Construction of Associative Vocabulary Learning System for Japanese Learners

Takehiro Teraoka
Faculty of Engineering,
Takushoku University
815-1, Tatemachi, Hachioji-shi,
Tokyo 193-0985, Japan
tteraoka@cs.takushoku-u.ac.jp

Tetsuo Yamashita
Institute of Japanese Language Education,
Takushoku University
3-4-14, Kohinata, Bunkyo-ku,
Tokyo 112-8585, Japan
tyamashi@ner.takushoku-u.ac.jp

Abstract

In recent years, natural language processing (NLP) technology has been applied to support systems and teaching materials for learning Japanese in the second language acquisition. Research on Japanese language learning using language resources and techniques related to NLP is divided into two main types: "dictionary/corpus" and "learning-support system". While some corpora and learning-support systems are useful, most relay on word co-occurrence information. To improve learning efficiency with a different approach, we aim to enable second language learners of Japanese to acquire vocabulary along with associative information of Japanese native speakers by using the Associative Concept Dictionary for Verbs (Verb-ACD). In this study, we have constructed a vocabulary learning system that generates question and answer sets and other correct candidate sets by extracting and combining associative information from the Verb-ACD. We investigated how accurately the proposed system could generate the question and answer sets for Japanese learners and found that it had the generation accuracy (0.70). We conclude that the various associative relationships among the generated words are important and require co-occurrence information to check their consistencies.

1 Introduction

The number of Japanese language learners in the world has increased from about 580,000 in 1984 to approximately 3.67 million in 2015. While this number decreased a little from 2012 to 2015, the number of institutions for Japanese language education and the number of teachers increased slightly in a total of 137 areas including 130 countries and seven regions. These numbers cover only "school and other institutions teaching the Japanese language as language education", and learners who study Japanese at other institutions or who self-study through various media (television, radio, books, the Internet, etc.) are not included. All told, the true number of Japanese language learners is estimated to be much higher (The Japan Foundation, 2017).

Recently, natural language processing (NLP) technology has been applied to systems for learning languages and to teaching materials for acquiring language knowledge. Research on Japanese language acquisition with language resources and techniques related to NLP is divided mainly into "dictionary/corpus" and "learning-support system" categories. The former includes the Japanese Learner's Written Composition Corpus[1] (Lee et al., 2013), which summarizes the composition data of Japanese learners, the Japanese Educational Vocabulary (Sunakawa et al., 2012), which contains 18,000 words for Japanese vocabulary education based on the Balanced Corpus of Contemporary Written Japanese (BCCWJ) (Maekawa et al., 2014) and a Japanese textbook corpus, and the Tsukuba Web Corpus (TWC)[2], which includes 1,138 million words collected from the Web.

The latter category includes the Learning Item Analysis System[3], which outputs learning-item and

[1] http://sakubun.jpn.org
[2] http://nlt.tsukuba.lagoinst.info/
[3] http://www.intersc.tsukuba.ac.jp/ kyoten/en/lias.html

level-judgment results for input Japanese texts, and the text readability measurement system JReadability[4], (Hasebe and Lee, 2015) which outputs the degree of reading difficulty for input Japanese sentences. Thus, there are various learning-support systems and teaching materials for Japanese learners that have been constructed and developed. However, while these systems and materials contain a large amount of data and are generally useful, most use word co-occurrence information that consists of co-occurrence frequency in sentences taken from Japanese documents (i.e., news corpora and Web texts). In other words, they use the collocation of written words.

In this study, to improve learning efficiency, we constructed an Associative Vocabulary Learning System that is based on Japanese native speakers' associative information for basic verbs. This information is extracted from the Associative Concept Dictionary for Verbs (Verb-ACD) (Teraoka et al., 2010), which we previously constructed and extended from large-scale association experiments. On the basis of our analysis of the difference between word co-occurrence information and word associative information (Teraoka, 2018), we expect our learning system to enable learners of Japanese to acquire vocabulary along with the word associations of Japanese native speakers.

