CORPUS-BASED COLLOCATION RESEARCH TARGETED AT JAPANESE LANGUAGE LEARNERS

Irena SRDANOVIC
University of Ljubljana
irenasrdanovic@gmail.com

Abstract
This paper discusses corpus-based research on collocations, introduces various tools for querying and extracting Japanese collocations and presents an analysis of Japanese collocations using language corpora and related tools. First, major corpus query tools such as Sketch Engine, NINJAL-NLP, Natsume, Chunagon, which can be used by learners and teachers of Japanese language, are briefly described. Focus then shifts to adjectival and nominal collocates and the resource "Collocation data of adjectives and nouns" which consists of adjective headwords and their nominal collocates extracted from two large corpora, BCCWJ and JpTenTen: 500 adjectives and 9,218 collocate nouns, and 500 adjectives and 23,220 collocate nouns from each corpus respectively. Finally, it is shown that corpus-based resources can be used in the creation of reference materials for learners of the Japanese language. The benefits of empirical research into collocations are also shown by comparing the obtained results with collocations in textbooks for Japanese as foreign language.

Keywords: Japanese language; collocations; second-language acquisition; language learning; corpora and tools

Povzetek

Ključne besede: japonski jezik; usvajanje drugega jezika; učenje jezika; kolokacije; korpusi in orodja

ISSN: 2232-3317, http://revije.ff.uni-lj.si/ala/
DOI: 10.4312/ala.4.2.25-36
1. Introduction

The importance of collocation research has been recognized especially within corpus linguistics with the development of empirical research methods. There are also an increasing number of Japanese language resources and studies on collocations, for example about which collocations are easily acquired and which are burdensome for foreign language learners or about the need to systematically teach collocations to learners and provide reference materials on collocations to ease the process of learning (Oso and Takizawa 2003, Himeno 2012, Srdanović 2013a).

In this paper, I first briefly describe various tools for searching and extracting Japanese collocations in corpora (Chakoshi, Case frame, Sketch Engine, Natsume, NINJAL-NLP, Chunagon). Next, I focus on collocations of the type adjective+noun and present the resource "Collocation data of adjectives and nouns" which consists of adjectival headwords and their noun collocates extracted from two large corpora, BCCWJ and JpTenTen: 500 adjectives and 9,218 collocate nouns, and 500 adjectives and 23,220 collocate nouns from each corpus respectively.

I then will discuss the types of information that can be provided in a hypothetical dictionary of collocations for Japanese language learners, such as placing emphasis on collocations that are difficult for language learners to predict, displaying lexical-map information, corpus-informed information on register and special usages, and so on. Finally, the comparison of these research results with collocations in textbooks for Japanese as foreign language indicates the importance of empirical research into and systematic treatment of collocations, which can then be applied in the creation of corpus-based Japanese language learning materials.

2. Corpus query tools for collocation analyses

Kawahara and Kurohashi (2006) developed the Case Frame \(^1\) online search functionality that initially only extracted verbs and their case frames from the corpus but was later on expanded. Currently it provides a valuable predicate-argument profile of target words from a large web corpus, although the corpus data is limited to extracted sentences only. The first corpus query tool specifically created for Japanese collocation extraction, Chakoshi,\(^2\) was developed at Nagoya University, and is a Japanese text search and collocation extraction system (Fukuda 2007). However, the public version of this tool is limited to searches of the Aozora bunko corpus of literary texts and the Nagoya kaiwa corpus.

---

\(^1\) Case Frame: http://lotus.kuee.kyoto-u.ac.jp/cf-search/

\(^2\) Chakoshi: http://tell.cla.purdue.edu/chakoshi/public.html
In 2007, the first corpus-query system with detailed lexical profiles of search words for Japanese appeared (Srdanović et al. 2008), Sketch Engine. Sketch Engine was originally created for English (Kilgarriff et al. 2004), but then more than a hundred corpora and sketch grammars for other languages were gradually added to the system. As described by McEnery and Hardie (2012), the tool belongs to the fourth generation of corpus-query tools with various advanced search functionalities. The major functionality breakthrough was called word sketches, “one-page automatic, corpus-based summaries of a word's grammatical and collocational behaviour” (Kilgarriff et al. 2004).

