A Research Project on Language Resources for Learners of Japanese

Irena SRDANOVIĆ
Department of Asian and African Studies
Faculty of Arts
University of Ljubljana

Abstract
This research note presents the ongoing postdoctoral project Language Resources for Slovene Learners of the Japanese Language *supported by the Slovenian Research Agency ARRS. The project addresses the phenomenon of (un)predictability of collocations and aims to develop language resources for Japanese language learners*.

Keywords: research project, corpora, language resources, Japanese, collocations, second language learners

要旨
本稿は「スロベニアの日本語学習者の言語リソース」というポストドクター研究プロジェクトを紹介する。プロジェクトは、コロケーションの（不）予測性の現象に対処し、日本語学習者のための言語リソースの開発を目指している。

キーワード：研究プロジェクト、コーパス、言語リソース、日本語、コロケーション、第二言語学習者

1. Research background

With the development of corpora and their application to language research and learning, it has been demonstrated on the basis of empirical evidence, that words do not function in isolation but are co-selected with other words to produce meaning, and consequently awareness of the importance of collocations and

context has grown steadily (Hunston and Francis 1998: 45-72; Partington 1998; Sinclair 1991; Stubbs 2001).

Furthermore, certain studies emphasize the importance of collocation learning. Kjellmer (1991: 111-127) argues that it is important to study word co-occurrences in order to approach the proficiency of a native speaker. It is necessary to shift the emphasis from individual words to their co-occurrences and to stop teaching vocabulary items alone. In his research on the 'naturalness' of collocation, James (1998) confirms that non-native speakers often produce unnatural collocations and underscores the importance of learning word co-occurrences. His research points out that various combinations of collocations are frequently either overused or underused by language learners compared with how native speakers use the same collocations. Nation (2001) describes the so-called 'unpredictable collocations', i.e. collocations that are difficult for language learners to predict based on knowledge of their native or other foreign languages. According to Nation (2001), the learning burden associated with a collocation is connected to its predictability and the less predictable the collocation the greater the learning burden for a foreign language learner.

Although the need for a systematic treatment of collocations in language learning has been recognized in general, the greatest advances in the field have been made for the English language, with many other languages falling far behind. Japanese language education, for instance, still relies to a large extent on the intuition and experience of teachers and language specialists, and there is still a significant lack of materials that could be acquired through the use of cutting-edge technology and methodologies which would yield innovative and objective language descriptions. Furthermore, there is still a need to obtain new theoretical and practical insights into the treatment of collocations from the points of view of similarity and divergence between first and second languages, and the nature of 'unpredictable' and 'predictable' collocations, which is not possible to gain from a mere monolingual approach.

2. The goal and phases of the project: language components for the JAP and JSP modules

2.1. Project goal

The main focus of the proposed research project[1] is to develop language resources for Slovenian learners of the Japanese language. The project applies empirical methods of corpus linguistics and the latest language technologies through the perspective of lexical learning, and stresses the need for a systematic treatment of collocations or words that typically co-occur in a text. Moreover, the project attempts to provide new theoretical insights into the treatment of collocations from the points of view of predictability and unpredictability; learning burden; similarity and divergence between first and second languages; and finally, differences in collocation usage at the academic level or in the context of a specific profession.

2.2. Project phases

The proposed research project aims to develop two kinds of modules for Slovenian learners of Japanese and will be conducted in three phases.

Phase I - The preparation phase consists of a needs analysis and the preparation of specifications for language resources, and the creation of collocation syllabi and a dictionary model (JAP and JSP).

Phase II - The second phase consists of the creation of two modules:

> **Module 1** - corpus-based language resources for Slovene learners of Japanese for Academic Purposes (JAP) based on general language skills and containing the following components:
>
> > (1a) a collocation query system containing information on various language proficiency levels;
> >
> > (1b) a collocation syllabus covering theoretical aspects of the treatment of collocations emphasizing points of similarity and divergence between first and second languages;

(1c) a model for the Japanese-Slovene-Japanese dictionary of collocations.

Module 2 - corpus-based resources for learners of Japanese for Specific Purposes (JSP), with the following components:

(2a) Domain-specific corpus and word list creation;

(2b) Domain-specific collocation syllabi and dictionary.

