2017 HAWAII UNIVERSITY INTERNATIONAL CONFERENCES
Science, Technology & Engineering, Arts, Mathematics & Education June 8 - 10, 2017
Hawaii Prince Hotel Waikiki, Honolulu, Hawaii

The Acquisition of a Japanese Practical Formulaic Sequences List from a Closed Caption TV Corpus

Mochizuki, Hajime
Institute of Global Studies
Tokyo University of Foreign Studies
Japan

Shibano, KohJi
Research institute for Languages and Cultures of Africa and Asia
Tokyo University of Foreign Studies
Japan

Dr. Hajime Mochizuki
Institute of Global Studies
Tokyo University of Foreign Studies
Japan.

Prof. Kohji Shibano
Research Institute for Languages and Cultures of Africa and Asia
Tokyo University of Foreign Studies
Japan.

The Acquisition of a Japanese Practical Formulaic Sequences List from a Closed Caption TV Corpus

Synopsis:

This paper describes the specific results of some statistical tests on data derived from our CCTV corpus to acquire a practical list of formulaic sequences in the corpus and confirmed the statistical significant distribution of target expressions in a specific genre of TV programs, from the results of chi-squared tests. We also investigate the characteristics of the acquired practical formulaic sequences that express communicative functions. The top 100 practical formulaic sequences for each genre are classified based on their characteristics. This paper also report the results of our analyses of the expression types of acquired formulaic sequences as key phrases for smooth communication.

The Acquisition of a Japanese Practical Formulaic Sequences List from a Closed Caption TV Corpus

Hajime Mochizuki
Institute of Global Studies
Tokyo University of Foreign Studies
Tokyo, Japan

Kohji Shibano
Research Institute for
Languages and Cultures of Asia and Africa
Tokyo University of Foreign Studies
Tokyo, Japan

Abstract: This paper describes the specific results of some statistical tests on data derived from our CCTV corpus to acquire a practical list of formulaic sequences in the corpus and confirmed the statistical significant distribution of target expressions in a specific genre of TV programs, from the results of chi-squared tests. We also investigate the characteristics of the acquired practical formulaic sequences that express communicative functions. The top 100 practical formulaic sequences for each genre are classified based on their characteristics. This paper also report the results of our analyses of the expression types of acquired formulaic sequences as key phrases for smooth communication.

Introduction

A Corpus is a large and structured set of texts. The use of corpora has become the most important resource for research studies and applications related to natural language, and a variety of research studies and applications for corpus-based computational linguistics, knowledge engineering, and language education has been reported in recent years (Flowerdew, 2011; Newman et al., 2011). In our project, we have been collecting a large-scale spoken language corpus from closed caption TV (CCTV) data transmitted through digital terrestrial broadcasting since December 2012 (Mochizuki and Shibano, 2014). The size of our corpus has reached over 166,000 TV programs with more than 65 million sentences and 655 million words as of February 2016.

We anticipate the existence of sequences of words that occur frequently in the corpus, such as collocations, idioms, and greeting expressions. These sequences of words are referred as formulaic sequences (FSs) (Wray, 2002). Recently, it is generally accepted that FSs serve an important function in discourse and are widespread in language (Conklin and Schmitt, 2008). In our recent research, to extract the FSs, we calculated the sequences of n words (*n*-grams) in the corpus, where *n* value ranged from one to nine. We calculated a total of 3,544,847,579 *n*-grams. After a sorting and merging process, we acquired over 33 million significant *n*-grams as FSs candidates (Mochizuki and Shibano, 2016). However, an extensive list of FSs is hard to use as a practical learning material. For that reason, investigating FSs more closely and creating a list of useful FSs are necessary as our future works.

