2018 HAWAII UNIVERSITY INTERNATIONAL CONFERENCES
ARTS, HUMANITIES, SOCIAL SCIENCES & EDUCATION JANUARY 3 - 6, 2018
PRINCE WAIKIKI HOTEL, HONOLULU, HAWAII

ANALYZING USEFULNESS OF DIALOGUES FROM CLOSED CAPTION TV CORPUS AS AN EXAMPLE OF CAN-DO STATEMENTS FOR LANGUAGE LEARNING

MOCHIZUKI, HAJIME
SHIBANO, KOHJI
INSTITUTE OF GLOBAL STUDIES
TOKYO UNIVERSITY OF FOREIGN STUDIES
TOKYO, JAPAN

Dr. Hajime Mochizuki
Prof. Kohji Shibano
Institute of Global Studies
Tokyo University of Foreign Studies
Tokyo, Japan

Analyzing Usefulness of Dialogues from Closed Caption TV Corpus as an Example of Can-do Statements for Language Learning

Synopsis:

This paper describes a clustering method by using Doc2vec, SVD, and k-means method, in order to find discourse segments extracted from a closed caption TV corpus using formulaic sequences related to can-do statements for language learning. We report a feature of discourse segments in each classified group, and analyze usability of the acquired discourse segments whether discourse segments can be used as sample dialogues for can-do statements.

Analyzing Usefulness of Dialogues from Closed Caption TV Corpus as an Example of Can-do Statements for Language Learning

Abstract:

This paper describes the specific results of some analyses regarding our closed caption TV corpus to investigate usefulness of discourse segments in the corpus. We will confirm whether target discourse segments are preferred dialogues to be used as examples of can-do statements for language learning. This paper also describes a clustering method by using Doc2vec, SVD, and k-means method, in order to find discourse segments extracted from a closed caption TV corpus using formulaic sequences related to can-do statements. We report a feature of discourse segments in each classified group, and analyze usability of the acquired discourse segments whether discourse segments can be used as sample dialogues for can-do statements.

Introduction

Today, can-do statements are commonly used in language education to indicate language proficiency level and improvement. For example, the Common European Framework of Reference for Languages: Learning, teaching, assessment (CEFR) (Council of Europe, 2001), widely recognized guidelines in language education, defines can-do statements that aim to provide teachers and students with clear guidelines and descriptions to assess proficiency levels. However, typical can-do statements give few examples of dialogues for each can-do. Most can-do statements require teachers to prepare specific sample dialogues that can be used for active learning materials. Because this work is difficult and labor intensive, teachers should be provided rich dialogues matched to each can-do statement. One of the problems is that it will be necessary a large amount of dialogues in order to find appropriate examples matched to can-do statements.

On the other hand, we have been collecting closed caption TV (CCTV) corpus from December 2012 and as of March 2017, the size of our corpus has reached over 232,000 TV programs, over 89 million sentences. In our CCTV corpus, there are 13 genres: “Animation,” “Sport,” “Culture and Documentary,” “Drama,” “News,” “Variety,” “Film,” “Music,” “Hobby and Educational,” “Information and Tabloid Style,” “Welfare,” “Theater,” and “Other.” Each text in the corpus is classified to at least one genre according to the classifications provided in the EPG (Electric Program Guide). TV scripts from various genres of programs anticipate to contain many natural or near-natural dialogues and spontaneous sentences. We anticipate that discourse segments extracted from the CCTV corpus can be used as sample dialogues for each can-do statement.

Furthermore existence of high-frequency word sequences are anticipated in the corpus. Such word sequences include collocations, idioms, and greeting expressions that are considered useful fixed phrases. In applied linguistics, these expressions are treated as formulaic sequences (FSs) (Conklin and Schmitt, 2012; Jiang and Nekrasova, 2007; Schmitt, 2004; Vlach and Ellis, 2010; Wray, 2002; Wray, 2008; Wray and Perkins, 2000). Conklin and Schmitt have reported that FSs were read more quickly than non-formulaic phrases by both native and non-native speakers (Conklin and Schmitt, 2008). It is therefore helpful for second language learners to recognize and to use FSs in sentences. One of our aims is to tie a discourse segment that contained an FS to a can-do statement so that it can be used as an example of practical dialogue for language education.

