Expansion of Japanese Topic-Oriented Conversation Corpus: A Corpus of University Students' Conversation on the SDGs

Naoki NAKAMATA, Tzuhsuan MA
The University of Osaka, Kyoto University of Foreign Studies
n.nakamata.ciee@osaka-u.ac.jp, kennji.ma@gmail.com

Abstract

We expanded a topic-oriented corpus of Japanese conversation. In the first phase of the project, fifteen topics related to everyday life were selected. In the second phase, sixteen topics based on the SDGs were added. For each topic, the same pair of participants engaged in a five-minute conversation, making it possible to examine differences across topics while controlling for speaker variation. Combined with the first phase, the corpus now contains approximately 310 hours of conversation data and around four million words. In addition to transcription data, the corpus includes morphological analyses based on both short-unit and long-unit segmentation, together with Python code for searching. Metadata on participants are also provided, including self-reported familiarity with each topic and estimated vocabulary size. The corpus reveals several notable tendencies. For example, the topics with the highest average familiarity ratings were "education" and "gender." In contrast, science-related topics showed very low familiarity ratings and gender differences. In the first phase of the project, topic familiarity showed positive correlations with both token frequency and vocabulary diversity. We also provide lists of topic-specific keywords calculated using log-likelihood ratios (LLR). This corpus provides valuable foundational data for research on language, discourse, and education.

Keywords: language education, the SDGs, corpus design, specific words

1. Background

In response to the growing emphasis on content-oriented approaches in Japanese language education, we developed the Japanese Topic-Oriented Conversation Corpus (J-TOCC) (Nakamata2023, Nakamata(ed)2023). Unlike other Japanese conversation corpora such as Corpus of Everyday Japanese Conversation (CEJC)(Koiso(ed)2026), J-TOCC enables researchers to examine how conversation topics affect vocabulary, grammatical patterns, and discourse strategies.

A distinctive feature of J-TOCC is that pairs of university students who know each other well engage in five-minute conversations on each of 15 designated topics. Because the same speakers discuss different topics for the same amount of time, the corpus enables researchers to examine the effects of topic while minimizing the influence of individual differences.

The corpus has broad pedagogical applications. For content-oriented approaches such as Content-Based Instruction (CBI)(Brinton et al. 2003), Content and Language Integrated Learning (CLIL)(Coyle et al. 2010), and Task-Based Language Teaching (TBLT)(Ellis 2003), it helps identify the linguistic resources that are characteristic of individual topics, making it possible to design topic-specific teaching materials and learning activities. At the same time, it is also useful for more traditional structure-based syllabi, as it reveals the topics in which particular grammatical forms are most frequently employed, thereby providing meaningful contexts for grammar instruction.

Beyond language teaching, J-TOCC has been widely used as a research resource for investigating grammar(Miyoshi 2023), lexicons(An 2024), gender differences(Ishikawa 2023), discourse patterns(Nakamata 2024), register(Nakamata in print), and other aspects of spoken Japanese, demonstrating its value across a broad range of linguistic studies.

The first version of J-TOCC was released in 2020. It contained 15 topics, most of which were related to everyday life. As a result, the corpus lacked topics addressing social issues, which are frequently used in intermediate- and advanced-level Japanese language classes. To address this limitation, we recorded a new set of conversations on topics involving social issues and expanded the corpus as the second phase of J-TOCC.

2. Design of J-TOCC2

Hereafter, the data from the first phase will be referred to as J-TOCC1, and the data from the second phase as J-TOCC2.

2.1 Participants

All participants were pairs of univesity styudents who were close friends.

The numbers of pairs are shown in Table 1.

	J-TOCC1		J-TOCC2	
Gender pair	East	West	Wast	West
Male-Male	20	20	30	10
Male-Female	20	20	20	20
Female-Female	20	20	10	30
Total	120		120	

Table 1: Numbers of participants in J-TOCC

The participants were balanced by gender and region. Regional balance was taken into consideration because Japanese has large dialectal differences between eastern and western Japan.

A total of 240 pairs participated across J-TOCC1 and J-TOCC2. It should be noted, however, that the distribution of gender pairings differs between the two phases. The table also indicates that each topic consists of 120 conversation files.