Phase III - The final phase consists of the creation of a web page with a query function for collocations, and preparation for publication.

3. Resources and methods

For the creation of modules 1 and 2 (phase II), the project employs various types of Japanese and Slovene language corpora:

(a) written corpora: such as the Balanced Corpus of Contemporary Written Japanese (BCCWJ) developed at the National Institute for Japanese Language in Tokyo (Maekawa et al 2010: 1483-1486), various types of textbooks, Slovene corpora such as GigaFida, Nova beseda, etc. (Logar and Krek 2010); (b) spoken corpora: such as the Nagoya corpus; and (c) Japanese and Slovene web corpora: such as JpWaC, JpTenTen and slWaC. In addition, for certain components, corpora of other languages are used (e.g. the British National Corpus, BNC).

In order to extract the most frequent and the most salient collocations from the corpora, the research uses web-based tools that summarize collocational and grammatical relations, such as Sketch Engine. The Sketch Engine tool was initially developed for the English language (Kilgarriff et al. 2004: 105-116) and then extended to various European and non-European languages. The author of this paper developed the first Japanese version of the tool in collaboration with others (Srdanović et al. 2008). The tool was capable of extracting approximately fifty types of collocational and grammatical relations within the Japanese language. The tool presents a one-page summary of a given word's grammatical and collocational relations. It has proven to be extremely useful in the fields of lexicography[2] (Kilgarriff and Rundell 2002: 807-818), language teaching (e.g.

Chen et al. 2007), and linguistic research. Since the system is further adjusted for pedagogical purposes by adding suitable information concerning proficiency levels for collocation words, the JLPT (Japanese Language Proficiency Test) word list is used. The JLPT list is the most widely used vocabulary list for Japanese as a foreign language (Japan Foundation and Association of International Education Japan 2004). The list was used as a standard for the creation of Japanese language textbooks as well as tests for measuring the proficiency level of learners. It divides the Japanese vocabulary into four levels, where 1 is the most difficult and 4 is the least difficult level. The newest version of the word list uses five levels, but the data are not yet publicly available and is therefore not used in this project. In addition, level 0 is used to mark words that are not present in the existing word list. The research aims to adjust the description of collocations to pedagogical needs so that the information on the difficulty level of words is available.

Another resource used in the process of selection of collocations is the Slovene reference corpus and the word sketches for Slovene (Logar and Krek 2010). Grammatical and collocational relations in Slovene retrieved from the reference corpus provide additional tips on possible differences in collocational pairs in the two languages. The results are also compared to examples in the Japanese-Slovene parallel corpus available through the web page of the Japanese-Slovene dictionary JaSlo (Hmeljak 2002: 102-105). To describe the collocations in Slovene, other references, such as the monolingual Slovene dictionary SSKJ, etc., are used.

Special emphasis is placed on collocations recognized as unpredictable or difficult to predict for Slovene learners. At this stage of the project, the aim is to cover the collocations that have constituents of the lowest difficulty levels (JLPT 4 and 3) and that are taught in the first and second year of Japanese language studies. Frequently, collocations incorrectly learnt in the lower grades occur later in language use; it is, therefore, important to cover them systematically from the very beginning. For this reason, learners of Japanese language corpora are also used for searching frequent mistakes related to collocation usage.

For the JSP module, the research applies the BootCat technology, designed for the extraction of suitable texts from the World Wide Web and for the creation of specialized corpora and specialized word lists, as described in Baroni and Ueyama (2004: 13-16), and Baroni and Bernardini (2004). In addition, the research uses other important reference sources, such as textbooks and dictionaries.

4. Relevance of the project

The proposed research project bears upon the following areas of linguistics: corpus linguistics, foreign language learning, e-learning, corpus lexicography, applied linguistics and lexical semantics. Although all of the research areas touch upon the importance of collocations in language communication, collocations have still not been treated in the manner proposed in this project. This especially concerns the innovative combination of state-of-the-art methodology of corpus linguistics and linguistic technologies for extraction and grouping of the most salient collocational and grammatical relations with categories from both the theory and the practice of foreign language learning, such as learning burden, predictability of collocations, etc. Besides these innovative methods, the comparison of collocational relations in the two languages of Slovene and Japanese and the empirically based systematic treatment of the collocations are expected to provide new and contemporary linguistic findings from the abovementioned areas. This is especially expected with respect to new theoretical and practical insights into the nature of predictability and unpredictability of collocations for learners of a foreign language which would contribute to the development of science in the field of linguistics. The importance of the proposed theoretical framework will also be in its possible application to lexicography, creation of syllabi, and e-learning.