2 Verb-ACD

Verb-ACD consists of three elements: stimulus words, associated words from the stimulus words with semantic relations, and word distances among the two. To collect associative information on verbs, we conducted large-scale association experiments on the Web, where the stimulus words were basic verbs with ten semantic relations corresponding to deep cases: Agent, Object, Source, Goal, Duration, Location, Tool, Aspect, Reason, and Purpose. These verbs were selected from Japanese elementary school textbooks (Kai and Matsukawa, 2001) and the entries were prioritized as in basic Japanese dictionaries (Morita, 1989; Koizumi et al., 1989).

We used the linear programming method to quantify the word distance between a stimulus word and an associated one. As shown in Eq. (1), the distance $D(x, y)$ between a stimulus word x and associated word y is expressed with Eqs. (2)–(4):

$$D(x,y) = \frac{7}{10}IF(x,y) + \frac{1}{3}S(x,y) \quad (1)$$

$$IF(x,y) = \frac{N}{n(x,y)+\delta} \quad (2)$$

$$\delta = \frac{N}{10} - 1(N \geq 10) \quad (3)$$

$$S(x,y) = \frac{1}{n(x,y)}\sum_{i=1}^{n(x,y)} s_i(x,y). \quad (4)$$

Table 1 lists the deep cases and examples when the stimulus word is 運ぶ (*hakobu*) 'convey'. The distance consists of the inverse frequency of an associated word $IF(x, y)$ and the average associated word order $S(x, y)$. Each coefficient was obtained using the simplex method. Let N denote the number of participants in the experiment and $n(x, y)$ denote the number of those who responded with the associated word y to the stimulus word x. Let δ denote a factor introduced to limit the maximum value of $IF(x, y)$ to 10 and let $s(x, y)$ denote the associated word order of each participant. Three elements—the stimulus verbs, associated words, and the distances between them—were used to construct Verb-ACD (Teraoka et al., 2012).

Deep case	Associated words (Word distance)
Agent	I (3.60), Mover (4.21)
Object	Package (1.36), Furniture (7.78)
Source	House (1.45), School (3.81)
Goal	House (1.92), Station (3.73)
Duration	Morning (2.71), Midnight (5.88)
Location	Warehouse (3.73)
Tool	Car (1.62), Hands (3.47)
Aspect	Desperately (3.17)

Table 1: Example of associated words in Verb-ACD (stimulus word: 運ぶ (*hakobu*) 'convey').

There are currently 773 stimulus verbs in Verb-ACD, and the total number of participants was approximately 3,200. All participants were undergraduate and graduate students of Keio University or the Tokyo University of Technology. For this study, each stimulus verb was presented to 40 participants. There were approximately 305,000 associated words. When all overlapping words were elim-