Because no participants took part in both phases, a total of 480 individuals participated in the project. J-TOCC1

contains 1,800 conversation files, calculated as 120 pairs × 15 topics, while J-TOCC2 contains 1,920 files, calculated as 120 pairs × 16 topics.

Each conversation lasts five minutes. Collecting conversations from 120 pairs yields exactly ten hours of speech for each topic. Consequently, the corpus contains 10 hours of transcribed conversation per topic and 310 hours in total across J-TOCC1 and J-TOCC2.

2.2 Recording

The recordings for J-TOCC2 were conducted face-to-face between 2023 and 2024. After receiving instructions, the participants were presented with a board displaying an SDGs topic. They were given 30 seconds to think about the topic before engaging in a five-minute conversation, which was audio-recorded. The use of smartphones during the conversation was permitted. The order in which the topics were presented was randomized.

After completing all the conversations, the participants filled out the face sheet and consent form. The recordings were conducted after obtaining approval from the research ethics committee at the participants' respective universities.

2.3 Topic Selection

Table 1 shows the topics in J-TOCC1 : 11 daily topics and 4 topics related to society.

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 2: The Topics in J-TOCC1

For J-TOCC2, it was necessary to select topics suitable for discussion-based classes at the advanced level and for content-oriented approaches such as Content and Language Integrated Learning (CLIL). As a result, the Sustainable Development Goals (SDGs) were adopted as the conversation topics because they are universally relevant, provide valuable opportunities for research not only in Japanese language education but also in higher education for native speakers, and facilitate cross-linguistic comparative studies.

The corpus includes 16 SDG topics. Goal 17, *Partnerships for the Goals*, was excluded because it functions primarily as a means of achieving the other goals rather than as an independent topic. As a meta-level concept, it was considered relatively difficult for university students to discuss in a conversation task.

For each conversation, participants were shown only the Japanese SDG logo and title corresponding to the assigned topic. No additional instructions or explanations were provided. However, because the use of smartphones was permitted, some pairs searched SDG-related websites for subtopics and background information to facilitate their discussion.

No.	Topic
21	No Poverty
22	Zero Hunger
23	Good Health and Well-Being
23	Quality Education
24	Gender Equality
26	Clean Water and Sanitation
27	Affordable and Clean Energy
28	Decent Work and Economic Growth
29	Industry, Innovation and Infrastructure
30	Reduced Inequalities
31	Sustainable Cities and Communities
32	Responsible Consumption and Production
33	Climate Action
34	Life Below Water
35	Life on Land
36	Peace, Justice and Strong Institutions

Table3: The Topics in J-TOCC2

2.4 Background Information of Participants

After completing the recordings, all participants were asked to fill out a face sheet. It 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 familiality.

The level of topic familiality 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. By using the degree of topic knowledge, we can compare those who are familiar with a topic with those who are not.

In addition, participants also completed a Vocabulary Size Test[1] and reported their estimated vocabulary size. This information was newly collected for J-TOCC2 and was not included in J-TOCC1. With this information, we can examine how volabulary size affects the way of talking.

[1]https://www.kecl.ntt.co.jp/icl/lirg/resources/goitokusei/vocabulary_test/php/login.php

The corpus includes data other than the speaker's name as information on the speaker.

2.5 Transcription and Distribution

In J-TOCC1, the transcriptions were conducted by a professional transcription company. In contrast, the recordings for J-TOCC2 were transcribed using AI-based transcription software. The transcription and subsequent correction were carried out by students from the same region in which the recordings had been made, and in some cases from the same university as the participants. The transcription software was configured so that the audio data would not be used for AI training. Because the AI-generated transcripts were not completely accurate, all transcripts were manually corrected, and speaker identification was also manually verified.

The transcription policy was as follows. Fillers and backchannels were not restored when they occurred alone. However, compound expressions such as sō desu ne ("I see") and conjunctions such as demo ("but") were restored. For overlapping speech, substantive utterances were restored, whereas utterances consisting only of fillers or backchannels, such as sō sō sō, were not. If an utterance could not be identified with sufficient confidence, it was left untranscribed rather than being guessed.

As a result, the J-TOCC2 data are expected to contain fewer fillers and backchannels than J-TOCC1. Therefore, J-TOCC is not suitable for research focusing on fillers or backchannels.

