Building a Topic-Oriented Corpus and Its Application for Language Teaching

Naoki Nakamata
Osaka University
n.nakamata.ciee@osaka-u.ac.jp

Abstract

Topics play an important role in teaching methods, such as Content-Language Integrated Learning, Content Based Instruction, and Task-Based Language Teaching. To study the influence of topics on vocabulary, grammar, and discourse strategies in conversations, we constructed the *Japanese Topic-Oriented Conversation Corpus,* (J-TOCC), the main feature of which is that topics are fixed. University students were asked to engage in conversations on 15 topics, and each conversation was recorded for precisely 5 minutes. Eleven topics are related to daily life, and four topics re related to society. In total, 120 pairs participated; therefore, 10 hours of conversations were recorded for each topic. J-TOCC can clarify topic-specific words as well as topic-specific grammar and collocations in Japanese.

1 Background

In recent years, language education has emphasized "Can-Do" as an achievement goal for learners. "Can-Do" is connected to specific situations and topics. The CEFR descriptors consider whether learners can perform tasks with a range of topics. For example, in the overall comprehension section, item A1 includes the description "Can recognize concrete information (e.g. places and times) on familiar topics encountered in everyday life, provided it is delivered slowly and clearly," while B2 includes the description "Can understand standard language or a familiar variety, live or broadcast, on both familiar and unfamiliar topics normally encountered in personal, social, academic or vocational life" (Council of Europe 2020).

References to topics were adapted from the European context to the Japanese language education. The Japan Foundation developed a website, "Minna no Can-Do" (Everyone's Can-Do.) so that course organizers, teachers, and learners can set their own can-dos for each class or project, since the original descriptors are abstract enough to adapt to a wide range of learning styles.[1] To make concrete can-dos, this website attached topics such as "me and family" or "house and living environment" for each can-do. The Japanese Language Proficiency Test (JLPT), a large-scale Japanese language test, also adapts the terms "daily topics" or "various topics" to each level descriptor. (The Japan Foundation & Japan Educational Exchanges and Services 2020)

Furthermore, in recent years, teaching methods such as Content Based Instruction (CBI), Content-Language Integrated Learning (CLIL), and Task-Based Language Teaching (TBLT) have attracted attention in language education (Brinton et al. 2003, Coyle et al. 2010, Ellis 2003). Research into the development of teaching materials is being actively conducted. All these teaching methods focus on language as well as content, that is, topics.

At the same time, many classes in Japanese language education still use conventional sentence-structure stacked syllabi. However, even in classes using such a traditional approach, there is always a conversation activity, and the conversation activities always involve certain topics.

If we can clarify what vocabulary, grammar, and discourse strategy are required to perform the task for each topic, which topics can be performed with less linguistic knowledge, and which topics require more linguistic knowledge, it will be useful for language education, regardless of the kind of teaching method adopted.

[1] https://jfstandard.jp/cando/top/ja/render.do

2 Language Forms and Topics

It seems natural that the appearance of nouns or verbs is affected by topic. In the field of Japanese language education, a remarkable contribution is Yamauchi's topic-oriented vocabulary list (Yamauchi (ed.) 2013). This list includes 100 topics, and each topic contains plenty of verbs, nouns, and adjectives. However, the mapping of topics to words was mainly based on intuition and a corpus was partly used to build this list. Most words were manually divided into 100 topics after being copied from the list for the old Japanese Language Proficiency Test[2]. There is no guarantee that the word is really used frequently in conversation on each topic. Furthermore, Yamauchi (ed.) (2013) claims that while substantial words such as nouns, verbs and adjectives depend on topic, functional words does not depend on topics.

Though this claim is true apparently, Nakamata (2019) showed that some function words do depend on topic through a corpus survey. For example, potential forms appeared significantly more frequently in the "Eating" topic than in others. Expressions related to time such as tense marker *ta*, progressive marker *te-iru*, and time adverbs including those meaning "yesterday" and "in the last year" were extracted as topic-specific words in the "pop culture" topic. This shows an insight that grammar depends on topic, though the corpus used was one of conversation between native-Japanese-speakers and Japanese-learners. This characteristic could affect the results. Other problems are that the corpus size was limited and the topics were not strictly controlled.

To confirm if grammar really depends on topic, a new corpus containing conversations of native speakers on various topics should be built.