Therefore, in this work, we attempt to acquire a list of practical formulaic sequences from the substantial raw FSs candidates. We focus on keeping a balance between frequency and string-length: From the viewpoint of its benefit as a recognition unit, a longer string of FS is preferred than a shorter one whereas, from the view point of its benefit of high repeat usability and coverage, a more frequent FS is preferred. In this research, we investigate FSs with more than nine characters and more than nine occurrences in the corpus. From our preliminary investigation using these parameters, we expected to acquire an FS that can be considered as a prefabricated component. We adopt chi-squared tets to comfirm whether the target expressions have a statistical significant distribution in a specific genre of TV programs. From the results of chi-squared tests, a final list of practical FSs will consists of statistically significant FSs. In this paper, we also investigate the characteristics of the acquired practical formulaic sequences that express communicative functions. The top 100 practical formulaic sequences for each genre are classified based on their characteristics. The results of our analyses of the expression types of acquired formulaic sequences as a key phrase for smooth communication are also reported.

Extracting Formulaic Sequences from Closed Caption TV Corpus

In our previous research, we extarcted FSs from our closed caption TV corpus (Mochizuki and Shibano, 2016) and defined the FS as a continuous sequece of words in a sentence, though we recognized that FS can also includes discountinuous word sequences. By narrowing down the definition of FS, we extracted significant word-level *n*-grams form the CCTV corpus as FSs. A word-level *n*-gram means a sequence of *n* words that appears consecutively in the text. The number of *n*-grams for the value *n* from a sentence with *N* words is *N*-*n*+*1*.

In our previous research, we calculated the *n*-grams using our ruby program that changes the value of *n* from one to nine. We, therefore, found a total of $\sum_{n=1}^{9} N - n + 1$ *n*-grams from *N* word sentences. For example, the total number of n-grams from a 10-word sentence is $\sum_{n=1}^{9} 10 - n + 1$, which equals 10+9+8+7+6+5+4+3+2; thus, 54 *n*-grams were extracted. Our CCTV corpus has approximately 65 million sentences with an average of 10 words per sentence; this means that the total number of words is 655 million. Therefore, the total number of *n*-grams calculated from the CCTV corpus is approximately 65 million times 54, that is, 3,510 million *n*-grams. In fact, 3,544,847,579 *n*-grams were calculated from 168,175 text files for 39 months. In the next step, all *n*-grams were sorted and merged. As the result of the sort-and-marge process, 1,178,755,922 types of *n*-grams were extracted from the CCTV corpus.

Finally, significant *n*-grams were selected from over 1 billion types of *n*-grams. In our previous research, we defined significant *n*-grams as the longest n-gram that appears in various identical sentences. According to the definition, the extraction of significant *n*-grams resulted in the acquisition of 58,659,190 n-gram types from 1,178,755,922 *n*-gram types. Further, 32,590,395 *n*-grams among the 58,659,190 candidates appeared more than once in the CCTV corpus, and 26,068,795 *n*-gram types appeared only once.

Acquisition of a Practical Formulaic Sequences List

As mentioned in the previous section, we created a formulaic sequences list containing about 32 million types that appeared more than once in the CCTV corpus. Such an extensive list will be expected to contain useful expressions, but on the other hand a lot of meaningless or unfinished strings will also be included. It is hard to manage a mixture of wheat and chaff; therefore, our aim in this research is to aquire a practical FSs list that we can treat easily from the sizable list.

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). Among them, we use the three genres namely “Drama (D),” “Variety (V),” and “Information and Tabloid Style (I)” in this research because their sizes are relatively large and with samples of many dialogues.

To acquire a practical FSs list, we apply the following three procedures; (1) extract practical FS candidates from the extensive FS list; (2) examine each FS in the list whether the distribution of the FS in one genre differs statistically from the distribution the FS among other genres in the CCTV corpus; (3) acquire a practical FS list for each genre according to the results in the procedure (2).

