the association strength score (Weighted Dice coefficient score) of the collocation in the corpora. Color bars to the left of each collocate indicate the relative association measure score. Every part of each collocation suggested is highlighted. The "Dictionary" tab allows the user to search for a particular word (noun, verb, adjective or adverb) and check the words that strongly collocate with this input (corresponds to the same single-word query search of the Collocation Assistant). It also shows possible confusable words to the input (in "Check Also"), so the user can check the difference between them and the input. These possible confusable words are generated with the method described in Section 3.2.3.

[7] Demo available at http://cl.naist.jp/collocationassistant/



私は薬を食べました。

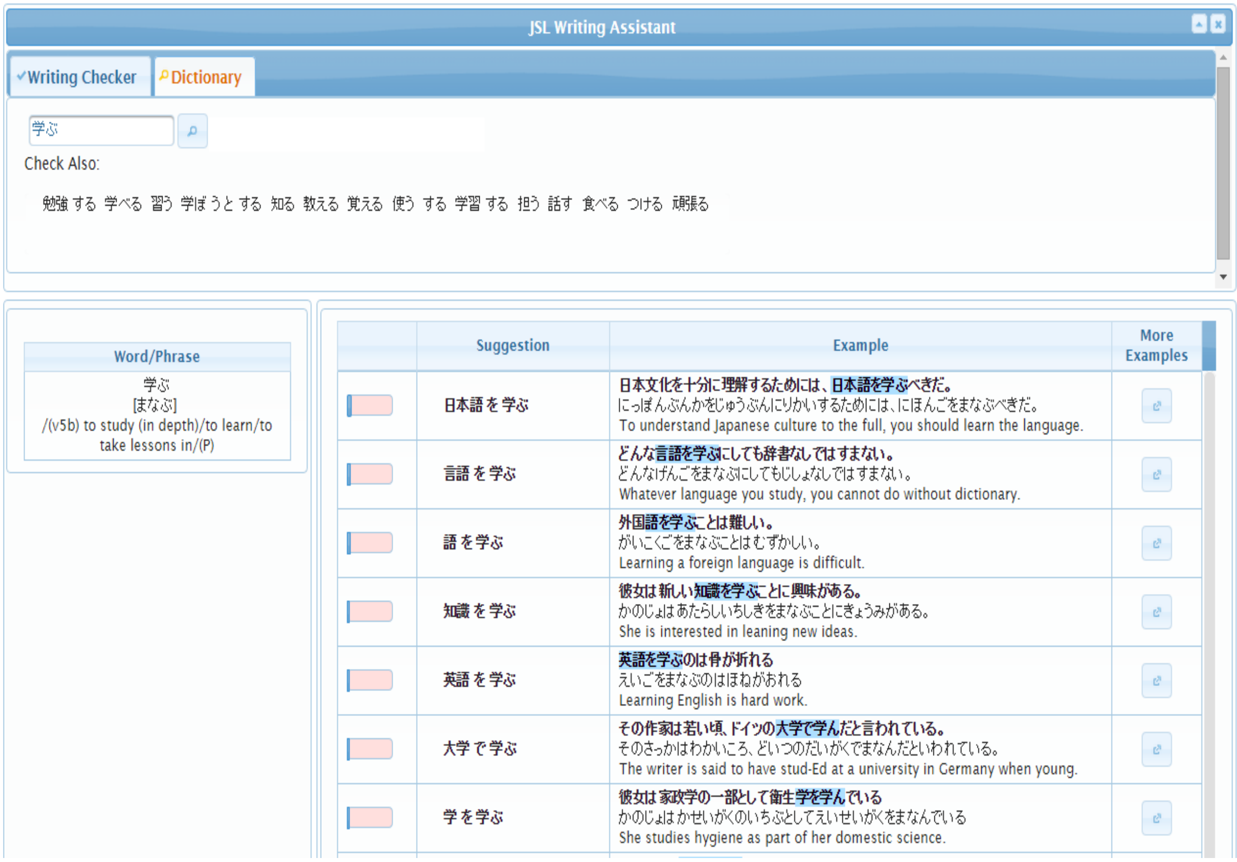
私は薬を食べました。

Your input	All	Choose a word to correct
薬を食べました。	⌕	薬 を 食べる

	Suggestion	Example	More Examples
▭	薬 を 飲む	薬を飲む くすりをのむ take medicine	⧉
▭	薬 を のむ	4時間おきにこの薬をのみなさい。 よんじかんおきにこのくすりをのみなさい。 Take this medicine every four hours.	⧉
▭	薬 を もらう	それで、処方薬をもらいに最初に行った薬局へ立ち寄りました。 それで、しょほうやくをもらいにさいしょにいったやっきょくへたちよりました。 So I dropped back in to the first drugstore to fill my new prescription.	⧉
▭	薬 を 塗る	母は私の膝の傷にその薬を塗ってくれた。 はははわたしのひざのきずにそのくすりをぬってくれた。 Mother applied the medicine to the sore on my knee.	⧉
▭	薬 を 試す	僕は、新しいタイプの薬を試しても良いと判断するまで、少なくとも6ヵ月は待ってみることにしている。 ぼくは、あたらしいたいぷのくすりをためしてもよいとはんだんするまで、すくなくとも6かげつはまってみることにしている。 I plan on waiting at least six months before possibly experimenting with the new class of drugs.	⧉
		私がその薬を買うと薬剤師の顔いっぱいに笑みが広がり、深々とおじぎをしました	

Sentence Examples
「この薬を飲まなければ行けませんか？」 「このくすりをのまなければいけませんか？」 Must I take this medicine ?
母は私に薬を飲ませた。 はははわたしにくすりをのませた。 My mother made me take some medicine.
彼女は子供をほうびでつってその苦い薬を飲ませた。 かのじょはこどもをほうびでつってそのにがいくすりをのませた。 She bribed her child to take the bitter medicine.