3 Design of J-TOCC

To examine how topic affects vocabulary, grammar and discourse strategy, the Japanese Topic-Oriented Conversation Corpus (J-TOCC), is designed. Its most important feature is that the same pairs of collaborators speak on 15 topics for 5 minutes each. Conditions except for topics are strictly controlled. The corpus is provided as a text file with an information table for participants on the author's website.[3]

All procedures described below were approved for research ethics review at the author's university. In addition, research ethics review committee approval was obtained at the university where the recordings were made, if necessary.

3.1 Topic Selection

The starting point for topic selection was the 100 topics listed in Yamauchi (ed.) (2013), and a target of around 15 topics was set based on the size of the corpus. 15 topics are enough when we developed a learning material since normal courses in a university contains 15 classes. Yamauchi (ed.) (2013) labeled familiarity and abstractness for each level. Based on these labels, 52 topics with high or middle familiarity were chosen first.

Subsequently, we selected 11 "daily topics" from the following 4 perspectives: a. Familiar to university students. b. Appropriate for beginner-level learners of Japanese c. Not too much information with participants' privacy appeared. d. Not too close to other topics. The exclusion criterion was set on the following basis: for example, although the topic of "music" was familiar to the university students, it was not adopted because, when the preliminary survey was conducted in Western Japan, one student could not be understood in the transcript.

No.	Topic	Group
1	Eating	Daily Topics
2	Fashion (Clothes)	
3	Travel	
4	Sports	
5	Manga and Games	
6	Housework	
7	School	
8	Smartphone	
9	Part-Time Job	
10	Animals	
11	Weather	
12	Dreams and Life Plan	Topics Related to Society
13	Manners (on Public Transport)	
14	Living Environment	
15	Declining Birthrate and Aging Population	

Table 1: Fifteen topics of J-TOCC

[2] The JLPT changed in 2010, after which there was an old and a new JLPT

[3] http://nakamata.info/database/

Furthermore, if lyrics were included in the speech, copyrights would be required, so they were excluded.

Additionally, the topic "home appliances and machines" in Yamauchi (ed.) (2013) was changed to "smartphones," which are the most familiar electrical appliances for university students.

Though these 11 topics are all "daily topics," to construct a corpus that handles "various topics" we selected 4 additional topics that are somewhat complex and socially oriented. These four topics include ones with low familiarity.

The 15 topics finally selected are shown in Table 1.

3.2 Participants

Participants were limited to close friends among university students in their 20s, to help them talk about various topics comfortably. All participants were undergraduate students, since graduate students would be easily identified if their research themes were included in the conversation.

The number of pairs was balanced between east and west Japan, and the same number of pairs was recorded in each of the three groups of "male–female," "male–male," and "female–female," respectively; 20 pairs (40 persons) participated per group, with no overlap of speakers. Participants were in 120 pairs, with 240 speakers.

3.3 Recording

The recordings took place between 2018 and 2019. Once a recording date was set, the participants were asked to come in pairs to a designated location and first receive an explanation from the recording artist. After the explanation, questions were checked, and recording began when there were no more questions.

First, a topic board was shown and the participants were given 30 seconds to think. This topic board contained detailed instructions along with one of the topics shown in Table 1. As an example, "01. Eating" has "for example, favorite food and eating out. *except for cooking." Then, after switching on the audio-recorder, the recording person first said, "Pair number, W-101. Topic, Eating. Start." Immediately after that, the participants started the conversation, and the recording person moved to a position out of sight of the participants and waited.

The time for the conversation was five minutes; this was indicated to the participants, but they were asked to keep on talking if they neared the end. After five minutes, the recorder was first switched off silently by the recording person, and then the conversation was stopped. Therefore, a conversation was not necessarily five minutes long, but each datum is exactly five minute long, since we transcribed exactly five minutes from the beginning of the conversation. After that, the next topic board was shown. We took breaks as needed, during which communication, eating, and drinking were allowed. Furthermore, the use of smartphones was allowed even during the recording of the conversation, and some pairs actually had a conversation while looking at their smartphones.

The order of topics was random for each pair; the order was counterbalanced because it was expected that the first topic would be awkward and less said about it than the later ones.

After the recording, the participants were asked to fill out a "face sheet" (see below) and an Informed Consent Form. In the consent form, we again confirmed that it would be acceptable to release the recorded files, and we confirmed with them the phrases that they wished to mask. The terms that were requested here were masked regardless of the masking criteria described in section 3.5.