[4]https://jreadability.net/

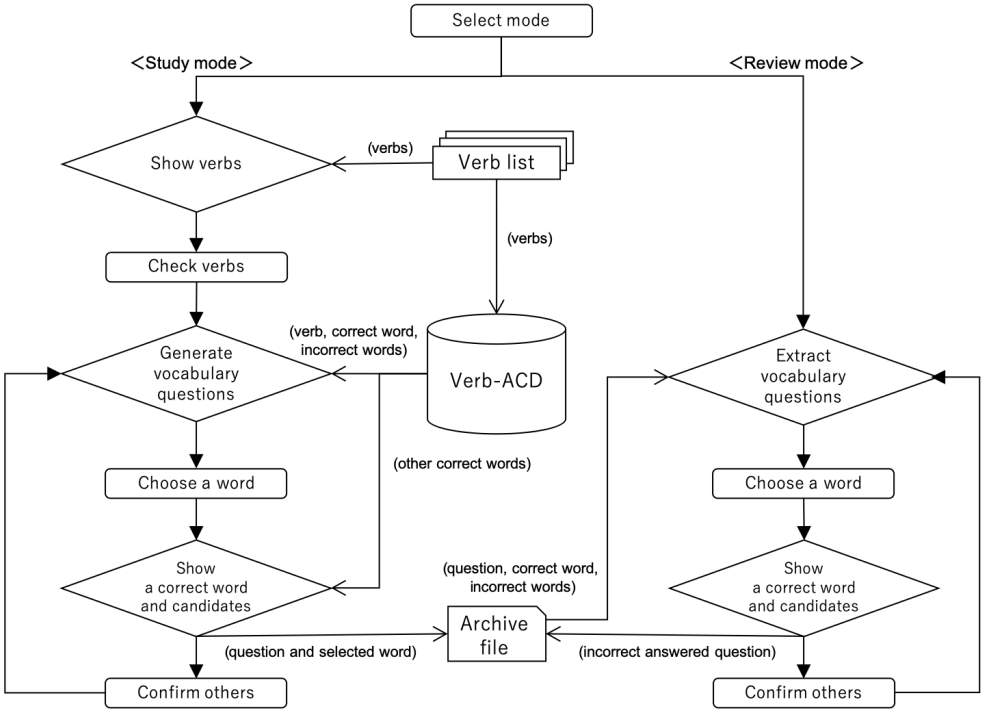

Figure 1: System outline.

inated, approximately 58,000 associated words remained.

3 Associative Vocabulary Learning System

3.1 System outline

Our proposed system runs on the Web, so learners need to operate it on a Web browser. Figure 1 shows the outline of our system. The squares and diamonds represent the learner's handles and system processes, respectively. Two types of arrows represent the processing flows and data movements. First, a learner select Study mode or Review mode. In Study mode, the learner answers vocabulary questions that the system automatically constructs from the Verb-ACD. In Review mode, the learner repeatedly answers the questions until the correct answers are chosen.

When the learner selects Study mode, one of the verb lists is displayed (see the left-hand flow, Fig. 1). After checking any verbs he or she already knows, the learner answers questions generated automatically by the system. These are multiple choice questions consisting of one correct word and three incorrect words. The learner chooses a word and the system shows both the answer and other correct words, namely, associated words extracted from the Verb-ACD. The learner checks these and then clicks the 'Next' button. The system sends the log data of the question and selected choice to an archive file and then generates the next question. These steps are repeated for all of the verbs in the list, so the same number of sets of questions and answers are stored in the archive file.

When the learner selects Review mode, the system extracts question-and-answer sets from the archive file (see the right-hand flow, Fig. 1). The same as in Study mode, the learner answers the question by choosing a word from multiple choices, and the system shows the correct answer along with other correct words. If the learner selects an incorrect word, the system sends the question information that includes the correct answer and the selected incorrect word to the archive file. If the learner selects a correct word, the question information is deleted from the archive file. These steps are repeated for every question. As long as the learner makes mis-

takes in this mode, the data in the archive file does not decrease. Thus, this mode enables the learner to repeatedly answer the same questions until no more mistakes are made.

These two modes enable the learner to study and review Japanese vocabulary with native speakers' associative information. The details of how we extract the associative information from the Verb-ACD and generate vocabulary questions with associated words will be described in 3.2 and 3.3.

3.2 Verb Lists

We constructed our system as a learning support tool for learners who have just started studying Japanese, so the verb lists in the first step of Study mode are based on basic vocabularies that were level 3 or 4 in older versions of the Japanese-Language Proficiency Test (JLPT) (The Japan Foundation and Association of International Education, Japan, 1994) and correspond to N3 and N4 levels in the current JLPT[5]. At this level, students are assumed to understand everyday Japanese to a certain degree.

As mentioned in 3.1, our system provides as many sets of questions and answers as there are verbs in a verb list per Study mode. To help learners concentrate, the number of verbs in a list is set to less than 30. Thus, when learners spend around one minute on each question, it will take them about 30 minutes to complete each study session. Moreover, making each learner check which verbs they already know enables us to confirm which selected answers for checked verbs are correct or incorrect. Specifically, by referencing the data on the checked verbs, we can judge whether the answers were correct by chance or not.

Table 2 shows an example of a verb list. Verbs are presented in kanji and hiragana characters. As shown in Table 2, there are mostly transitive verbs (e.g., 食べる（たべる）'eat', 言う（いう）'say', 飲む（のむ）'drink', 閉める（しめる）'close', 手伝う（てつだう）'help', 捕まえる（つかまえる）'catch', 利用する（りようする）'use'), but also some intransitive verbs (e.g., 怒る（おこる）'anger', 泣く（なく）'cry', 鳴る（なる）'sound') and exalted terms (e.g., 召し上がる（めしあがる）'eat', 仰る（おっしゃる）'say') that have the same meaning.