食前ごとにこの薬を飲みなさい。 しょくぜんごとにこのくすりをのみなさい。 Take this medicine before each meal.
６時間ごとにこの薬を飲みなさい。 ろくじかんごとにこのくすりをのみなさい。 Take this medicine every six hours.
具合が悪くなったらこの薬を飲みなさい。 ぐあいがわるくなったらこのくすりをのみなさい。 Take this medicine in case you get sick.

Figure 4.4: An example of collocation suggestions produced by the JSL Writing Assistant given the erroneous collocation 薬を食べました (*kusuri wo tabemashita*, lit. 'ate medicine') as input. (a) Collocation suggestions are shown on the left and an example sentence for each suggestion is shown on the right. In the example, 薬を飲む (*kusuri wo nomu*, 'to take medicine') is the correct collocation. (b) Further examples for each suggestion are shown when the user clicks on "More examples". In the example, further example sentences for the collocation 薬を飲む are displayed.



JSL Writing Assistant

学ぶ

Check Also:

勉強 する 学べる 習う 学ぼ うと する 知る 教える 覚える 使う する 学習 する 担う 話す 食べる つける 頑張る

Word/Phrase
学ぶ [まなぶ] /(v5b) to study (in depth)/to learn/to take lessons in/(P)

	Suggestion	Example	More Examples
	日本語 を 学ぶ	日本文化を十分に理解するためには、日本語を学ぶべきだ。 にっぽんぶんかをじゅうぶんにりかいするためには、にほんごをまなぶべきだ。 To understand Japanese culture to the full, you should learn the language.	
	言語 を 学ぶ	どんな言語を学ぶにしても辞書なしではすまない。 どんなげんごをまなぶにしてもじしょなしではすまない。 Whatever language you study, you cannot do without dictionary.	
	語 を 学ぶ	外国語を学ぶことは難しい。 がいこくごをまなぶことはむずかしい。 Learning a foreign language is difficult.	
	知識 を 学ぶ	彼女は新しい知識を学ぶことに興味がある。 かのじょはあたらしいちしきをまなぶことにきょうみがある。 She is interested in leaning new ideas.	
	英語 を 学ぶ	英語を学ぶのは骨が折れる えいごをまなぶのはほねがおれる Learning English is hard work.	
	大学 で 学ぶ	その作家は若い頃、ドイツの大学で学んだと言われている。 そのさっかはわかいころ、どいつのだいがくでまなんだといわれている。 The writer is said to have stud-Ed at a university in Germany when young.	
	学 を 学ぶ	彼女は家政学の一部として衛生学を学んでいる かのじょはかせいがくのいちぶとしてえいせいがくをまなんでいる She studies hygiene as part of her domestic science.	