The project is expected to contribute to the development of the emerging field of Language for Specific Purposes learning and teaching, which is an industry-relevant branch of language pedagogy. In the field, the novel possibilities for fast and efficient development of specialized corpora have been gradually recognized, but the advantages of these methods for extraction of collocations for the purpose of language learning have so far not been sufficiently exploited. The proposed project aims to enhance the existing approaches through a novel way of extracting domain-specific collocations and their innovative treatment from the pedagogical

perspective of learning burden and predictability. The theoretical and practical findings of the proposed research are of a wider importance to the research community and are expected to provide some new valid insights into the aspects of collocation usage which could also be applicable to other languages.

5. Ongoing research

The author of the present paper has carried out and published the following research studies during the present ongoing research project.

(a) Srdanović Irena (2013). 大規模コーパスを用いた形容詞と名詞のコロケーション記述的研究: 日本語教育のための辞書作成に向けて [Description of Adjective-Noun Collocations Based on Large Scale Corpora: Towards a Dictionary for Japanese Language Learners]. 国立国語研究所論集 [NINJAL Research Papers], No. 6.

This research concentrates on adjective-noun collocations and describes methods for creation of two resources based on two large-scale corpora of contemporary Japanese (BCCWJ and JpTenTen): "Adjective-Noun Collocation Data" and "Japanese Language Learner's Dictionary of Adjective-Noun Collocations". The highly frequent adjective *takai* serves as a model to show how the data from the first resource can be used as a basis for the creation of a dictionary for Japanese language learners.

(b) Srdanović Irena, Sakoda Kumiko (2013). Analysis of learner's production of adjectives using the Japanese language learner's corpus CJAS. *Acta linguistica asiatica* Vol. 3, No. 2., pp. 75-84.

The paper explores learner production of adjectives using the Japanese language learner's corpus C-JAS (Corpus of Japanese as a second language) and shows how the learner's production of adjectives develops in terms of form, correct/incorrect usages, and lexico-semantic coverage. The approach is relevant for discovering what combinations are predictable or unpredictable for language learners.

(c) Srdanović Irena (2013). Japanese *i*-adjectives as short and long unit words: implications for language learning. *PACLING 2013: Conference proceedings, September 24, 2013 Tokyo: Pacific Association for Computational Linguistics,* 8.

This paper examines Japanese language *i*-adjectives that are annotated as short and long-word units in the Balanced Corpus of Contemporary Written Japanese (BCCWJ). The large gap between the number of lemma as short and long-unit words (suw and luw), besides revealing the nature of lemma design in the new morphological dictionary UniDic and BCCWJ, gives some new insights into productivity in Japanese.

(d) Irena Srdanović et al. (2013). 百億語のコーパスを用いた日本語の語彙・文法情報のプロファイリング [Japanese language lexical and grammatical profiling using the web corpus JpTenTen]. 第 3 回コーパス日本語学ワークショップ. Tokyo, February 28 – March 1, 2013. *National Institute for Japanese Language and Linguistics (NINJAL), Department of Corpus Studies, Center for Corpus Development*, pp. 229-238.

This paper describes the creation of a Japanese language lexical and grammatical profiler that is based on the large-scale web corpus JpTenTen. The tool enables extraction of more than fifty different collocation patterns in Japanese and as such is a very relevant resource for the ongoing project.

(e) Srdanović Irena (2014). Corpus based collocation research targeted at Japanese language learners. *Acta linguistica asiatica*. Vol 4, No. 2., pp. 25-35.

This research shows how 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 a foreign language pointing out some learning difficulties that need to be addressed in the ongoing projects.