In the first step, as a criterion for practicality, we focus to keep the balance between frequency and string-length of the FS. We prefer a longer string of FS than a shorter one from the viewpoint of the unit's immediate recognizability. On the other hand, from the viewpoint of the benefit of high repeat usability and wide coverage for the whole corpus, we prefer more frequent FSs. In this research, we first count the character length of each FS and extract any FS with its length exceeding the threshold. We investigate FSs with more than nine characters and more than nine occurrences in the corpus. The number of FSs that consists of more than nine characters is 54,525 in genres D, V, and I. Among them, the number of FSs with frequencies more than nine is 6,734. Therefore 6,734 FSs are considered as candidates of practical FSs in our research.

In the second step, we conduct chi-squared tests for independence of each FS in the list for each genre among genres D, V, and I. We use a chi-squared test to determine whether the distribution of an FS in genre D, V, or I differed significantly from the distribution among other genres. We examine the results of the chi-squared tests between "D" and "V," "D" and "I," and "V and I." The numbers of FSs examined are 4,873, 5,040, and 5,610, respectively. From the results of the chi-squared tests, we found 288, 253, and 645 statistically significant FSs, respectively. Here, when the *p*-value for a chi-squared test is less than 0.05, we regard the FS as statistically siginificant.

In the third step, we create a practical FS list for each genre. We assign each statistically significant FS to at least one genre according to the distribution bias. For example, the result of the chi-squared test in an FS "こんなことになるなんて (I don't want to be like this)" between genre D and V is 188.27, with *p*-value is less than 0.001, a biased distribution in genre D. Therefore, we regard this FS as a practical FS for genre D. The numbers of types of practical FSs for genre D are 174 from genres "D" and "V," and 81 from genres "D" and "I."

The numbers of types of practical FSs for genre V are 114 from genres "D" and "V," and 79 from genres "V" and "I." The numbers of types of practical FSs for genre I are 172 from genres "D" and "I," and 566 from genres "V" and "I." Examples of the acquired FSs for genres D, V, and I are listed in Table 1, 2, and 3, respectively.

FS	chi-squared	*p*-value
申し訳ありませんでした。(I'm sorry)	527.20	> 0.001
本当に申し訳ありませんでした (I'm very sory)	112.65	> 0.001
おっしゃってください。(Please tell me …)	80.97	> 0.001
こんなことになるなんて (I can't believe that actually happened.)	79.31	> 0.001
いい加減にしてください (Please act properly.)	47.35	> 0.001
行ってらっしゃいませ (Have a nice trip)	42.43	> 0.001
お願いがあるんですけど (Would you do me a favor?)	18.44	> 0.001

Table 1: Examples of practical FSs for genre D.

FS	chi-squared	*p*-value
ありがとうございました。 (Thank you very much)	1247.20d	> 0.001
ご紹介したいと思います (I would like to introduce to you …)	270.58	> 0.001
そんなことないですよ。(That is not true)	145.56	> 0.001
ちょっと行ってみましょ (Let's go)	100.60	> 0.001
かもしれないんですけど (It might something, but …)	78.91	> 0.001
』で検索してください。 (look for … on a search engine)	70.36	> 0.001
歌っていただきましょう (Sing us a song, please)	47.34	> 0.001

Table 2: Examples of practical FSs for genre V.

FS	chi-squared	*p*-value
よろしくお願いします。 (Thank you for …)	12519.00	> 0.001
ありがとうございました。 (Thank you)	3924.80	> 0.001
ご紹介したいと思います (I would like to introduce to you …)	1214.00	> 0.001
とおっしゃっていました (You said that …)	762.95	> 0.001
いただければと思います。 (I hope you *DO*)	724.51	> 0.001
召し上がってください。 (Please eat / Please help …)	721.03	> 0.001
かということなんですが (Although it is that to/possible to)	468.20	> 0.001

Table 3: Examples of practical FSs for genre I.

The Features of the Practical Formulaic Sequences List

We investigated the features of the top 100 practical FSs in the order of their chi-squared values for genres D, V, and I, respectively. We classified the FSs with reference to Ek and Trim, 1990. Table 4 shows results of the classification of expression types.