In the previous research, we acquired a practical FSs list (Mochizuki, H. and Shibano, K., 2017a) and extracted over 18 million discourse segments contained practical FSs (Mochizuki, H. and Shibano, K., 2017b). As the following steps of our research, we plan to tie each FS to a can-do statement, and further to tie the can-do statement to specific dialogue units that include the FS to give sufficient examples of dialogues for each can-do statement. Therefore we have to recognize a content of each

discourse segment, and find a can-do statement matched it. We think discourse segments should be classified according to their contents, before tying them to can-do statements. In this research, we perform clustering to classify them into several categories by according to discourse contents of the segment clusters. We use Doc2vec, SVD, and k-means method for clustering segments.

In the following sections, we show the specific procedures of the following points: (1) creating discourse segments, (2) search discourse segments in which sentences include the given FS as a query, (3) clustering segments by using Doc2vec, SVD, and k-means method. We report a feature of discourse segments in each classified group, and analyze usability of the acquired discourse segments whether discourse segments can be used as sample dialogues for can-do statements.

Creating Discourse Segments from a Text in a Closed Caption TV Corpus

In this section, we give a brief explanation about our CCTV corpus and describe a method for creating discourse segments from a text in the CCTV corpus. In Tokyo, there are seven major terrestrial television service stations which organize a Japanese nationwide. A characteristic of Japanese TV stations is that each TV station provides a wide variety of programs selected from different genres. According to the classifications provided in the EPG (Electric Program Guide), there are at least 13 genres: “Animation,” “Sport,” “Culture and Documentary,” “Drama,” “News,” “Variety,” “Film,” “Music,” “Hobby and Educational,” “Information and Tabloid Style,” “Welfare,” “Theater,” and “Other” in our CCTV corpus. We have been continuing to build a Japanese spoken language corpus from TV programs with closed caption since December 2012. Therefore we record all TV programs with CCTV data during a 24-hour period. They are amount of 4,200 programs per month. In our previous research, the CCTV corpus consisted of more than 65 million sentences from over 166,000 texts as of February 2016[1]. Figure 1 shows a part of a text in the CCTV corpus.

863	1:04:05.57	1:04:08.57	あッ すみません
864	1:04:13.95	1:04:18.07	いろいろ
865	1:04:13.95	1:04:18.07	ご迷惑を おかけしてしまって
866	1:04:18.07	1:04:22.74	本当に申し訳ありませんでした
867	1:04:18.07	1:04:22.74	いえ こちらこそ
868	1:04:22.74	1:04:26.56	お力になれなくて すみません
869	1:04:26.56	1:04:29.36	家を出るんですか?
870	1:04:31.56	1:04:37.50	ここは信也さんが

Figure 1: An example of a closed caption sentence.

The first column in Figure 1 signifies the sentence number, the second column is the beginning time, the third column is the ending time for displaying a caption string, and the fourth column is a caption string. We used these display timings to divide a text into segments. In general, it is not appropriate to use a text from an entire TV program as a single dialogue unit. Since the whole text is too long, it is better to use a part of the text including a formulaic sequence as a single unit of dialogue. We can assume that if a scene of dialogue in a TV program is continuing, closed caption scripts will also display continuously. Therefore, we assume two adjoining sentences can be disconnected when the ending time of the former sentence and the beginning time of the later sentence is not the same, or

[1] As of March 2017, our corpus has reached over 232,000 TV programs, 89 million sentences, and 913 million words.

their display timings are not the same. For example, in Figure 1, the ending time of sentence No. 863 is different from the beginning time of sentence No. 864. Therefore, we mark a segment boundary between No. 863 and No. 864. Similarly, the ending time of sentence No. 869 is different from the beginning time of sentence No. 870. Thus, sentences between No. 864 and No. 869 will be a single discourse segment as in Figure 2.