3.4 Face Sheet and Topic Knowledge

The face sheet contains the following information: name, sex, place of language formation (the prefecture where the participant lived between the ages of 6 and 12; multiple answers are possible), and level of topic knowledge. The level of topic knowledge is defined as "how well you know each topic or how confidently you talked on each topic," and participants were asked to grade it with five levels. We wanted to know the "degree of detail," but the scale of detail for topics such as "11. Weather," would not be appropriate unless the speaker had studied to become (e.g.) a weather forecaster or climate scientist, so we used the definition above.

The corpus includes data other than the speaker's name as information on the speaker. By using the degree of topic knowledge, we can compare those who are familiar with a topic with those who are not.

Topic	Token	Type	TTR	Topic Knowledge
01. Eating	145,361	6,338	0.0587	3.78
02. Fashion (Clothes)	148,909	6,111	0.0543	2.87
03. Travel	147,386	6,660	0.0605	3.50
04. Sports	148,722	6,810	0.0605	3.35
05. Manga and Games	151,454	7,106	0.0631	3.70
06. Housework	147,950	6,141	0.0549	3.16
07. School	145,198	6,794	0.0616	3.71
08. Smartphone	144,745	6,145	0.0563	3.48
09. Part-Time Job	147,120	6,791	0.0611	3.70
10. Animals	148,636	6,455	0.0577	3.44
11. Weather	146,245	6,242	0.0564	2.91
12. Dreams and Life Design	141,101	6,034	0.0562	3.20
13. Manners (on Public Transport)	148,209	5,586	0.0490	3.27
14. Living Environment	141,934	5,578	0.0517	3.27
15. Declining Birthrate and Aging Population	136,718	5,950	0.0563	2.81
Sum	2,189,688	42,756	0.0258	3.34

Table 2: Word Size of Each Topic

3.5 Transcription

The recorded data were transcribed, and only the text files are published as J-TOCC. A company was hired to transcribe the recording; then, part-time students from universities different from the recording location were hired to check the transcription and mask personal information.

As regards masking, the main policy was to mask information that could lead to the identification of the participants or the university where the research was conducted. Names of famous people or famous places that everyone visits were excluded from the masking. All masking is indicated by square brackets, with the word "-name" inside the brackets.

For personal names, a distinction is made between [personal name, 1st person], [personal name, 2nd person], and [personal name, 3rd person]. In principle, the names of places are published as they are down to the level of the city, and the names of places lower than that are masked. Place names below the level of city which are not considered to be likely to lead to the identification of the speaker, such as the names of travel destinations, are left as they are. An example of masking is shown below.

E-213-1F：受ける．い，【人名：２人称】発なのそれ．

E-213-2M：そうそう，そうそう．で，あの【店名】行って飲んで．（E-213-1F：へー）で，昨日はあの，【店名】で【人名：３人称】と食べ飲み放して．

E-213-1F：【店名】って食べ飲み放あんの？

Translation

E-213-1F: Funny. Is, is that from **[person's name: 2nd person]**.

E-213-2M: Yes, yes, yes, yes. And then we went to that **[restaurant's name]** and had a drink. （E-213-1F: Heh.） And yesterday, I had all you can eat and drink with **[person's name: 3rd person]** at **[restaurant's name]**.

E-213-1F: Does **[restaurant's name]** have all you can eat and drink?

4 Analysis

4.1 Word Size

The final size of J-TOCC is 10 hours per topic, for a total of 150 hours, since 120 pairs spoke on 15 topics for 5 minutes each. The most important feature of J-TOCC compared to other corpora is that the same speakers talk about 15 different topics. Since all other conditions except for the topic are controlled, the effect of the topic can be isolated.

Table 2 shows the word size of each topic. In the column of token frequency there seems to be a slight gap between the maximum value of "05. Manga and Games" and "15. Declining Birthrate and Aging Population" This is because "05. Manga and Games" includes several parentheses for titles. If these are excluded, the difference between the maximum and minimum is as small

as 8,431, indicating that each topic has around 110,000 words.

Whereas, type frequency differs a little more depending on the topic. As a trend, all the topics related to society (12–15) have fewer type frequency than the daily topics (1–11.) Type–Token Ratio (TTR) also shows the same trend.

The rightmost column shows the average of the topic knowledge scores of the participants. The topic that the Japanese university students were most familiar with and confident in talking about was "05. Manga and Games." In contrast, some topics, such as "02. Fashion" and "11. weather," had low values. The correlation coefficient between the level of topic knowledge and the token frequency was .132, which was almost negligible, while the correlation coefficient with the type frequency was .600, which was moderate.

