Abstract: This paper demonstrates a method for creating Japanese learners dictionary of i-adjective-noun collocations. After an introduction of the importance of collocations and the necessity of their inclusion in Japanese language learning, we present various corpora types and corpus query tools that are used to obtain variety of collocational usage in different types of discourse. The Japanese language proficiency word-list is also used to provide levels of difficulty for dictionary headwords and their collocations. Further on, the process of collocation selection, analysis, and description of the i-adjective-noun collocation dictionary is presented.

Keyword: collocations, Japanese language, learners’ dictionaries, i-adjective-noun collocations, corpora

INTRODUCTION

In the last decade, various Japanese language electronic resources that obtain information on collocations have emerged. Computer assisted language learning systems, such as Natsume (Nishina, 2011), and corpus query tools, such as Chakoshi (Fukada, 2007) and Sketch Engine (Kilgarriff et al., 2004; Srdanović et al., 2008), extract collocations from one or multiple contemporary Japanese language corpora of various types.

At the same time, the first Japanese language collocation dictionaries have appeared, although with no use of such corpora. The first Japanese learners’ collocation dictionary (Himeno, 2004) covers collocations of verbs and collocation of na-adj ectives based on literary works, newspapers and other data. There are also a couple of dictionaries that are not targeted at learners of Japanese as a foreign language but cover noun-particle-verb collocations (Ogino et al., 2003; Kindaichi, 2006) or a wider spectrum of collocations based on modern and contemporary language (Onai, 2010).

The need for a systematic treatment of collocations in language learning has been widely recognized. For example, Kjellmer (1991) argues that it is necessary to shift the emphasis from individual words to their co-occurrences, and to stop teaching vocabulary items alone in order to approach the proficiency of a native speaker. James’s (1998) research notes that various combinations of collocations are frequently either overused or underused by language learners compared to how native speakers use the same collocations and points out the importance of learning word co-occurrences. Nation (2001) describes »unpredictable collocations« - the collocations that are difficult for language learners to predict based on knowledge of their native or other foreign languages since the constituents of the collocations in their native language are different from those in the foreign language. However, there is still a significant lack of collocation syllabi, collocation dictionaries for Japanese language learners, and other learning materials that cover Japanese language collocations (Srdanović, 2011).

Since collocational relations of i-adjectives and nouns have not been covered so far in Japanese language learners’ dictionaries and are present only in Onai (2010, but not based on large-scale electronic corpora, this paper presents the method for obtaining such kind of information from various corpora types and using various tools.
RESOURCES

There are various corpora types and various tools that are available for exploring Japanese collocational relation: (a) written Japanese language corpora, such as the Balanced Corpus of Contemporary Written Japanese (BCCWJ), the Mainichi Newspaper data, various types of textbooks, various specialized corpora etc.; (b) formal and informal spoken corpora, such as Oikawa or Nagoya corpus; and (c) the Japanese web corpora, such as JpWaC, Yahoo Chiebukuro, JpTenTen, Wikipedia etc. The more data is used, the richer is the obtained information on collocations and their distribution across various types of texts; moreover the results are more comparable.

In order to extract the most frequent and the most salient collocations from the corpora, the web-based tools that summarize collocational and grammatical relations, such as Sketch Engine, Chakoshi Natsume, and recently developed NINJAL-LWP can be used. There are also some limitations to be aware of – some corpora are available only inside the specific query system and some systems can be used only with the specific corpus or corpora incorporated into them.

The Sketch Engine tool was initially developed for the English language (Kilgarriff et al., 2004) and then extended to various European and non-European languages. In order to create the collocation dictionary models, the module Word sketches is especially relevant since it extracts collocational relations from the corpus and offers a summary of words' statistically most salient and most frequent grammatical and collocational relations classified into various collocational types. The tool has proven to be extremely useful in the fields of lexicography (e.g., Oxford University Press, Macmillan (Kilgarriff & Rundell, 2002), language teaching and linguistic research. The Japanese module (Srdanović et al., 2008) is capable of extracting approximately 50 types of collocational and grammatical relations within the Japanese language including the targeted adjective on –i and noun collocation type. It is based on the JpWaC, which is a well-balanced large-scale Japanese language web corpus containing 400 million words/tokens.

Similarly to Sketch Engine, the online system NINJAL-LWP, which was recently developed by National Institute for Japanese Language and Linguistics and Lago Institute of Language, supports lexical profiling including extraction of collocational relations classified into various collocational types (Pardeshi & Akasegawa, 2012). It is created for query of the Japanese balanced written corpus BCCWJ, which was built during the five-year project »MEXT Grant-in-Aid for Scientific Research Priority Area Program: Japanese Corpus« in cooperation between the National Institute for Japanese language and other research and pedagogical institutions in Japan (Maekawa et al., 2010).

Further on, Natsume Writing Support System is a CALL system that provides the ability to search for various collocational relations across various types of corpora and subcorpora including various subcorpora of BCCWJ. The system currently does not provide information on the targeted i-adjective-noun collocations; it provides detailed information for the noun-particle-verb collocation type.