Figure 4.5: Dictionary function of the JSL Writing Assistant showing collocations for the verb 学ぶ (*manabu*, ‘to learn’). Under “Check Also”, possible confusable words are shown.

Chapter 5

Conclusion

Collocations have been acknowleged as essential to reach native-like fluency. However, there are no explicit rules in learning collocations, which makes it harder for language learners to acquire the correct usage. In this study, we proposed a system that targets potential collocation errors made by JSL learners by a combination of a large learner corpus and statistical association measures. The system can directly provide collocation suggestions to potential collocation errors, aiming to improve JSL learners' word usage.

5.1 Pedagogical Implication

The Writing Assistant as described in this thesis is suitable for JSL learners to check if they commited collocation errors and then apply the suggestions given by the system to revise their text. The system can be used independently or it can be integrated into the writing component of some bigger CALL systems. For example, the system can also be used by teachers as a way to obtain better understanding about learners' errors and help them provide better feedback to the students.

5.2 Other possible applications

5.2.1 Reading Assistant

Some useful tools have been proposed to support learners' reading skills. For Japanese, two examples of existing tools are the Reading Tutor [1] and Rikaichan [2]. The main functionalty of both tools is to provide word by word reading and translation given a Japanese text as input. One limitation is that both tools do not handle collocations and in some cases, the word by word information provided might not be enough to understand the meaning of the whole expression. As an example, we give as input the collocation お茶を入れる (*ocha wo ireru, 'make tea'*). Both tools provide several senses for the component words 手 (*te*, 'hand') and 入れる (*ireru*, 'to put in'), but none of them match to the meaning of 入れる in the expression (Figure 5.1 and Figure 5.2) .

We believe that the integration of our tool to similar reading support systems might be helpful. For example, it can detect target collocations and show bilingual sentence examples that can help learners understand the meaning of the collocation. Moreover, it can show similar expressions to the input text, so the learner can clarify the difference in meaning of possible confusable expressions.

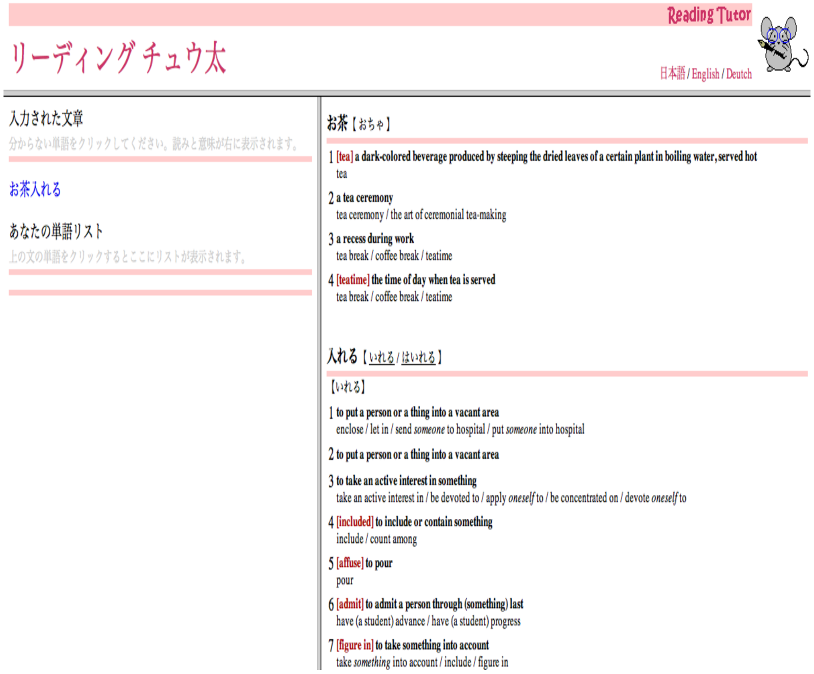
5.2.2 Collocation Assessment

Tests on vocabulary are widely used as a method to evaluate language proficiency. However, elaborating a test from scratch can be time consuming and labor intensive for teachers and instructors. Although several works have been proposed on the automatic generation of vocabulary assessment, there are only few works focusing on the usage of collocations.