863	1:04:05.57	1:04:08.57	あッ すみません
864	1:04:13.95	1:04:18.07	いろいろ
865	1:04:13.95	1:04:18.07	ご迷惑を おかけしてしまって
866	1:04:18.07	1:04:22.74	本当に申し訳ありませんでした
867	1:04:18.07	1:04:22.74	いえ こちらこそ
868	1:04:22.74	1:04:26.56	お力になれなくて すみません
869	1:04:26.56	1:04:29.36	家を出るんですか?
870	1:04:31.56	1:04:37.50	ここは信也さんが

Figure 2: An example of discourses segments.

After dividing the texts into discourse segments, there were 18,771,787 discourse segments. Table 1 shows the total numbers of texts, words, sentences, and segments by genre.

	A	S	C	D	N	V	F
texts	13,947	6,193	20,002	15,709	31,641	27,717	1,000
words	29,842,244	34,278,679	52,967,834	81,793,987	157,191,145	145,861,881	8,609,937
sentences	4,043,830	3,011,801	4,586,022	10,591,826	10,118,969	17,501,766	1,210,028
segments	512,199	1,434,464	1,443,846	1,816,371	5,247,840	4,666,491	234,489
	M	H	I	W	T	O	Total
texts	2,670	19,961	24,071	2,448	711	31	166,101
words	7,255,132	39,356,229	88,686,760	6,104,123	3,423,377	72,633	655,443,960
sentences	752,461	4,646,833	7,839,272	52,1223	364,305	5,622	65,193,958
segments	161,698	657,620	2,455,793	122,456	4,234	14,266	18,771,787

Table 1: Statistics of the CCTV Corpus. In this table, A, S, C, D, N, V, F, M, H, I, W, T, and O refer to "Animation," "Sport," "Culture and Documentary," "Drama," "News," "Variety," "Film," "Music," "Hobby and Educational," "Information and Tabloid Style," "Welfare," "Theater," and "Other," respectively.

Search Discourse Segments for FSs

As mentioned above, as the following steps of our research, we plan to tie each FS to a can-do statement, and further to tie the can-do statement to specific dialogue units that include the FS to give sufficient examples of dialogues for each can-do statement. Therefore we have to recognize a content

of each discourse segment, and find a can-do statement matched it. We think discourse segments should be classified according to their contents, before tying them to can-do statements.

In order to analyze usefulness of dialogues from our CCTV corpus, we use the limited FSs as a given query to select discourse segments in this research. We especially use a pair of two FSs that are 'いらっしゃいませ (*Irasshai mase*)' and 'ください (*kudasai*)'. The FS 'いらっしゃいませ' is one of popular greeting expressions and is used to express speaker's feelings of welcome. For example, when customers come to a store, a sales-person will speak '*irasshai mase*' to the customers at first. If you visit a friend's house, your friend and his family will speak '*irasshai mase*' to you. The second FS 'ください' means 'thanks to do' 'please do' or 'please something' and is used to express speaker's wants or desires. It is difficult to explain or fully translate meanings of '*kudasai*' as it is a fragment and is semantically ambiguous. The meaning of it changes by the precedent word(s). If you buy one of something, you can say '*1tu kudasai* (Can I have one?).' If you want someone to sit down, you can say '*suwatte kudasai* (Please sit down).'

Though 'irasshai mase' is only used by hosts or sales-people, 'kudasai' can be used by both hosts and guests, and both sales-people and customers. Therefore we expect that discourse segments including both FSs have a wide variety of situations and functions of dialogues.

As mentioned above there are 13 genres in our CCTV corpus. Among them, we use four genres that are "Drama (D)," "Information/Tabloid Style (I)," "Variety (V)," and "Animation (A)" in this paper. We search discourse segments on the four genres that have both two FSs. The numbers of searched segments are 738 from genre D, 222 from genre I, 948 from genre V, and 96 from genre A, respectively.