[5]https://www.jlpt.jp/e/index.html

Verb（Japanese syllabary）	Remark
食べる（たべる）'eat'	transitive
召し上がる（めしあがる）'eat'	transitive, exalted
言う（いう）'say'	transitive
仰る（おっしゃる）'say'	transitive, exalted
怒る（おこる）'anger'	intransitive
泣く（なく）'cry'	intransitive
鳴る（なる）'sound'	intransitive
飲む（のむ）'drink'	transitive
閉める（しめる）'close'	transitive
手伝う（てつだう）'help'	transitive
捕まえる（つかまえる）'catch'	transitive
利用する（りようする）'use'	transitive

Table 2: Example of verb list.

3.3 Generation of Questions and Answers

The System automatically generates 1) questions and answer sets that include one correct word and three incorrect words and 2) other correct candidate sets by extracting associative information from the Verb-ACD. This processing consists of three steps.

First, after the learner checks which verbs he or she knows, the system obtains all the verbs in the list and extracts words that have been associated with these verbs (i.e., stimulus words) by means of the semantic relations of the Agent and Object in Table 1. Next, the system generates the question and answer sets. When a question is about the Agent, the question sentence consists of the following three components: a blank with a particle (e.g., は 'ha' or が 'ga'), an associated word of the Object with a particle (e.g., を 'wo') in which the distance from the verb is the shortest with the Object, and the verb from the list, as shown in the example below.

( ) が 意見（いけん）を おっしゃる。(in Japanese)
'( ) states an opinion.'

Next, as incorrect words, the system extracts three associated words with short distances from the other verb. This other verb has neither a strong nor weak relation with the verb in the question sentence because the three incorrect words that are associated words from this other verb must be absolutely different from the correct word. If they are not different, the learner cannot make a choice and cannot understand the differences between the correct

	Question and answer set	Other correct candidates set	All (Both sets)
Agent	0.68 (19/28)	0.64 (18/28)	0.54 (15/28)
Object	0.81 (67/83)	0.90 (75/83)	0.76 (63/83)
Total	0.77 (86/111)	0.84 (93/111)	0.70 (78/111)

Table 3: Results of accuracy in automatically generating the question and answer sets and other correct candidates sets.

word and the three incorrect ones. Therefore, this other verb has the moderately low similarity with the predicate verb in the sentence. Then, the system extracts three incorrect words of this other verb from the Verb-ACD. Here, we utilize the nearby or less than 0.3 cosine distance between two verbs by using Word2vec (Mikolov et al., 2013), the same as in our previous studies (Teraoka, 2018).

Finally, the system displays the question sentence along with four randomly arranged answers consisting of the correct word and the three incorrect words. After the learner chooses an answer, the correct word is shown at the top of the display, and the other correct candidate words are simultaneously shown in ascending order of the distance between them and the verb.

4 Experiment

4.1 Evaluation of Question and Answer Sets

To investigate how accurately the system can generate the question and answer sets for Japanese learners, we prepared 111 sets consisting of two types. This number means four times of the Study mode. The first type is a question sentence where learners choose a target word with Agent (as in the example in 3.3), and the other is a question sentence where learners choose a target word with Object.

In this evaluation, a specialist of Japanese language education who has been teaching Japanese as a second language to international students at a Japanese language institution judged whether the question and answer sets consisting of one question sentence, one correct word and three incorrect words, and other correct candidate words were suitable as vocabulary learning materials. The system provided question and answer sets for the N3 and N4 levels (as mentioned in 3.2), so we asked the specialist to consider these levels in judging them.

When the specialist judged that there was nothing wrong with the set or the other correct candidates, the generation by the system was considered successful. In contrast, when any one of these was judged inappropriate, it was considered unsuccessful. On the basis of these judgments, we calculated the accuracy in automatically generating the question and answer sets and other correct candidates.

4.2 Results

Table 3 shows the results of accuracy in generating the question and answer set and other correct candidates. When a target word with a blank in a question sentence was an Agent of the verb, the system appropriately generated 19 question and answer sets and 18 other correct candidate word sets in all 28 cases. The system with Agent had slightly higher than average accuracies of generating question and answer sets (0.68) and other correct candidate sets (0.64). For both sets, only 15 cases were generated accurately, and the accuracy with Agent was about half of all 28 cases (0.54).