(f) Srdanović Irena, Kaseda Harumi (2014). Selecting advanced Japanese language vocabulary for tourism. *The 14th International Conference of the EAJS, Department of Asian and African Studies, Faculty of Arts, University of Ljubljana, Ljubljana, August 27-30, 2014. Book of abstracts.* [Ljubljana]: Ljubljana University Press, Faculty of Arts, p. 653

The paper describes how a specialized web corpus for the domain of tourism can be created, and explores the usage of the corpus for creation of advanced Japanese language vocabulary.

[1] This is a basic post doctorate project supported by the Slovenian Research Agency ARRS. The project was proposed in response to the public call for (co)financing of research projects in 2012 - call in 2011, proposal no. 2011II/814, but started a year later due to circumstances at ARRS. The duration of the project is two years, beginning August 2013 and ending August 2015. The project leader is the author of this research note.

[2] e.g. Oxford University Press, Macmillan.

Bibliography

BARONI Marco and BERNARDINI Silvia (2004). BootCat: Bootstrapping Corpora and Terms from the Web. *Proceedings of the Fourth Language Resources and Evaluation Conference LREC2004* (Lisbon).

BARONI Marco and UEYAMA Motoko (2004). Retrieving Japanese Specialized Terms and Corpora from the World Wide Web. In Ernst Buchberger (Ed.). *Proceedings of KONVENS* (Vienna), ÖGAI.

CHEN A., RYCHLY P., HUANG C.-R., KILGARRIFF A., and SMITH S. (2007). A Corpus Query Tool for SLA: Learning Mandarin with the Help of Sketch Engine. *Practical Applications of Language Corpora* (Lodz, Poland).

HMELJAK SANGAWA K. (2002). Slovar japonskega jezika za študente japonščine [The Japanese Language Dictionary for Students of Japanese]. *Zbornik Konference o jezikovnih tehnologijah* [The Proceedings of the Conference on Language Technologies] *SDJT'02*, 14-15 October 2002 Ljubljana: Inštitut Jožef Stefan.

HUNSTON Susan and FRANCIS Gill (1998). Verbs Observed: A Corpus-driven Pedagogic Grammar. *Applied Linguistics*, Vol. 19, No. 1.

JAMES C. (1998). *Errors in Language Learning and Use*, London: Longman.

Japan Foundation and Association of International Education Japan. (2004). *Japanese Language Proficiency Test: Test Content Specifications*. Tokyo: Bonjinsha.

KILGARRIFF Adam and RUNDELL Michael (2002). Lexical Profiling Software and its Lexicographic Applications: A Case Study, *Euralex 2002 Proceedings*.

KILGARRIFF Adam, RYCHLY Pavel, SMRŽ Pavel and TUGWELL David (2004). The Sketch Engine. *Euralex 2004 Proceedings*.

KJELLMER G. (1991). A Mint of Phrases. In Karin AIJMER and Bengt ALTENBERG (eds). *English Corpus Linguistics*. London: Longman.

LOGAR BERGINC N., KREK S. (2010). New Slovene Corpora within the "Communication in Slovene" Project. *Slavicorp Conference*. (Warsaw).

MAEKAWA K., YAMAZAKI M., MARUYAMA T., YAMAGUCHI M., OGURA H., KASHINO W., OGISO T., KOISO H., DEN Y. (2010). Design, Compilation, and Preliminary Analyses of Balanced Corpus of Contemporary Written Japanese. *Proceedings of LREC*. (Malta).

NATION P. (2001). *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.

PARTINGTON Alan (1998). *Patterns and Meanings: Using Corpora for English Language Research and Teaching*. Amsterdam: John Benjamins.

SINCLAIR John McHardy. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

SRDANOVIĆ ERJAVEC Irena, ERJAVEC Tomaž and KILGARRIFF Adam (2008). A Web Corpus and Word-Sketches for Japanese. *Shizen Gengo Shori (Journal of Natural Language Processing)* Vol. 15, No. 2. <www.jstage.jst.go.jp/article/imt/3/3/3_529/_article>

STUBBS Michael (2001). *Words and Phrases: Corpus Studies of Lexical Semantics*. Oxford; Malden, Mass.: Blackwell Publishers.

TOGNINI-BONELLI Elena (2002). Between Phraseology and Terminology in the Language of Economics. *Phrases and Phraseology - Data and Descriptions*. Bern, Switzerland: Peter Lang.