We believe that the proposed tool can be helpful in the generation of collocation tests. For example, the tool can be used to identify target collocations within a text and, after blanking out one of the component words, generate the possible distractors. The similar expressions to the collocation can be used as distractors.

[1] http://language.tiu.ac.jp/

[2] http://rikaichan.mozdev.org/



入力された文章

分からない単語をクリックしてください。読みと意味が右に表示されます。

お茶入れる

あなたの単語リスト

上の文の単語をクリックするとここにリストが表示されます。

お茶【おちゃ】

1 [tea] a dark-colored beverage produced by steeping the dried leaves of a certain plant in boiling water, served hot
tea

2 a tea ceremony
tea ceremony / the art of ceremonial tea-making

3 a recess during work
tea break / coffee break / teatime

4 [teatime] the time of day when tea is served
tea break / coffee break / teatime

入れる【いれる / はいれる】

【いれる】

1 to put a person or a thing into a vacant area
enclose / let in / send *someone* to hospital / put *someone* into hospital

2 to put a person or a thing into a vacant area

3 to take an active interest in something
take an active interest in / be devoted to / apply *oneself* to / be concentrated on / devote *oneself* to

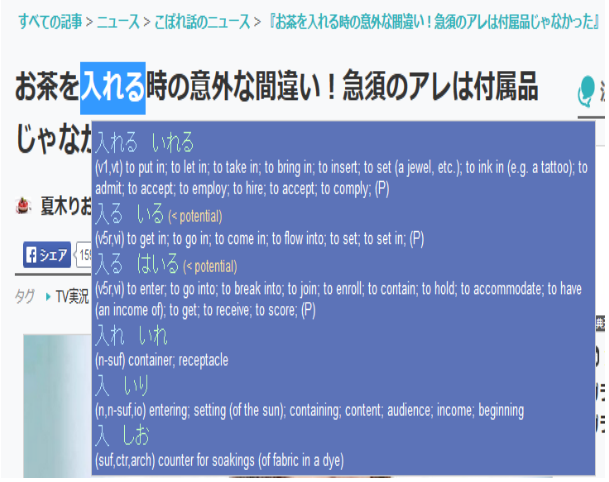
4 [included] to include or contain something
include / count among

5 [affuse] to pour
pour

6 [admit] to admit a person through (something) last
have (a student) advance / have (a student) progress

7 [figure in] to take something into account
take *something* into account / include / figure in

Figure 5.1: The interface of the Reading Tutor sytem.


すべての記事 > ニュース > こぼれ話のニュース > 『お茶を入れる時の意外な間違い！急須のアレは付属品じゃなかった』
シェア
タグ
TV実況

お茶を入れる時の意外な間違い！急須のアレは付属品じゃなか

夏木りお

入れる　いれる
(v1,vt) to put in; to let in; to take in; to bring in; to insert; to set (a jewel, etc.); to ink in (e.g. a tattoo); to admit; to accept; to employ; to hire; to accept; to comply; (P)

入る　いる (< potential)
(v5r,vi) to get in; to go in; to come in; to flow into; to set; to set in; (P)

入る　はいる (< potential)
(v5r,vi) to enter; to go into; to break into; to join; to enroll; to contain; to hold; to accommodate; to have (an income of); to get; to receive; to score; (P)

入れ　いれ
(n-suf) container; receptacle

入　いり
(n,n-suf,io) entering; setting (of the sun); containing; content; audience; income; beginning

入　しお
(suf,ctr,arch) counter for soakings (of fabric in a dye)

Figure 5.2: Example showing the use of Rikaichan for reading a webpage in Japanese.

5.3 Future Work

Education is an exciting area where NLP still has much to contribute. In our particular case, there are undoubtedly many issues that haven't been analyzed or have only been partly addressed. Therefore, many avenues still need to be explored to improve the quality of collocation suggestions. One direction is to verify if other features can increase the performance of our system. For example, one limitation of our experiments is the limited contextual information considered (e.g. for the verb suggestion task, we only considered the noun, the particle and the verb written by the learner). In the future, considering a wider context size might help. Another feature that can be added is information regarding the L1 of the learner. Some studies have shown that L1 is one of the main causes of collocation errors and including such information improves the system performance (see [4] as example).