On the other hand, the system appropriately generated 67 question and answer sets and 75 other correct candidate word sets in all 83 cases when a target word with a blank in a question sentence was the Object of the verb. As such, the system on Object had the highest accuracies of generating question and answer sets (0.81) and other correct candidate sets (0.90). The system successfully generated 63 cases consisting of both correct sets, and the accuracy on Object was clearly higher than that on Agent (0.76).

In terms of appropriately generating all of the question and answer sets and other correct candidate sets, the system showed accuracies of 0.77 and 0.84, respectively. The overall accuracy for both sets was therefore reasonably high (0.70).

	Question sentence	Correct word	Incorrect words	Other correct candidates
Agent	( ) が 意見（いけん）を おっしゃる。 ‘( ) states an opinion.’	先生（せんせい） ‘Teacher’	国 ‘Country’ 街 ‘Town’ 村 ‘Village’	上司 ‘Boss’ 社長 ‘Company president’ 教授 ‘Professor’
Object	親（おや）が ( ) を育てる（そだてる）。 ‘Parents raise ( ).’	子供（こども） ‘Child’	鈴 ‘Bell’ ベル ‘Bell’ 鐘 ‘Chime’	植物 ‘Plant’ 動物 ‘Animal’ 犬 ‘Dog’

Table 4: Examples of correctly generated question and answer sets and other correct candidates. Parentheses and kana characters in incorrect words and other correct candidates are omitted.

	Question sentence	Correct word	Incorrect words	Other correct candidates
Agent	( ) が 曇る（くもる）。 ‘( ) is cloudy.’	空（そら） ‘Sky’	子供 ‘Child’ 私 ‘I’ 後輩 ‘Junior’	**天気 ‘Weather’** 表情 ‘Facial expression’ 顔色 ‘Complexion’
Object	ピッチャーが ( ) を投げる（なげる）。 ‘A pitcher throws ( ).’	ボール ‘Ball’	絵の具 ‘Paints’ 材料 ‘Material’ 粉 ‘Powder’	**ゴミ ‘Garbage’** **さじ ‘Spoon’** **槍 ‘Spear**

Table 5: Examples of incorrectly generated question and answer sets and other correct candidates. Parentheses and kana characters in incorrect words and other correct candidates are omitted. The bold and underline mean the inappropriate place.

4.3 Discussion

Table 4 lists examples of the question and answer sets and other correct candidate sets that were generated correctly. Both examples consist of three main trends. First, the target word with Agent or Object in the question sentence has a strong relation to the verb. For example, the verb おっしゃる in the table is a Japanese exalted word meaning ‘say’ that is frequently used in daily conversations with, e.g., a superior. Therefore, the correct word (i.e., 先生 ‘Teacher’) and other correct candidates (i.e., 上司 ‘Boss’, 社長 ‘Company president’, and 教授 ‘Professor’) were suitable for the sentence. Second, some incorrect candidates (e.g., 鈴 ‘Bell’, ベル ‘Bell’, and 鐘 ‘Chime’) were extracted from other verbs that had less than 0.30 cosine distance from the predicate verb in the sentence, and they were clearly different from the correct word (i.e., 子供 ‘Child’). Thus, it is important for learners to understand why they are different. Third, a word (e.g., 親 ‘Parents’) that was extracted with a different semantic relation and combined with a sentence had a relation to the correct word (e.g., 子供 ‘Child’). In other words, independently extracted associated words that were the correct word and another word in the sentence had co-occurrence relationships.

Table 5 shows examples of incorrectly generated question and answer sets and other correct candidate sets. In this table, the bold and underline of other correct candidates mean incorrect. When the question sentence and other correct word were ( ) が曇る ‘( ) is cloudy’ and 天気 ‘Weather’, these semantic relationships were suitable at first glance. However, if this word 天気 was in the parentheses of this sentence, this sentence 天気が曇る was strange as a Japanese sentence even though the translated English sentence ‘Weather is cloudy’ makes sense. In this case, the sentence 天気は曇りです is the correct Japanese sentence. In other words, even if the semantic relationship between the word and the sentence matches, the expression as a Japanese sentence was wrong. The accuracy for Agent was lower than that for Object due to composing such inappropriate sentences with other correct candidates and parentheses in the sentence.