Clustering Discourse Segments

Searched segments in the previous step are expected to have a wide variety of situations and functions of dialogues. Therefore we think discourse segments should be classified according to their contents before analyzing their usefulness as examples for can-do statements. Here, we perform clustering to classify them into several categories by according to discourse contents of the segment clusters. In our clustering, each segment is represented by word embedding using Doc2vec (Le and Mikolov, 2014) as an extension of word2vec (Mikolov et al., 2013) for learning document embeddings. Doc2vec is an unsupervised algorithm that learns fixed-length feature representations from variable-length pieces of texts, such as sentences, paragraphs, and documents.

In this research, we use FSs that consists of multiple pragmatically related words instead of single words for calculating Doc2vec. Document embeddings calculated by Doc2vec will be compressed by Singular Value Decomposition, SVD algorithm. Then compressed vectors are classified by *k*-means method that is a well-known squared errorbased clustering algorithm. To use k-means algorithm requires the number of clusters in the data to be pre-specified. Therefore we set the value of *k* to five in this research. The following figures (Figure 3 to 6) show the results of the clustering for genre D, I, V, and A respectively.

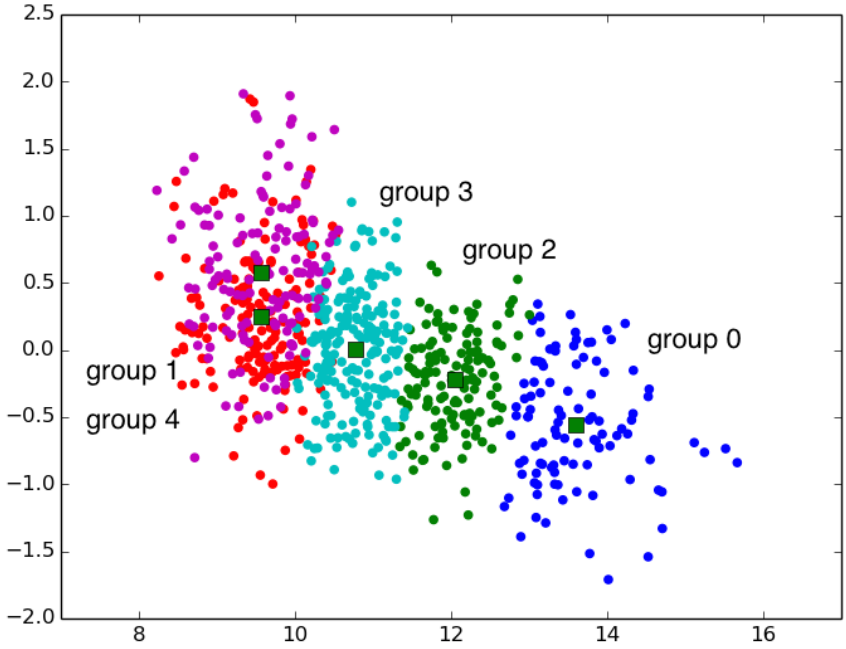

Figure 3: The clustering result of segments on genre D.