Other improvement that can be made is to include a module that handle the kanji-kana conversion. Learners sometimes write in hiragana instead of kanji, for example. Kana can sometimes introduce semantic ambiguity when processing the input text. One alternative is given an input word written in kana, prompt all kanji possibilities to the user and ask him/her to choose one of the words. Another alternative is to include all these possible words when generating the confusion set. However, the precision of the system might be affected.

Although recently learner data have become more readily and publicly available, a learner corpus will never be large enough to cover all possible error cases. One possible solution is to further explore how to apply distributed representation models. According to Mikolov et al. [41], the main advantage of these models is that they make generalization to novel patterns easier and model estimation more robust, which can help alleviate the data sparseness problem. Moreover, it has been shown that these models can capture linguistic regularities [40]. However, further investigation is necessary in order to verify if these models can make generalizations taking into consideration the errors commonly made by language learners.

In this work, we only considered binary collocations and another investigation that should be extended is how to consider longer collocations. For example, some noun-verb pairs have incomplete meaning since they are part of complex collocations. One possible solution is to extend the existing association measures to more than two items ([53], for example, extends measures such as PMI and Dice coefficient in order to deal with candidates of arbitrary size) or to propose new measures. Another approach, proposed by Seretan [58], is to identify long collocations by relying on previously extracted binary collocations. She extends the notion of collocation from co-occurrence of words to co-occurrence of collocations. In her method, co-occurrence of two collocations means that they combine syntactically by sharing a common term in the input sentence (e.g. *natural language* and *language processing* in the collocation *natural language processing*).

Another interesting, yet challenging direction is how to provide useful feedback to the learner. As observed by Heatst [18], a challenge for NLP is how to build *writing coaches*, a system that watches alongside a learner as they write an essay, giving hints and scaffolding the way a tutor would - not giving the answer explicitly, but showing the path and letting the learner fill in the missing information.

Finally, a more extensive evaluation with JSL learners in practical learning scenarios (e.g. in the real classroom) is necessary to further verify the usefulness of the applications proposed.

In summary, we have presented a method to correct potential collocation errors made by JSL learners by a combination of a large learner corpus and statistical association measures. The experimental results showed that this method is feasable for suggesting collocations with good ranking quality compared to existing methods. In addition, we showed how to try to utilize our method to develop useful end-user applications for pedagogical application.

References

[1] Bai, M. H., You, J. M., Chen, K. J., & Chang, J. S. (2009, August). Acquiring translation equivalences of multiword expressions by normalized correlation frequencies. In Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2 (pp. 478-486). Association for Computational Linguistics.

[2] Bengio, Y. 2008. Neural net language models. Retrieved from: http://www.scholarpedia.org/article/Neural_net_language_models

[3] Breen, J. (1995). Building an electronic japanese-english dictionary. In *Japanese Studies Association of Australia Conference*.

[4] 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. doi: 10.1080/09588220802090337.

[5] 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.

[6] Cho, Y.S. (2013). Software Review: Lang-8. *CALICO Journal*, 30(2), pp 293-299. doi: 10.11139/cj.30.2.293-299.

[7] 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, USA. doi:10.3115/981623.981633

[8] Cowie A. P., 1978. The place of illustrative material and collocations in the design of a learner's dictionary. In *P. Strevens, editor, In Honour of A. S. Hornby*, pages 127—139. Oxford University Press.

[9] Curran, J. R.(2003). From distributional to semantic similarity. (Doctoral dissertation). Retrieved from http://www.inf.ed.ac.uk/publications/thesis/online/IP030023.pdf

[10] 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). Association for Computational Linguistics, Stroudsburg, PA, USA.

[11] Deane, P. (2005).A nonparametric method for extraction of candidate phrasal terms. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics* (pp. 605-613). Association for Computational Linguistics.

[12] Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational linguistics*, 19(1), 61–74. MIT Press Cambridge, MA, USA.

[13] Evert, S. (2008). Corpora and collocations. *Corpus Linguistics. An International Handbook*, 2, 223-233.

[14] Evert, S. (2008). A lexicographic evaluation of German adjective-noun collocations. *Towards a Shared Task for Multiword Expressions (MWE 2008)*, 3.