The corpus is publicly available on the website of the first author.[2] Anyone may use it after agreeing to the terms of use. In addition to the plain-text transcripts, the website provides morphologically analyzed files for the entire corpus and search programs that run on Google Colab, allowing even novice users to search the corpus with morphological information and analyze by topics, regions, and gender.

3. Results

3.1 Word Size

The number of words in the first phase is shown in Table 1, and the number of words in the second phase is shown in Table 2.

A comparison of J-TOCC1 and J-TOCC2 reveals that the overall amount of speech is lower in J-TOCC2. This tendency was also noticeable during the recording sessions, where participants in J-TOCC2 appeared to pause more frequently before responding, resulting in less fluent conversations than those in J-TOCC1.

One possible explanation is the difference in the nature of the topics. J-TOCC1 primarily consists of everyday-life topics. In contrast, all 16 topics in J-TOCC2 are related to the SDGs, requiring participants to discuss social issues throughout the recording sessions. Such topics may have been more demanding and therefore less conducive to spontaneous conversation.

The discriptive statistics support this observation. The average number of words per topic is 145,979 in J-TOCC1 and 120,288 in J-TOCC2, meaning that J-TOCC2 contains only 82.4% as many running words as J-TOCC1. The difference is even greater for vocabulary diversity. The average number of word types per topic decreases from 6,316 in J-TOCC1 to 3,995 in J-TOCC2, corresponding to only 63.2% of the value for J-TOCC1. In other words, the reduction in lexical variety is greater than the reduction in the amount of speech, suggesting that participants in J-TOCC2 tended to reuse the same vocabulary more frequently.

No.	Topic	Token	Type
1	Eating	145,361	6,338
2	Fashion (Clothes)	148,909	6,111
3	Travel	147,386	6,660
4	Sports	148,722	6,810
5	Manga and Games	151,454	7,106
6	Housework	147,950	6,141
7	School	145,198	6,794
8	Smartphone	144,745	6,145
9	Part-Time Job	147,120	6,791
10	Animals	148,636	6,455
11	Weather	146,245	6,242
12	Dreams and Life Plan	141,101	6,034
13	Manners (on Public Transport)	148,209	5,586
14	Living Environment	141,934	5,578
15	Declining Birthrate and Aging Population	136,718	5,950
	Subtotal of Phase 1	2,189,688	42,756
21	No Poverty	119,415	3,837
22	Zero Hunger	120,742	4,016
23	Good Health and Well-Being	120,330	3,990
24	Quality Education	123,864	3,704
25	Gender Equality	124,456	3,767
26	Clean Water and Sanitation	122,999	3,873
27	Affordable and Clean Energy	118,141	3,781
28	Decent Work and Economic Growth	118,343	3,787
29	Industry, Innovation and Infrastructure	115,146	4,213
30	Reduced Inequalities	117,181	3,857
31	Sustainable Cities and Communities	120,456	4,175
32	Responsible Consumption and Production	122,353	4,182
33	Climate Action	121,129	4,177
34	Life Below Water	122,883	4,159
35	Life on Land	118,869	4,170
36	Peace, Justice and Strong Institutions	118,295	4,231
	Subtotal of Phase 2	1,924,559	16,647

Table 4: Word size of J-TOCC

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

A similar tendency can be observed by comparing the total number of word types in each corpus with the average number of word types per topic. This ratio is 6.77 for J-TOCC1 but only 3.48 for J-TOCC2, indicating that the lexical differences among topics are much smaller in J-TOCC2. One possible explanation is that discussions of SDG-related issues share academic words (Coxhead 2000) across different topics, resulting in greater lexical overlap throughout the corpus.

Furthermore, the coefficient of variation—calculated by dividing the standard deviation by the mean—is used as a measure of variability. When this was calculated for token frequency and type frequency in the first and second periods, the coefficient of variation for token frequency was 0.025 in the first period, whilst for type frequency it was 0.022 in the second period. As for type frequency, the value for the first period was 0.072, whilst that for the second period was 0.040. Although both are homogeneous, it can be said that the differences in size

3.2 Topic Familiarity

Figure 1 & 2 show the average topic familiarity.