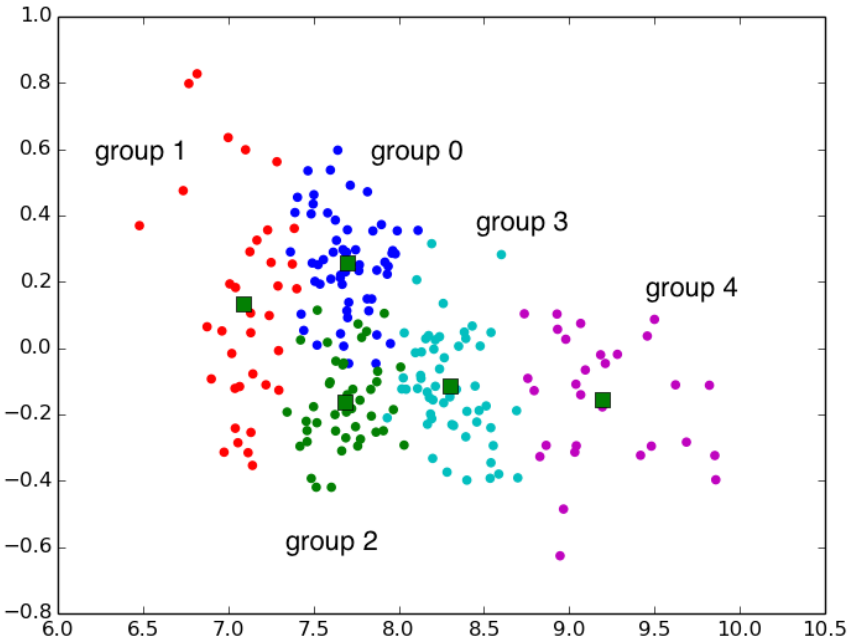

Figure 4: The clustering result of segments on genre I.

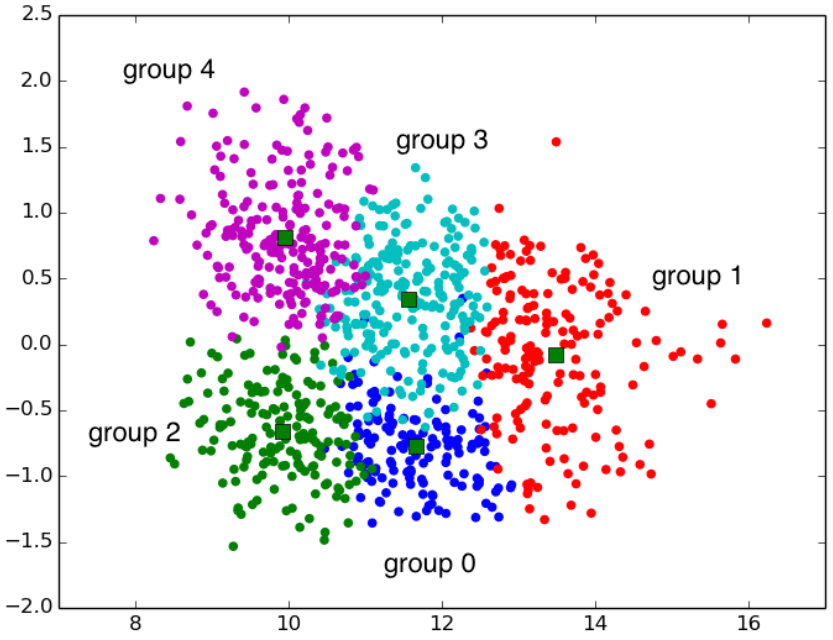

Figure 5: The clustering result of segments on genre V.