Topic	Freq.	Topic	Freq.
01. Eating	1,321	09. Part	1,841
02. Fashion	1,869	10. Animals	1,927
03. Travel	1,187	11. Weather	1,857
04. Sports	2,065	12. Dreams	1,431
05. Manga	2,573	13. Manner	2,120
06. Housework	1,550	14. Living	1,088
07. School	1,824	15. Future	1,466
08. Smartphone	2,291		

Table 3: Frequency of progressive marker *teru* on each topic.

The level of topic knowledge was self-reported; it was considered that familiarity with a topic is reflected not in token frequency but in type frequency.

4.2 Does Grammar Depend on Topic?

One of the main purposes of building J-TOCC was to confirm if function words are affected by topic. As a case study, token frequency of progressive marker *teru* is surveyed. The original form of this marker is *te-iru*; however, it is well-known that this form shrinks to *teru* in conversation. The survey indicated that 98% of markers are shrunken forms. Table 3 shows the results.

The number of occurrences of *teru* varies widely by topic, with a difference of about 2.5 times between the minimum and maximum values. *Teru* occurred most frequently in "05. Manga and Games," followed by "08. Smartphone" Nakamata (2019) similarly reported that it occurred frequently in the topic of "pop culture," which included Mangas and TV dramas. Considering that more people watch videos on smartphones and other devices than before, these results from J-TOCC can be said to verify Nakamata's (2019) claim. Of course, this is only a simple survey of one linguistic form, but it indicates the possibility of clarifying the relationships among various grammatical forms and topics.

4.3 Vocabulary List

Based on J-TOCC, we built two vocabulary lists and published them. One list contains raw frequency and Log Likelihood Ratio (LLR),

No.	High LLR	High UR
1	*taberu* (eat)	*iu* (say)
2	*kuu* (eat)	*taberu* (eat)
3	*nomu* (drink)	*omou* (think)
4	*tanomu* (order)	*wakaru* (understand)
5	*futoru* (get fat)	*deru* (go out)
6	*tsukuru* (cook)	*chigau* (different)
7	*yaku* (bake)	*tsukuru* (cook)
8	*suku* (get hungry)	*kuu* (eat)
9	*narabu* (queue)	*hairu* (be in)
10	*aburu* (grill)	*shiru* (know)

Table 4: High LLR and UR verbs in "Eating"

No.	High LLR	High UR
1	*raamen*	*koto* (event)
2	*sushi*	*yatsu* (thing)
3	*niku* (meat)	*hito* (person)
4	*yakiniku*	*kanji* (feeling)
5	*curry*	*hou* (one)
6	*gohan*	*raamen*
7	*omuraisu*	*gohan*
8	*cheese*	*ie* (home)
9	*aji* (taste)	*hontoo* (really)
10	*tabemono* (food)	*mono* (thing)

Table 5: High LLR and UR nouns in "Eating"

No.	High LLR	High UR
1	*oishii* (tasty)	*suki* (like)
2	*suki* (like)	*oishii* (tasty)
3	*umai* (good)	*sugoi* (great)
4	*amai* (sweet)	*ooi* (much)
5	*daishuki* (love)	*tashika* (certainly)
6	*karai* (hot)	*maji* (really)
7	*aburakkoi* (oily)	*umai* (good)
8	*yasui* (cheap)	*yabai* (bad)
9	*shoppai* (salty)	*sonna* (like that)
10	*koi* (strong flavor)	*iya* (dislike)

Table 6: High LLR and UR adjectives in "Eating"

which indicates which words are specific to one topic as compared to the other 14 topics. The other list contains how many times each participant used a specific word in each topic. User Ratio (UR) is easily calculated from the latter list, which indicates what percentage of participants used the word, given a topic.

Tables 4-6 show verbs, nouns, and adjectives with high LLR and verbs, nouns, and adjectives with high UR on the topic of "Eating."[4]

5 Application to Education

As an application to education, three vocabulary builders have been developed based on the vocabulary list of J-TOCC. Some vocabulary builders have utilized corpora to select words, whereas our builder utilized the corpus to arrange the selected words in topic groups. First, the topic in which each word shows the highest LLR is determined from the vocabulary list mentioned in 4.3. Topics are automatically attached for more than 80% of words for N3 (intermediate) level, and manually attached for the remaining 20% of words. After that, writers composed a short story with 3-8 words that belonged to the target topic. Words related to one topic were in the same chapter in most builders published previously, and therefore, utilization of the corpus could show interesting results. For example, *seikai* (right answer), which is used in the "School" topic, is also a specific word in the "Fashion (Clothes)" in J-TOCC. This word is used in the context where speakers wonder what kind of clothes are adequate for time, place and occasion, which may reflect the Japanese cultural convention that people pursue the "right answer" on what clothes should be worn, especially in formal situations under strong peer pressure.