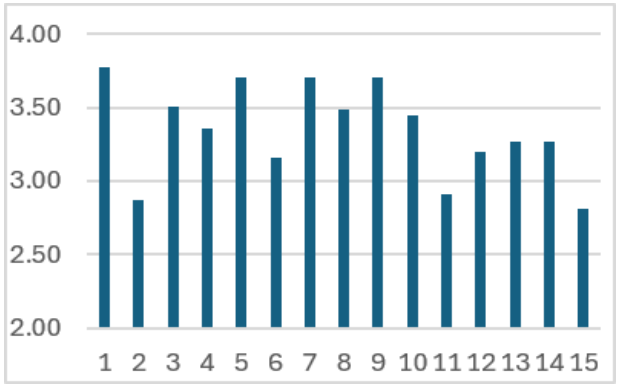

Figure 1: The average of Topic Familiarity (Phase 1)

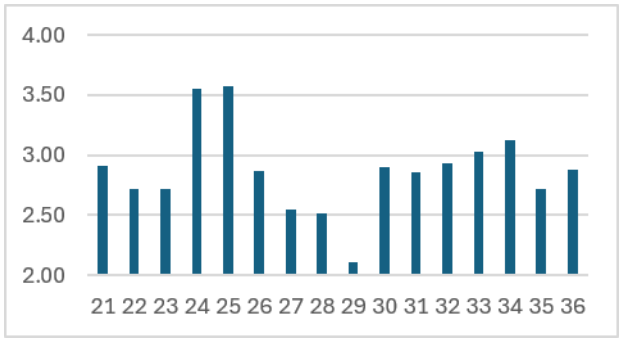

Figure 2: The average of Topic Familiarity (Phase2)

Among the SDGs topics (Phase 2), ’24. Quality Education’ and ‘25. Gender Equality’ show the highest topic knowledge, which are as high as common topics in Phase 1 such as ‘1. Eating’ and ‘5. Manga and Game.’ Nakamata (in printing) surveyed Phase 1 data and found that conversations of topics with high familiarity contain more elements of ‘narrative.’ This means that participants talked more about their own experiences. It should be clarified if these two topics in the SDGs also contain narratives or experiences.

’29. Industry, Innovation and Infrastructure’ shows the lowest familiarity by far.’27. Affordable and Clean Energy’, which is related to science and technology, also shows low familiarity.

3.3 Specific Words

This section examines the specific words associated with a given topic, using '21. No Poverty' as an example. Table 5, 6, and 7 show the specific words of nouns, verbs, and adjectives[3] within this topic. The Log Likelihood ratio (LLR) was calculated as designating ’21. No Poverty’ as the target corpus, and the other 15 topics from the SDGS as reference corpora.

Words	Translation	LLR
Hinkon	poverty	4,323
o-kane	money	439
seekatsu-hogo	public assistance	308
kodomo	child	213
soutaiteki-hinkon	relative poverty	202
hoomulesu	homelessness	153
zettaiteki-hinkon	absolute poverty	141
kodomo-shokudoo	children’s cafeterria	117
sihon-syugi	capitalism	109
bokin	fundraising	90

Table 5. Noun-specific words in ’21. No Poverty’

Words	Translation	LLR
nakusu	get rid of	494
iru	be, exist	104
komaru	have troubles, be in need	81
ageru	give, raise	75
kasegu	earn	59
ikiru	live, survive	48
naku-naru	be lost, disappear	44
nuke-dasu	escape	44
sikyuu-suru	provide	38
morau	receive, get	38

Table 6. Verb-specific words in ’21. No Poverty

[3] Japanese has two types of adjectives : i-adjectives and na-adjectives. Both were included in the analysis.

Words	Translation	LLR
mazusii	poor	436
sootaiteki	relative, relatively	104
yuuhuku	rich	100
bussitu-teki	material	49
kinsen-teki	financial, monetary	42
nai	no	42
binboo	poor	31
zettai-teki	absolute, absolutely	30
seesin-teki	mental, mentally	22
keezai-teki	economic, economically	21

Table 7. Adjective-specific words in ’21. No Poverty

‘Nakusu’ and ‘hinkon’ are extremely specific because they were included in the goal statement in Japanese and were presented to the participants. Among nouns, it is notable that ‘soutaiteki-hinkon’ (relative poverty) and ‘zettaiteki-hinkon’ (absolute poverty) are frequently appeared in the corpus, indicating that university students recognize the importance of distinguishing these two concepts when discussing poverty. For verbs, those related to giving and getting were highly specific. Among adjectives, it was remarkable that five of the top ten words were Sino-Japanese words ending with the suffix *-teki*, which are frequently used in technical discussions.