At this stage, Sketch Engine can be searched for automated summaries of Japanese collocational relations within the 400-million word JpWaC (Srdanović et al. 2008) and the two versions of JpTenTen (Srdanović et al. 2013) - the full one, a 10-billion token corpus with short-unit-word annotation, and the smaller sample one, a 200-million token corpus with long-unit-word annotation (suw and luw). This tool enables inclusion and creation of other corpora for personal or group use, so the author got permission to use BCCWJ with Sketch Engine as well. Other functionalities include SketchDiff - comparison of the collocational behavior of two similar words, thesaurus, corpora comparison, parallel corpora concordancing, and lexicographic tickbox (Kilgarriff et al. 2004).

The Japanese word sketches inspired researchers in Japan to create online tools that automatically summarize collocations. Unlike the multilanguage-oriented Sketch Engine, these tools are for Japanese only, but have other specific functionalities and are available for free. These tools, Natsume and NINJAL-LWP, are described below.

Natsume is a computer-assisted language learning system that supports learners’ Japanese writing skills by providing assistance in automatically summarized collocation relations drawn from multiple Japanese (sub-) corpora of different genres (Nishina 2011). An advantage of this is that comparison of the collocational behavior of words in different genres is provided. Further, the tool is convenient for searching synonyms and thereby exploring their similarities and differences through collocates. Recently, the tool has been provided in a set with the learner corpus Natane and Nutmeg, a system for automatic correction of learners’ errors (Hodošček and Nishina 2012).

In 2010, the online corpus-query system NINJAL-LWP was developed, initially for BCCWJ (Pardeshi and Akasegawa 2010), later incorporating the Tsukuba web corpus (TWC) (Imai et al. 2012). Similarly to Sketch Engine and Natsume, the tool provides lexical profiling, with a comprehensive picture of the collocational and

---

3 Sketch Engine: http://www.sketchengine.co.uk/. The system is owned by Lexical Computing Ltd. One-month free trial is provided.

4 Later on, another term appears for such types of tools, “lexical profiler”. The name aims to grasp their advantages in search functionalities over standard concordance tools.

5 Natsume: http://hinoki.ryu.titech.ac.jp/natsume/

6 NINJAL-LWP: http://nlb.ninjal.ac.jp/
grammatical patterns of lexical entries. The system also visually compares the frequencies of search words in each of the subcorpora of BCCWJ, as well as the distribution of words within dialogues or within prose inside the book subcorpus.

With the compilation of the Balanced Corpus of Contemporary Written Japanese (BCCWJ), a new corpus-query tool Chunagon⁷ was developed (Maekawa et al. 2013). This tool offers three major search functionalities for searching the morphologically annotated BCCWJ, namely the short-unit word, long-unit word and string search functionalities. By specifying a keyword and its surrounding context, the tool enables extraction of collocational relations as well. The extracted data can be downloaded in tabular format and then further analyzed within Excel or similar tools.

The corpus query tools introduced above have provided an essential contribution to the study of Japanese corpus linguistics and also have great potential for application in the fields of Japanese language education and second language acquisition.

3. Collocation data and possible applications in language education

This section focuses on adjectival and nominal collocational data and presents the results of the extraction of such collocations from Japanese language corpora. Based on the analysis of the obtained data, possible applications to Japanese language education are discussed, by comparing the results with the collocation data currently present in Japanese language textbooks.

4. "Collocation data of adjectives and nouns"

This collocation data resource was created for most frequent 500 adjectives in both, BCCWJ and JpTenTen. Once the most frequent adjectives were detected, for each of the adjectives their most frequent noun collocates were extracted using the Sketch Engine tool.⁸ For highly frequent adjectives, up to 100 nominal collocates and for the rest of the adjectives up to 50 nominal collocates were extracted. In order to gain clean data, collocates below a frequency of 5 tokens in JpTenTen and a frequency of 2 tokens in BCCWJ were not included.

Both annotation data types, short-unit and long-unit words, were considered in this task, but since the JpTenTen with long-unit-word annotation is smaller in size and therefore fewer collocations can be obtained from it, long-unit-word annotated data was used only for the most frequent compound adjectives that were not obtainable from the short-unit-word annotated data. For example, kyōnibukai ‘interesting’ is annotated as

---

⁷ Chunagon: https://chunagon.ninjal.ac.jp/
⁸ Many thanks to the Sketch Engine team for technical support on this task.
one long-word-unit and two short-word-units: kyōmi ‘interest’ + fukai ‘deep’ and was therefore extracted from the corpus annotated with long-unit-words.⁹

One of the main drawbacks of obtaining collocation data from short-unit-word annotated data is that some units that are in general considered as a word are divided into two or more than two units reflecting their morphological derivation. For example, the word kenkyūsha is annotated as two short-word-units: kenkyū ‘research’ + sha ‘[a morpheme denoting a person]’, which can result in obtaining kenkyū when the actual use of the noun collocate was kenkyūsha. To overcome this issue, the three most frequent one-, two- and three-grams of word-units on the right and left of a collocate (right and left contexts of a collocate) were extracted and included in the final data, as shown in Table 1.