Expression type of FS	D	V	I
Greeting expressions	3	14	1
Expressing gratitude	12	28	8
Expressing wants/desires	13	14	16
Expressing congratulation	0	5	1
Expressing thoughts or guesses	2	1	6
Expressing intentions	7	11	21
Long or complex nouns	5	9	6
Person’s name	0	3	4
A, but B	3	5	13
Offering an apology	17	0	0
Expressing disapproval	17	0	0
Expressing condition	3	0	0
Paying attention, expressing an angry	7	0	0
Offering or inviting someone to do something	0	0	3
Confirming something	0	0	3
Other things	11	10	18
Total number	100	100	100

Table 4: Classification of Expression Types of Practical FSs

In the drama genre labeled “D,” the most freqent expressions are “offering an apology” and “expressing disapproval” followed by “expressing wants or desires,” and “expressing gratitude.” It is natural for the many expressions related to one's emotions because the orignal source of this genre consists of scripts from dramas. In the variety genre labeled “V,” the most frequent expression is “expressing gratitude” followed by “greeting expressions,” “expressing wants or desires,” “expressing intentions,” and “long or complex nouns.” We assumed that the prevalence of the many descriptions for greetings and expressing gratitudes is attributed to te fact that this genre of TV programs had many guests having many discussions and dialogues. In the information and tabloid style genre labeled “I,” the most frequent expression is “expressing intentions” followed by “expressing wants or desires,” “A, but B,” and “expressing gratitude.” There are relatively many expressions stating opinions or attitudes, because the TV programs of this genre include many topics or news with discussions and dialogues by the casts. In all three genres, “expressing gratitude,” “expressing wants or desires,” and “expressing intentions” occur frequently. Their occurrences are 12, 13, 7, and 8, 16, 21, and 28, 14, 11 times, respectively.

Conclusions

In this paper, we reported the analysis results of the expression types of acquired formulaic sequences as key phrases for smooth communication. We described the specific results of some statistical tests performed on significant FSs extracted from the CCTV corpus to acquire a practical list of formulaic sequences in the corpus. We confirmed whether the target expressions have a statistical significant distribution in a specific genre of TV programs from the results of chi-squared tests. We also investigated characteristics of acquired practical formulaic sequences that express communicative functions, and classified them. The results showed the following conclusions: (1) There were many expressions related to one's emotions in the drama genre: (2) there were many descriptions of greetings and expressions of gratitude in the variety genre: and (3) there were relatively many expressions stating opinions or attitudes in the information genre.

Acknowledgments

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

References

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.

Ek, J. A. V. and Trim, J. L. M. (1990). Threshold 1990. Cambridge, Cambridge University Press.

Flowerdew, L. (2011). Corpora and Language Education. Palgrave Macmillan.

Mochizuki, H. & Shibano, K. (2014). Building Very Large Corpus Containing Useful Rich Materials for Language Learning from Closed Caption TV. World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Volume 2014, No. 1, (pp. 1381-1389). Association for the Advancement of Computing in Education (AACE), New Orleans.

Mochizuki, H. & Shibano, K. (2015a). Development of a Closed Caption TV Corpus Retrieval System to Seek Video Scenes Containing Useful Expressions for Language Learning. The EdMedia World Conference on Educational Media and Technology, (to appear, nine pages). Association for the Advancement of Computing in Education (AACE), Montreal, Canada.

Mochizuki, H. & Shibano, K. (2015b). Detecting Topics Popular in the Recent Past from a Closed Caption TV Corpus as a Categorized Chronicle data, the 7th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (KMIS) 2015, *to appear*, Lisbon, Portugal.

Newman, J., Baayen, H. & Rice, S. (2011). Corpus-based Studies in Language Use, Language Learning, and Language Documentation. (Language and Computers Studies in Practical Linguistics), Rodopi.

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