[15] 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. doi: 10.1080/09588220802343561.

[16] Gamon, M. (2010). Using mostly native data to correct errors in learners' writing: a meta-classifier approach. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics* (pp. 163-171). Association for Computational Linguistics, Stroudsburg, PA, USA.

[17] Harris. Z. (1954). Distributional structure. *Word*,10 (23), 146-162.

[18] Hearst, M. A. (2015) Can Natural Language Processing Become Natural Language Coaching?. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, (pp. 1245-1252). Association for Computational Linguistics, Stroudsburg, PA, USA.

[19] Hill, J. 2000. Revising priorities: From grammatical failure to collocational success. In *Teaching Collocation: Further Developments in the Lexical Approach*, ed. Michael Lewis, 88-117. Hove: Language Teaching Publications.

[20] Hodoscek, B. (2013). Contextually aware writing assistance system for Japanese. PhD Thesis. Tokyo Institute of Technology, Japan.

[21] Japan Foundation: survey report on Japanese language education abroad. 2013. Retrieved from https://www.jpf.go.jp/j/project/japanese/survey/result/dl/survey_2012/2012_s_excerpt_e.pdf

[22] Johns, T. (1991). Should you be persuaded—two samples of data-driven learning materials. In T. Johns & P. King (Eds.), *Classroom Concordancing* (Vol. 4). Birmingham: Center for English Language Studies.

[23] 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.

[24] Justeson, J. S., & Katz, S. M. (1995). Technical terminology: some linguistic properties and an algorithm for identification in text. *Natural language engineering*, 1(01), 9-27.

[25] Kathleen R. McKeown and Dragomir R. Radev. Collocations. In Robert Dale, Hermann Moisl, and Harold Somers, editors, *A Handbook of Natural Language Processing*. Marcel Dekker, 2000.

[26] Kilgarriff, A., Marcowitz, F., Smith, S., & Thomas, J. (2015). Corpora and Language Learning with the Sketch Engine and SKELL. Revue française de linguistique appliquée, 20(1), 61-80.

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

[28] Kudo, T. & Matsumoto, Y. (2002). Japanese dependency analysis using cascaded chunking. In *Proceedings of the 6th Conference on Natural Language Learning*, (pp.1-7). Association for Computational Linguistics, Stroudsburg, PA, USA, 1-7. doi:10.3115/1118853.1118869.

[29] Kullback, S. & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics* 22(1), 79–86.

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

[31] 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 Publishers.

[32] 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). Association for Computational Linguistics, Stroudsburg, PA, USA. doi:10.3115/1034678.1034693.

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

[34] Li, D., & Yu, D. (2014). *Deep Learning: Methods and Applications.* Foundations and Trends in Signal Processing, Now Publishers.

[35] Liou, H., Chang, J., Chen, H., Lin, C., Liaw, M., Gao, Z., Jang, J., Yeh, Y, Chuang, T. and You, G. (2006) Corpora processing and computational scaffolding for a Web-based English learning environment: The CANDLE Project. *CALICO Journal*, 24 (1), 77-95.

[36] 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). Association for Computational Linguistics, Stroudsburg, PA, USA.

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

[38] Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., Den, Y. (2014). Balanced Corpus of Contemporary Written Japanese.*Language Resources and Evaluation*, 48 (2), 345-371.

[39] Manning, C. D., & Schütze, H. (1999). *Foundations of statistical natural language processing.* MIT press.

[40] Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic Regularities in Continuous Space Word Representations. In HLT-NAACL (pp. 746-751).

[41] Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.

[42] Mitchell, T. M. (1997). Machine learning. WCB.

[43] Mizumoto, T., Komachi, M., Nagata, M., & Matsumoto, Y. (2011. Mining Revision Log of Language Learning SNS for Automated Japanese Error Correction of Second Language Learners. In *Proceedings of the 5th International Joint Conference on Natural Language Processing* (pp. 147-155).

[44] Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24, 223–242.

[45] Östling, R. and Knutsson, O. 2009. A corpus-based tool for helping writers with Swedish collocations , In Proceedings of the Workshop on Extracting and Using Constructions in NLP, Nodalida, Odense, Denmark.

[46] Oyama, H., & Matsumoto, Y. (2010). Automatic error detection method for japanese particles. Ritsumeikan Asia Pacific University Polyglossia Vol, 18, 55-63.