Table 1 exemplifies the obtained results of adjectives and their noun collocates along with their contexts. For the sake of space, only a few nominal collocates of a few adjectives are displayed.

Finally, the obtained results based on large-scale resources and empirical methods clearly revealed the adjectives that have lexical constraints in their attributive role. For example tebayai “quick”, tokorosemai “crowded” are not used in their attributive adjectival form, preceding modified nouns, but rather in other forms, such as as an adverb preceding a verb (tebayaku katazukeru “to tidy up quickly”, tokorosemashi to narande iru “to line up crowdedly”).

5. Towards a dictionary of collocations for Japanese language learners

The resource “Collocation data of adjectives and nouns,” extracted from two large-scale corpora, can serve as a good basis for the creation of various resources that aim to describe Japanese i-adjectives and their behavior as modifiers of nouns. This section presents a number of major steps in analysis of the obtained collocation data, aimed at the creation of a dictionary of collocations for Japanese language learners. The analysis consisted of the following steps:

- Defining main entries for the dictionary

Srdanović (2013b) showed that the 25 most frequent i-adjectives represent more than 62% of the overall usage of i-adjectives in Japanese texts. Since our learner dictionary aims to cover very detailed information on adjective+noun collocates with thorough supplemental data, the initial version of the dictionary aimed only to cover these most frequent i-adjectives.

---

⁹ For research on adjectives annotated as short- and long-word-units, refer to Srdanović (2013b).
<table>
<thead>
<tr>
<th>Adjective</th>
<th>Noun 1</th>
<th>Noun 2</th>
<th>Noun 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Example 1</td>
<td>Example 2</td>
<td>Example 3</td>
<td>Example 4</td>
</tr>
</tbody>
</table>

**Table 1:** Example from the resource "Collection data of adjective and nouns"
Comparing collocations extracted from the two large-scale corpora

While we can use collocations from only one reliable large-scale corpus, there are some advantages in comparing the two large-scale corpora. The comparison reveals the idiosyncrasies of each corpus and helps us to exclude irrelevant collocates. On the other hand, results that are derived from both the corpora can be confirmed as significant. For example, the comparison of the noun collocates of the adjective *takai* ‘high’ in the two large-scale corpora of contemporary Japanese, BCCWJ and JpTenTen,\(^\text{10}\) showed a very high level of similarity, which indicates the reliability of the corpora and the significance of the data (Srdanović 2013a: 149-51). A few differences among the first 50 collocates showed some more specific usages, such as *takai nobi/takai nobiritsu* ‘high growth’, which are often used in economics and specifically appear in the White book subcorpus of BCCWJ.

- Arranging the list of collocates

The list of noun collocates were arranged so that functional words such as *koto, mono, hazu, wake, tame* etc. are excluded. Then, the right and left contexts of collocates were checked to avoid the inclusion of incomplete collocates, and also in order to consider morphological elements that were excluded from the annotated short-unit word, for example, *(sei ga )takai kata* ‘a tall person’, *takai gijutsu(ryoku)* ‘high skills/technology’.

- Grouping collocates based on their difficulty level

The twenty most frequent adjectives appear in the beginners’ level of the old JLPT list and the Tsukuba vocabulary list, but they naturally appear in combination with their noun collocates at varying levels of Japanese language study. Some of the collocates do not even appear in the JLPT list or the super advanced level of Tsukuba vocabulary list.

Grouping collocates into difficulty levels assists teachers and learners when considering collocates for inclusion at a specific level of language learning. Lower level collocates are not necessarily for inclusion at lower levels since in some cases a collocation requires a high level of proficiency even though its constituents are basic words. Learners’ needs, motivation and other relevant factors also need to be considered when creating a foreign language learning syllabus. For example, while *takai* ‘high/tall/expensive’ and *kaimon* ‘shopping’ are both basic level words, their patterns of collocation are rather complex, abstract, and too specific to be included at the beginner level of a general Japanese language program. In contrast, higher level collocates would never be used in lower levels since their constituents are by definition difficult. This relationship between the level of difficulty of a collocate versus the difficulty of its constituents needs further clarification and calls for further collocation data analysis.