A list of specific words for 31topics has already been published on the first author’s website, providing important concepts for various topics treated in Japanese language education. From another perspective, however, they are the concepts Japanese university students need when discussing each of these topics. By translating the words in this list into English, we can provide useful information for Japanese students in English classes and multicultural co-learning courses. In this way, the topic-oriented corpus holds great potential for applications.

4. Summary

The key concept behind a topic-base corpus is to collect an equal amount of conversation on a specified topic. This enables researchers to examine topic-specific patterns of vocabulary, grammar, and language use.

Considering background information such as topic proficiency and estimated vocabulary size also brings us new insights into linguistic competence and performance at the personal level.

Furthermore, such corpora can be constructed for other languages as well.

Acknowledgement

This work was supported by JSPS KAKENHI Grant Number JP22H00668.

Bibliographical References

An, R. (2024). ‘Junbi’ to ‘Youi’ no Tsukaiwake nitsuite:Meishi-yoohoo to Kango-sahen-yoohoo o Rei ni [Usage of “Junbi” and “Youi”], *Toua Daigaku Kiyou* 38, pp.45-65.

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

Coxhead, A.(2000). A New Academic Word List. *TESOL Quarterly*, 34(2), pp.213-238.

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.

Ishikawa, S. (2025). Koopasu Kiban-gata Seisa Kenkyuu no Houhou-ron : Danjo-daigakusee no Kaiwa ni Miru Seisa o Megutte [A Meyjodology of Corpus-based Study off Gender Difference : Gender Differences Appeared in Convesations among Male and Female University Students] In Moriyama, Yu., Kato, D., and Ishikawa, S. (eds.) *Onna-kotoba Otoko-kotoba o Koete [Beyond Male and Female Language]*. Hitsuji Shobo, Tokyo

Koiso, H.(ed.) (2026). *Nichijoo Kaiwa Koopasu no Sekkei Koochiku Rikatsuyoo [The Design, Construction, and Utilization of an Everyday Conversation Corpus]*, Kenkyusha, Tokyo.

Miyoshi, Y. (2023). Wadai kara Miru ’Shiyoo to omou/omotteiru’ to ‘Tsumori(da)’ no Tsukaukata[The Usage of ’Shiyoo to omou/omotteiru’ and ‘Tsumori(da)’ in from the Perspective of Topics], *Nihongo/Nihongo-kyoiku Kenkyuu* 14, pp.121-136.

Nakamata, N. (2023). Building a Topic-Oriented Corpus and Its Application for Language Teaching. In *Proceeding of The 37th Pcific Asia Conference on Language, Information and Computation*, pp.210-226.

Nakamata, N. (ed.)(2023). *Wadai-betsu Koopasu ga Hiraku Nihongo, Nihongo-kyooiku-kenkyuu [Japanese Language and Japanese Language Education Research Pioneered by Topic-oriented Corpus]*. Hitsuji Shobo, Tokyo

Nakamata, N. (2024). Zatsudan no Can-Do-ka o Mezashita Wadai no Ruikei no Bunseki: ‘Taberu Koto’ o Rei toshite [An analysis of Topic Pattern for Making Can-dos of Small Talks: In the Case of ‘Eating’]. In Saito, K. and Shu, T.(eds.) *Danwa Bunsyoo Tekusuto no Hitomatomari-sei [Unitarity of Discourse, Document, and Text]*. pp.165-188, Izumi Shoin, Osaka.

Nakamata, N. (in print). Wadai-betsu Koopasu ni taisuru Buntai Bunseki to Wadai no Bunrui:Tiiki-sa Wadai-sa Seisa[Clustering Topics Based on a Stylistic Analysis against Conversation Corpus: Variation across Place, Topic, and Sex] *Shakaigengokagaku*, 28(2).