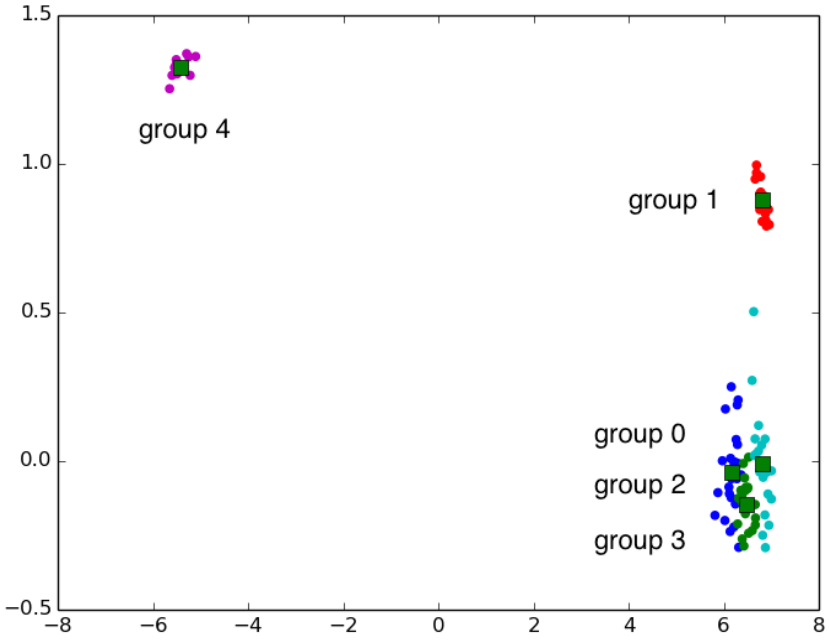

Figure 6: The clustering result of segments on genre A.

Analyzing Usefulness of Discourse Segments

Here we investigate usefulness of clustered segments as a learning material limited to genres D and I. Because there are overlapping segments by rebroadcasting, we remove the same segments from all segments before investigate the usefulness. In order to judge the usefulness of each segment, we use the following two criteria.

(1) a dialogue exists in the segment;
(2) both a situation and a function of a dialogue in the segment are clear;

Figure 3 shows the clustering result of 738 segments on the drama genre labeled D. There are five clusters in Figure 3 from group 0 to group 4. Because there are rebroadcast programs, we remove the same segments from all segments. The numbers of different segments in each group are 70, 112, 125, 167, and 96, respectively. The ratio of overlapping in genre D is 22.9% because the total number of different segments is 569. It is consider that the overlapping ratio in this genre is relatively higher because TV dramas tend to be re-broadcasted frequently in general. The numbers of segments meet the first criterion in each group are 44, 81, 71, 78, and 64 segments, respectively. Table 2 shows results of the classification of situations and functions of dialogues that met the second criterion in genre D.

A situation and a function of a dialogue in the segment	g0	g1	g2	g3	g4
ordering meals at a restaurant	15	19	16	13	30
giving or asking the description of cuisines at a restaurant	5	0	7	6	0
ordering something at a shop	3	7	7	5	0
giving or asking the description of products at a shop	0	0	4	8	2
guide customers to the seats at a shop	0	7	0	8	0
guide guests to the room at a hotel'	0	3	11	0	0
greeting with guests at a reception desk of a hotel	1	0	0	13	5
greeting with customers at a shop	0	0	0	8	4

Table 2: Classification of Situations and Functions of Dialogues in Genre D.
'g*n*' in the table means group numbers.

As shown in Table 2, a restaurant, a shop, and a hotel commonly appeared as situations of dialogues in genre D. Similarly, 'ordering something' 'giving or asking the description of something' and 'greeting with someone' also commonly appeared as functions of dialogues. Especially a dialogue of 'ordering meals at a restaurant' appeared frequently in all clusters. A dialogue of 'ordering something at a shop' also appeared frequently.

Figure 4 shows the clustering result of 222 segments on the information / tabloid style genre labeled I. After removing the same segments, the numbers of different segments in each group are 49, 35, 42, 56, and 28, respectively. The ratio of overlapping in genre I is 5.4% because the total number of different segments is 210. It is consider that the overlapping ratio in this genre is relatively lower because information or tabloid style programs tend to be given as a live broadcast. The numbers of segments meet the first criterion in each group are 22, 26, 28, 44, and 14, respectively. Table 3 shows results of the classification of situations and functions of dialogues that met the second criterion in genre I.

A situation and a function of a dialogue in the segment	g0	g1	g2	g3	g4

ordering meals at a restaurant	4	3	2	4	5
giving or asking the description of cuisines at a restaurant	3	10	11	16	5
ordering something at a shop	1	0	0	1	0
giving or asking the description of products at a shop	3	3	3	6	0
guide customers to the seats at a shop	0	0	0	0	0
guide guests to the room at a hotel'	1	3	0	0	0
greeting with guests at a reception desk of a hotel	2	0	2	3	0
greeting with customers at a shop	0	3	3	0	1

Table 3: Classification of Situations and Functions of Dialogues in Genre I.
'g*n*' in the table means group numbers.