- Providing translations to a target language and furigana

---
\(^{10}\) The comparison is done thus; the relative frequency of the first 100 nominal collocates for the adjective *takai* is calculated for each of the two corpora. The collocates are then sorted by the highest score of the sum of their relative frequencies in both corpora. Then, if a collocate appears in the list of the first 50 collocates in one corpus, but does not appear in the other corpus’ list of the first 100 nominal collocates, it is specifically marked as different and further checked.
This study is limited to translating the obtained collocations into English, but can serve as a basis for preparing translations in other languages. \(^{11}\) Besides English translations, the dictionary provides furigana in order to assist learners in reading unknown Japanese words.

- Detecting collocations that are difficult to predict or unpredictable for learners

Nation (2001) describes the so-called »unpredictable collocations« - collocations that are difficult for language learners to predict based on knowledge of their native or other foreign languages. Because a learner’s prior knowledge is based on the constituents of collocations in their native language, if these constituents are different from those in semantically similar expressions in Japanese, the learner is likely to make a mistake and produce an utterance that is unnatural in Japanese. Therefore, it is important to grasp these types of collocations and pay special attention when introducing them to language learners by taking into account also how the collocates appear in their native language. The analysis of the collocates of takai (Srdanović 2013a) revealed a few different types of unpredictable collocates for learners of Japanese who are English native speakers. For example, sei ga takai hito/kata ‘a high person’ is unpredictable since the additional elements sei ga ‘the back is’ must be added to make the collocation complete in Japanese (lit. a person with a high back). Although the collocation takai kabe has the same constituent ‘high’ (takai) and ‘wall’ (kabe) as in English, the same expression has a second more abstract meaning ‘high barrier’ with different and unpredictable constituents. These unpredictable usages, different from expressions in learners’ mother tongues, need to be explicitly explained. Finally, it is important to note that unpredictability is relative to the language background of learners and needs to be considered separately for each language.

- Discovering complex patterns, typical usages, genre specifics etc.

Further, the corpus-based analysis reveals complex patterns, typical usages, and usages specific to a particular sub-corpus/genre, which are then described in the learners’ dictionary. This type of information is specifically searched for in the case of unpredictable collocations. The following is an example of a detailed description on differences in usage between takai koe ‘high-pitched voice’ and ōkī koe ‘loud voice’.

<table>
<thead>
<tr>
<th>高い声</th>
<th>「高い声」と「大きい声」は、意味・用法が違う。</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>「高い声」+で+「歌う、鳴く、叫ぶ」（with a high-pitched voice）</td>
</tr>
<tr>
<td></td>
<td>「大きい声」+で+「言う、話す、歌う、叫ぶ、挨拶する」（with a loud voice, loudly）</td>
</tr>
<tr>
<td></td>
<td>● カラオケで高い声 がでなくなって辞めた。</td>
</tr>
</tbody>
</table>

\(^{11}\) This could possibly be a basis for theoretical investigations into the cross-linguistic behavior of collocations.
以前出ていた高い声が出なくなったらどう思いますか？

- 大きい声で言ってください。（注：このよく使われる表現は「高い」で使えない。） |

大きい声では言えないような話も多い。

- Providing lexical maps of noun collocates

Finally, the dictionary provides images of lexical maps of noun collocates, which depict how collocates can be grouped together as concepts observed from a cognitive linguistics perspective. The analysis of takai and usage of its noun collocates revealed three major groups of collocates: collocates that represent relative position (e.g. takai yama ‘high mountain’, takai ki ‘high tree’, takai tokoro ‘high place’), relative quantity (takai wariai ‘high percentage’, takai nedan ‘high price’) and relative quality (takai shinraisei ‘high reliability’, kanshin ‘high/great interest’).

For more information on how entries are structured in the dictionary, refer to Srdanović (2013a). This study introduces the adjective takai and its noun collocates as an example from the ongoing dictionary project.