Another example in Table 5 shows incorrect other correct candidates (e.g., ゴミ 'Garbage', さじ 'Spoon', and 槍 'Spear'). Unlike the example in the previous paragraph, even if these candidates were in the parentheses of the sentence ピッチャーが ( ) を投げる 'A pitcher throws ( )', their expressions as Japanese sentences were appropriate. Compared to the relationship between this sentence and the correct word ボール 'Ball', that between these candidates and this sentence had less relationship because the word ピッチャー 'Pitcher' in this sentence and the correct word ボール 'Ball' had a co-occurrence relation in a specific scene (e.g., a baseball game). Thus, the relationship between the sentence and these correct candidates was different from that between the sentence and the correct word.

Overall, the question sentences, answer sets consisting of correct words and three incorrect words, and other correct candidate sets generated by the system depended on various relationships between words. Specifically, the relationships between the verb and the correct word, between the verb and the word in the sentence, between another verb and incorrect words, and between the verb and other correct candidates were based on associative information extracted from the Verb-ACD and were important components for generating them appropriately. Meanwhile, the sentence structure must be changed by combining associative information. Furthermore, co-occurrence relations between the words in the sentence, the correct word, and other correct candidates were also important to compose the question and answer sets and the other correct candidate sets.

5 Conclusion

Our proposed vocabulary learning system showed a high accuracy for generating question and answer sets by extracting associative information from the Verb-ACD. The results of our evaluation demonstrated that various relationships between generated words were important, while inappropriate sentences and incorrectly generated other candidates were due to not considering such co-occurrence. In future work, we will add co-occurrence information to the system for checking the consistencies of generated words.

Acknowledgments

We would like to thank all the experiment participants from Keio University and the Tokyo University of Technology for their help with associated words and stimulus and semantic relations. This work was supported by JSPS KAKENHI Grant Number 18K12434.

References

Yoichiro Hasebe and Jae-Ho Lee. 2015. Introducing a Readability Evaluation System for Japanese Language Education. In *Proceedings of the 6th International Conference on Computer Assisted Systems for Teaching & Learning Japanese (CASTEL/J)*, pages 19–22.

Mutsuro Kai and Toshihiro Matsukawa. 2001. *Method of Teaching Lexicon-Vocabulary*. Mitsumura Tosho Publishing. (in Japanese).

Tamotsu Koizumi, Michio Funakoshi, Kyoji Honda, Yoshio Nitta, and Hideki Tsukamoto. 1989. *Japanese Basic Verb Dictionary of Usage*. TAISHUKAN Publishing. (in Japanese).

Jaeho Lee, Yayoi Miyaoka, and Hyunjung Lim. 2013. Learner's corpus and language test: Relationship between the language test score and text information content of the written composition. *Assessment and Evolution in Language Education (AELE)*, 3:22–31. (in Japanese).

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. 2014. Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient Estimation of Word Representations in Vector Space. *CoRR*, abs/1301.3781.

Yoshiyuki Morita. 1989. *A Dictionary of Basic Japanese*. Kadokawa Gakugei Shuppan Publishing. (in Japanese).

Yuriko Sunakawa, Jae-Ho Lee, and Mari Takahara. 2012. The Construction of a Database to Support the Compilation of Japanese Learners Dictionaries. *Acta Linguistica Asiatica*, 2(2):97–115.

Takehiro Teraoka, Jun Okamoto, and Shun Ishizaki. 2010. An associative concept dictionary for verbs and its application to elliptical word estimation. In *Proceedings of the 7th International Conference on Language Resources and Evaluation (LREC)*, pages 3851–3856.

Takehiro Teraoka, Ryuichiro Higashinaka, Jun Okamoto, and Shun Ishizaki. 2012. Automatic Detection

of Metonymies using Associative Relations between Words. In *Proceedings of the 34th Annual Conference of the Cognitive Science Society (CogSci)*, pages 2417–2422.

Takehiro Teraoka. 2018. Analysis of Associative Information for Second Language Learning of Japanese. In *Proceedings of the 4th Asia Pacific Corpus Linguistics Conference (APCLC)*, pages 434–439.

The Japan Foundation. 2017. *Survey Report on Japanese-Language Education Abroad 2015*.