Finally, Chakoshi is another tool that enables search of collocations, although not in a form of classified collocational types. It is publicly available for search of two corpora: Aozora bunko and Nagoya conversational corpus.

In addition, for the collocation dictionary model the Japanese language proficiency test (JLPT) word-list is employed to provide information on the difficulty level of constituents of collocations (Japan Foundation and Association of International Education Japan, 2004). Based on the list, the collocation constituents are divided into 4 levels, where 1 is the most difficult and 4 is the least difficult level, and an additional level 0 is used to mark words that
are not present in the existing word-list. The newest version of the word-list uses 5 levels, but it is not yet publicly available and is therefore not used in this research. Besides the original JLPT reference material, the level checker module of the Reading Tutor tool is used to provide an automatic level annotation of the collocation constituents (Kawamura, 1999).

PROCEDURE

The process of collocation selection, analysis and description for the i-adjective-noun collocation dictionary is performed in the following steps:

1. Defining the list of headwords: The first step is to extract and validate the word-list of i-adjectives as candidates for collocational dictionary headwords. The candidate list obtained from the large-scale web corpus, JpWaC, is compared to the JLPT. Among 500 most frequent i-adjectives that appear in the web corpus 188 are not present in the JLPT list. Two of them, kyōmibukai `very interesting` and habahiroi `broad`, are among the first hundred most frequent i-adjectives. The list of headwords is still to be compared to the word-list of i-adjectives in BCCWJ. The words from the list are divided based on their level and ordered alphabetically in each group. The comparison of word-lists resolves certain shortcomings of the JLPT word-list and of the corpora.

2. Analyzing the Japanese word sketches for a headword: The second step is to use the i-adjective candidates as keywords, extract the most important noun collocations from the corpus, based on the frequency and statistical salience, classify and order them based on difficulty levels. The Table 1 shows the twenty most frequent i-adjective-noun collocations for the adjective aoi `blue, green` obtained from Japanese word sketches and the JpWaC corpus. The collocations are divided into 0 to 4 levels and ordered based on their frequency. The first number in brackets presents the corpus frequency and the second the statistical relevance of the collocation.

Table 1 The Most Frequent i-Adjective-Noun Collocations for the Adjective aoi `blue, green` Obtained from the Japanese Word Sketches

| 青い, aoi (6415) | 4 空(602;9.5)、鳥(251;9.2)、海(235;7.8)、月(166;4.4)、頭(84;5.1)、色(68;5.3)、花(64;5.8)、ハンカチ(22;6.9) |
| | 3 光(92;6.5)、森(44;6)、桜(41;4.8)、葉(26;5.9) |
| | 2 眼(41;7.3)、果実(17;6.4) |
| | 1 芝(93;8.9)、山脈(35;7.7)、ユニフォーム(14;6.1) |
| | 0 蜻蛉(30;7.3)、炎(46;7)、珊瑚礁(16;6.7) |

3. Comparing collocations in different resources for the Japanese language (JpWaC, BCCWJ & other corpora): The Japanese word sketches provide collocations that are present in the JpWaC corpus. These can be compared to other results, such as those obtained by the Natsume system, which uses various corpora, by NINJAL-LWP, which uses BCCWJ, or by Chakoshi, which uses Nagoya conversational corpus and Aozora bunko. The comparison allows for specifying collocations that are most frequent in most of the corpora and most representative, and it furthermore provides additional information on collocations that are specific for a particular type of text. This is covered as information on usage differences in the dictionary. As additional resources, the existing Japanese language dictionaries will be employed.

4. Referring to the Japanese language learners’ corpora: Learners’ corpora are employed to check for frequent mistakes by language learners – first for headwords and then for their
collocations, which is a valuable indicator for unpredictable collocations as the most difficult
to master by language learners and very relevant data to be included into the dictionary.

5. Encoding information: The prepared database of collocations is encoded using the
XML format, which facilitates processing, the operability of the data, and its adoption to
syllabus or dictionary creation.

REFERENCES

Nihongo kagaku 22, 161-172.

Kenkyusha


nihongo kyoioku, 34, 1-22.


105-116.

collocation dictionary). Tokyo, Gakushu kenkyusha.

127.

Maekawa, K., Yamazaki, M., Maruyama, T., Yamaguchi, M., Ogura, H., Kashino, W., Ogiso,

University Press.

Natsume no kouchiku to hyoka. (Development and evaluation of the writting support system
Natsume using, the balanced corpus). Tokutei ryōiki denkō “Nihongo kōpasu” Heisei 22
nendo kōkai wakkushoppu yokōshū. Tokyo: Monbukagakusyō kagakukyoku
tokuteiryōiki denkōyu ’Nihongo koopasu’ sookatu ban, 215-224.

valency). Tokyo, Sanseido.


sakusei: koopasu braujingu shushttmu NJNIAL-LWP no tokuchou to kinou. Tokutei ryouiki
denkōyu Nihongo koopasu Gendai Nihongo kakikotoba kinkou koopasu kansei kinen
yokoushuu, 205-216.

Japanese. Shizen gengo shori (Journal of Natural Language Processing) 15/2, 137-159.

Srdanović, I. (2011). Colloccational Relations in Japanese Language Textbooks and Computer-