Utilization of J-TOCC for education is not limited to this, but open for any type of classroom and learning. We have published the corpus text and vocabulary list; however, these data files are not easy to use for the preparation of everyday class materials. Therefore, we are now building a website where users can check the words, collocations, and grammatical features which tend to be used on a specific topic, or, conversely, topics in which a specific word is frequently used. The website will also have a search system by which users can search J-TOCC directly with morphological information (e.g., POS).

[4] Terms without translation in Table 5 are names of food in Japanese.

6 Conclusion

This paper has presented the J-TOCC's design and the results of analysis based on it. Creating a topic-oriented conversation corpus enables us to show that topics affect various aspects of linguistic forms, including grammar. This study has clarified only one part of the relationship between topics and language form, however; further research will provide more useful information for language education on relationships among topics and vocabulary, grammar, and discourse strategy. Although this study was conducted on Japanese language, similar survey can be conducted for other languages as well.

Furthermore, since J-TOCC contains the same number of samples for each gender combination and recording location, it could be used for studies such as a comparison of speaking styles between east and west Japan, or a comparison of male and female speaking styles. When conducting such studies, it will be possible to eliminate the influence of the topic and compare the attributes of the speakers by surveying only a specific topic. This is difficult with conventional large-scale corpora, and the key characteristics of J-TOCC, that the topic of each conversation is specifically defined, makes it possible.

As a short report, there is a discussion in Japan that people in western Japan use onomatopoeia more frequently than those in eastern Japan. The results of the survey using this corpus showed that onomatopoeia that begins with a voiced sound and has a lively image in Japanese is used more in western Japan, although there is no significant difference in onomatopoeia as a whole (Ota 2023).

For future development, a website is under construction that will allow users to retrieve information from the corpus at ease, as mentioned Section 5. Furthermore, a limitation of J-TOCC is that it focuses on everyday topics. For intermediate and advanced learners, a corpus on more difficult social issues would be beneficial. Therefore, we are building a corpus on each of the Sustainable Development Goals (SDGs) topics.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP18H00676 and JP22H00686.

We also express our thanks to all participants who helped us record the conversation.

References

Brinton, M., Snow, M. A., and Wesche, M. B. 2003. *Content-Based Second Language Instruction (Michigan Classics ed.)*. University of Michigan Press, Ann Arbor.

Council of Europe. 2020, *Common European Framework of Reference for Languages: Learning, teaching, assessment – Companion Volume*. Council of Europe Publishing, Strasbourg, www.coe.int/lang-cefr

Coyle, D., Hood, P., and Marsh, D. 2010. *CLIL: Content and Language Integrated Learning*. Cambridge University Press, Cambridge.

Ellis, R. 2003. *Task-Based Language Learning and Teaching*. Cambridge University Press, Cambridge.

Nakamata, N. 2019. Vocabulary depends on topic, and so does grammar. *Journal of Japanese Linguistics*, *35*(2), 213-234. https://doi.org/10.1515/jjl-2019-2011

Ota, Y. 2023. Daigakusee no Zatsudan niokeru Onomatope no Siyoo-keekoo: Chiiki-sa, Danjo-sa ni Tyakumokushite, [Onomatopoeia Usage Trends in University Students' Chatting: Focusing on Regional and Gender Differences] In Nakamata (ed.) Wadai-betsu Koopasu ga Hiraku Nihongo, Nihongo-kyooiku-kenkyuu [Japanese Language and Japanese Language Education Research Pioneered by Topic-oriented Corpus]. Hitsuji Shobo, Tokyo.

The Japan Foundation & Japan Educational Exchanges and Services. 2020. Nihongonooryokushiken-gookakusha to senmonka no hyooka niyoru reberu-betsu Can-do risuto: watashi ga nihon-go de dekiru koto. [JLPT Can-do Self-Evaluation List] https://www.jlpt.jp/about/pdf/cdslist_all_2020.pdf

Yamauchi, H. (ed.). 2013. *Jissen Nihongokyooiku Sutandaado* [Jissen's standard of Japanese language education]. Hitsuji Shobo, Tokyo.