[47] 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.

[48] Ozaki, S. (2011). Teaching collocations effectively with the aid of L1. *The Language Teacher*, 35(3), 37-40.

[49] 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.

[50] Pecina, P. (2005). An extensive empirical study of collocation extraction methods. In *Proceedings of the ACL Student Research Workshop* (pp. 13-18). Association for Computational Linguistics.

[51] Pecina, P. (2010). Lexical association measures and collocation extraction. Language resources and evaluation, 44(1-2), 137-158.

[52] Ramisch, C. (2012, July). A generic framework for multiword expressions treatment: from acquisition to applications. In *Proceedings of ACL 2012 Student Research Workshop* (pp. 61-66). Association for Computational Linguistics.

[53] Ramisch, C., Villavicencio, A., & Boitet, C. (2010, May). mwetoolkit: a Framework for Multiword Expression Identification. In LREC.

[54] Ramisch, C. (2012). A generic and open framework for multiword expressions treatment: from acquisition to applications (Doctoral dissertation, Université de Grenoble (France); Universidade Federal do Rio Grande do Sul (Brazil)).

[55] Rozovskaya, A. & Roth, D. (2010). Generating confusion sets for context-sensitive error correction. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing* (pp. 961-970). Association for Computational Linguistics, Stroudsburg, PA, USA.

[56] Sahlgren, M. (2005). An introduction to random indexing. In *Methods and applications of semantic indexing workshop at the 7th international conference on terminology and knowledge engineering*, TKE (Vol. 5).

[57] Seretan, V. (2008). Collocation Extraction Based on Syntactic Parsing. PhD Thesis. University of Geneva, Switzerland.

[58] Seretan, V. (2011). *Syntax-based collocation extraction.* Text, speech and language technology series, 44. Springer-Verlag New York, Inc., New York, NY, USA.

[59] Shei, C.C., & Pain, H. (2000). An ESL writer's collocational aid. *Computer Assisted Language Learning*, 13(2), 167–182.

[60] Shermis, M. D., & Burstein, J. (2013). *Handbook of automated essay evaluation: Current applications and new directions.* Routledge.

[61] Shimbun, M. (1991). 1994. *Mainichi shimbun CD-ROM.*

[62] Shoji, K. (2010) Common Japanese Collocations: A Learner's Guide To Frequent Word Pairings. Kodansha International Ltd.

[63] Smadja, F. (1993). Retrieving collocations from text: Xtract. *Computational linguistics*, 19(1), 143–177. MIT Press Cambridge, MA, USA.

[64] Smadja, F., McKeown, K. R. & Hatzivassiloglou, V. (1996). Translating collocations for bilingual lexicons: A statistical approach. *Computational linguistics*, 22(1), 1–38. MIT Press Cambridge, MA, USA.

[65] Suzuki, H., & Toutanova, K. (2006). Learning to predict case markers in japanese. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics* (pp. 1049-1056). Association for Computational Linguistics.

[66] Tetreault, J., Foster, J. & Chodorow, M. (2010). Using parse features for preposition selection and error detection. In *Proceedings of the 48th annual meeting of the Association for Computational Linguistics* (pp. 353–358). Stroudsburg, PA, USA.

[67] Turian, J., Ratinov, L., & Bengio, Y. (2010). Word representations: a simple and general method for semi-supervised learning. In *Proceedings of the 48th annual meeting of the association for computational linguistics* (pp. 384-394). Association for Computational Linguistics.

[68] Varghese, A., Varde, A. S., Peng, J., & Fitzpatrick, E. (2015). A Framework for Collocation Error Correction in Web Pages and Text Documents. *ACM SIGKDD Explorations Newsletter*, 17(1), 14-23.

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

[70] Wible, D., Kuo, C., Tsao, N., Liu, A. & Lin, H. (2003). Bootstrapping in a Language Learning Environment, *Journal of Computer-Assisted Learning*, 19 (1), pp. 90-102. SSCI, LLBA.

[71] 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). Association for Computational Linguistics, Stroudsburg, PA, USA.