6. Evaluation of Japanese language textbooks for their collocation data

This section summarizes the results of the analysis of the adjective takai and the use of its nominal collocates in the beginner Japanese language textbooks Minna no Nihongo honbun 1 and 2. The results are compared to the collocation data obtained from the corpora and processed for the target learners’ collocation dictionary described in the previous section. The observations are as follows:

- The adjective takai is introduced in the beginner textbooks in various grammatical forms and patterns including the attributive form (rentai). This adjective is presented in its predicative form (shūshi) more than twice as often as in its attributive form. The behavior of the adjective in the corpus indicates, however, that these two forms of takai have approximately the same distribution.12
- Takai ‘high, tall’ in its attributive form is introduced with nouns such as yama ‘mountain’, biru ‘building’ and in the pattern sei ga takai hito ‘a tall person’, which only covers only the semantic domain of relative position. The use of corpora revealed two other major domains (relative quantity and relative quality) covered by the collocates of takai in its attributive form.
- Among the collocates not covered in the textbooks are: tokoro ‘place’, tatemono ‘building’, ki ‘tree’, kaimono ‘shopping’, (o)kane ‘money’ which belong to JLPT

12 I presented the analysis of the distribution of forms of the adjective takai from JpTenTen at the 27th Paris Meeting on East Asian Linguistics, at CRLAO / INALCO. They are summarized as: takai+N (18%), takai+suffix (18%), N+takai (10%), Nga/no+takai+N (16%), takaku+V (9%), Nga+takaku(te)[cont] (8%), Nga+takai[concl] (18%), Nwa+takai[concl] (3%).
4th level and correspond to a low beginner level. Also, *oto* ‘sound’, *kabe* ‘wall/barrier’, *basho* ‘location’, *wariai* ‘percentage’, *gijutsu*(ryoku) ‘technology/skill’, *kyōiku* ‘education’ belong to JLPT 3rd level and correspond to an upper beginner level.

Some of the above mentioned collocates are less predictable by Japanese language learners and therefore require special treatment. For example, *sei ga takai hito* ‘a tall person’ is not correct if used only with the constituents *takai hito* ‘lit. tall + person’]. Learners make such mistakes, as noted by Srdanović and Sakoda (2013) and therefore Japanese language learning materials need to make learners aware of this.

To summarize, there is a need for more systematic treatment of collocations in textbooks, in relation to forms, relevant semantic domains, levels of difficulty and predictability. Also, a well-balanced large-scale spoken corpus, once created for Japanese, would be a valuable resource for further analysis.

7. Conclusion

This paper introduced some major tools, resources and methods that can be used in analysis of collocations. “Collocation data of adjective and nouns”, which was obtained from two large-scale Japanese language corpora has been described. The resource is valuable as an exhaustive collection of the collocational data of i-adjectives and nouns and can be used as a basis for the creation of various empirically based materials that would aim to describe Japanese i-adjectives when they act as nominal modifiers. Also discussed was a more in-depth analysis of “Collocation data of adjective and nouns” aimed at the creation of a dictionary of collocations for Japanese language learners. Finally, the comparison of the data with the collocation information in textbooks revealed how the material could be improved based on the data obtained from corpora.

The Japanese language corpora and tools, collocation resources, research methods and results that were introduced in this paper will hopefully contribute to future creation of corpus-based Japanese learners’ dictionaries, textbook materials and syllabi.

**Literature**


---

13 Some of the items, although belong to beginner level, are more demanding to master as collocations and hence not needed at this level, for example *takai kaimono* ‘expensive purchase’, *takai okane* ‘lots of money’.


CONSTRUCTION OF A LEARNER CORPUS FOR JAPANESE LANGUAGE LEARNERS: NATANE AND NUTMEG

Kikuko NISHINA
Tokyo Institute of Technology
nishina.k.aa@m.titech.ac.jp

Yagi YUTAKA
Picolab Co., Ltd.
yagi@picolab.jp

Bor HODOŠČEK
Osaka University
bor@lang.osaka-u.ac.jp

Takeshi ABEKAWA
National Institute of Informatics
abekawa@nii.ac.jp

Abstract
Japanese language learners aim to acquire reading, listening, writing and speaking skills. We at the Hinoki project (https://hinoki-project.org/) have recently been working on the Natsume collocation search system (https://hinoki-project.org/natsume/), the Natane learner corpus to support Natsume (https://hinoki-project.org/natane/) and the Nutmeg writing support system (http://hinoki-project.org/nutmeg/). In order to test the effectiveness of Nutmeg, we conducted an online experiment with 36 participants who used the system’s register misuse identification feature to correct four writing assignments. Results show that Nutmeg can be an effective tool in correcting common register-related errors, especially those involving auxiliary verbs. However, the accuracy of verb and adverb identification was too low, suggesting the need for improvements in the variety of corpora used for identifying register misuse.

Keywords: writing support system; learner corpus; academic writing; register; errors; performance evaluation experiment

Povzetek