As shown in Table 3, a restaurant, a shop, and a hotel commonly appeared as situations of dialogues in genre I. Similarly, 'ordering something' 'giving or asking the description of something' and 'greeting with someone' also commonly appeared as functions of dialogues. Especially both dialogues of 'ordering meals at a restaurant' and 'giving or asking the description of cuisines at a restaurant' appeared frequently in all clusters. A dialogue of 'giving or asking the description of products at a shop' also appeared frequently.

Conclusions

This paper described the specific results of some analyses regarding our closed caption TV corpus to investigate usefulness of discourse segments in the corpus. We confirmed whether target discourse segments are preferred dialogues to be used as examples of can-do statements for language learning. This paper also described a clustering method by using Doc2vec, SVD, and k-means method, in order to find discourse segments extracted from a closed caption TV corpus using formulaic sequences related to can-do statements.

The results of investigation show that a restaurant, a shop, and a hotel commonly appeared as situations of dialogues in both genres D and I. Expressions of 'ordering something' 'giving or asking the description of something' and 'greeting with someone' also commonly appeared as functions of dialogues. Especially a dialogue of 'ordering meals at a restaurant' appeared frequently in both genres D and I. A dialogues of 'ordering something at a shop' and 'giving or asking the description of products at a shop' also appeared frequently. At least from the limited result of this paper, it seemed that we can find many useful discourse segments that include both a situation and a function of a dialogue clearly. However, since similar types of dialogues overlap in different clusters, our clustering method should be improved in the future.

Acknowledgments

This research was supported by the Grant-in-Aid for Scientific Research (A) (No. 26240051) and (B) (No. 15H02794) of JSPS.

References

Council of Europe (2001). *Common European Framework of Reference for Languages: Learning, Teaching, assessment (CEFR)*. http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Conklin, K. & Schmitt, N. (2012). The Processing of Formulaic Language. Annual Review of Applied Linguistics, vol.32, pp.45-61.

Jiang, N. & Nekrasova, M. T. (2007). The processing of Formulaic Sequences by Second Language Speakers. The Modern Language Journal, vol. 91, iii., pp. 433-445.

Schmitt, N. (Ed.). (2004). Formulaic sequences. Amsterdam: Benjamins.

Vilach, S. R. & Ellis, C. N. (2010). An Academic Formulas List: New Methods in Phraseology Research. Applied Linguistics 31, 4. pp.487-512.

Wray, A.(2002). Formulaic Language and the Lexicon. Cambridge UK: Cambridge University Press.

Wray, A. (2008). Formulaic Language: Pushing the Boundaries. Oxford University Press.

Wray, A. and Perkins, M. R. (2000). The functions of formulaic language: an integrated model. Language & Communication 20, pp. 1-28.

Conklin, K. & Schmitt, N. (2008). Formulaic sequences: Are they processed more quickly than nonformulaic language by native and nonnative speakers? Applied Linguistics 29, 1. pp.72-89.

Le, Q. & Mikolov, T. (2014). Distributed representations of sentences and documents. In Proceedings of the 31st International Conference on Machine Learning (ICML 2014), pages 1188–1196, Beijing, China.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient Estimation of Word Representations in Vector Space. In Proceedings of Workshop at International Conference on Learning Representation (ICLR) 2013, Scottsdale, AZ.

Mochizuki, H. & Shibano, K. (2017a). The Acquisition of a Japanese Practical Formulaic Sequences List from a Closed Caption TV Corpus, In Proceedings of 2017 STEAM Education (Science, Technology, Education, Arts and Math), 6 pages, 2017, Honolulu, USA.

Mochizuki, H. & Shibano, K. (2017b). Searching Discourse Segments for Formulaic Sequences in a Closed Caption TV Corpus for Language Learning. World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Volume 2017, No. 1, (pp. 19-27). Association for the Advancement of Computing in Education (AACE), Vancouver, Canada.