Appendix

A Comparison of Different Distributional Similarity Measures

Table 5.1 shows the results of the evaluation of several distributional similarity measures (Cosine, Dice measure, Jensen-Sahnnon Divergence and Kullback-Leibler divergence). The Jensen-Shannon divergence obtained the highest MRR values.

	Verb Suggestion				Noun Suggestion			
System	*P@1*	*P@5*	*Recall*	*MRR*	*P@1*	*P@5*	*Recall*	*MRR*
Cosine + WD	0.46	0.79	0.58	0.35	0.53	0.78	0.35	0.23
Dice + WD	0.51	0.76	0.51	0.46	0.37	0.74	0.34	0.17
KL-Div + WD	0.55	0.75	0.43	0.28	0.58	0.79	0.24	0.16
JS-Div + WD	0.54	0.80	0.73	0.49	0.38	0.67	0.47	0.23

Table 5.1: The precision, recall and MRR of different distributional similarity measures and Weighted Dice measure combinations applied to object-verb constructions.

Note: WD stands for Weighted Dice, KL-Div stands for Kullback-Leibler divergence and JS-Div stands for Jensen-Shannon Divergence.

List of Publications

Journal Paper

- Lis Pereira and Yuji Matsumoto. Leveraging a Learner Corpus for Automated Collocation Suggestion for Learners of Japanese as a Second Language. CALICO Journal. (Accepted, to be published) [**Corresponds to Chapter 3 and Chapter 4**]

International Conferences (refereed)

- Lis Pereira and Yuji Matsumoto. 2015. Collocational Aid for Learners of Japanese as a Second Language. In Proceedings of the 2nd Workshop on Natural Language Processing Techniques for Educational Applications (NLP-TEA-2), Beijing, China. [**Corresponds to Chapter 4**]

- Lis W.K. Pereira, Elga Strafella, Kevin Duh and Yuji Matsumoto. 2014. Identifying Collocations using Cross-lingual Association Measures. In Proceedings of the EACL 2014 Workshop on Multiword Expressions, Gothenburg, Sweden.

- Lis W.K. Pereira, Elga Strafella and Yuji Matsumoto. 2014. Collocation or Free Combination? – Applying Machine Translation Techniques to identify collocations in Japanese. Proceedings of the 9th Language Resources and Evaluation Conference (LREC 2014), Reykjavik, Iceland.

- Lis W.K. Pereira, Erlyn Manguilimotan and Yuji Matsumoto. 2013. Automated Collocation Suggestion for Japanese Second Language Learners.

In Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL), Student Research Workshop, pp.52-58, Sofia, Bulgaria. [**Corresponds to Chapter 3**]

Non-refereed Publications

- Lis W.K. Pereira, Erlyn Manguilimotan and Yuji Matsumoto. 2013. Data Coverage vs. Data Size: A comparison of two large-scale corpora in Collocation Suggestion for Japanese Second Language Learners. In Proceedings of the Nineteenth Annual Meeting of the Association for Natural Language Processing (NLP-2013), pp.74-76, Nagoya, Japan, March 2013. [**Corresponds to Chapter 3**]

- Lis W.K.Pereira, Erlyn Manguilimotan and Yuji Matsumoto. 2013. Collocation Suggestion for Japanese Second Language Learners, In Proceedings of the 210th Meeting of the Information Processing Society of Japan, Special Interest Group on Natural Language Processing, Vol.2013-NL-21, No.3, pp.1-5, January 2013.

Acknowledgements

I am deeply grateful to Matsumoto Sensei for accepting me as student in his lab and for providing continuous help and encouragement through all the PhD course.

I also thank the commitee members of this thesis for the helpful comments and advice.

I am grateful to Ms Yuko Kitagawa and all the staff of NAIST International Affairs Division, for all the help and support during these years in Japan.

I heartily thank Marcos Yokoyama, Marina Oikawa, Olivia Nakaema, João Monzani, Gustavo Garcia, Diego Reinoso, Xiaodong Liu, Budi Irmawati, Elga Strafella and Erika Carmargo. Thanks for everything!

I specially thank Diego Reinoso and Xiaodong Liu for helping me with the journal and thesis drafts.

I am also thankful to my family and friends in Brazil. Special thanks to Eder Cruz and Marilúcia Bezerra for the help and friendship.

Finally, I would like to thank the Japanese Government for providing financial support for this PhD course.