Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 5587–5594
Marseille, 20-25 June 2022

Design and Evaluation of the *Corpus of Everyday Japanese Conversation*

Hanae Koiso[†], Haruka Amatani[†], Yasuharu Den[‡], Yuriko Iseki[†], Yuichi Ishimoto[†], Wakako Kashino[†], Yoshiko Kawabata[†], Ken'ya Nishikawa[†], Yayoi Tanaka[†], Yasuyuki Usuda[†], Yuka Watanabe[†]

[†] National Institute for Japanese Language and Linguistics
10-2 Midoricho, Tachikawa, Tokyo 190-8561, Japan
{koiso, h-amatani, iseki, yishi, waka, kawabata, nishikawa, yayoi, usuda, yuwatanabe}@ninjal.ac.jp

[‡] Graduate School of Humanities, Chiba University
1-33 Yayoicho, Inage-ku, Chiba 263-8522, Japan
den@chiba-u.jp

Abstract

We have constructed the *Corpus of Everyday Japanese Conversation* (CEJC) and published it in March 2022. The CEJC is designed to contain various kinds of everyday conversations in a balanced manner to capture their diversity. The CEJC features not only audio but also video data to facilitate precise understanding of the mechanism of real-life social behavior. The publication of a large-scale corpus of everyday conversations that includes video data is a new approach. The CEJC contains 200 hours of speech, 577 conversations, about 2.4 million words, and a total of 1675 conversants. In this paper, we present an overview of the corpus, including the recording method and devices, structure of the corpus, formats of video and audio files, transcription, and annotations. We then report some results of the evaluation of the CEJC in terms of conversant and conversation attributes. We show that the CEJC includes a good balance of adult conversants in terms of gender and age, as well as a variety of conversations in terms of conversation forms, places, activities, and numbers of conversants.

Keywords: Corpus of everyday Japanese conversation, corpus design, corpus evaluation

1. Introduction

The characteristics of languages and behaviors in conversations occurring in daily life are an important research theme. Although many corpora of Japanese conversations have been constructed, most have targeted conversations in artificially created settings in terms of topics and recording situations or were biased in terms of conversational situation. In addition, there is no corpus of Japanese conversations that provides video data, and some do not even provide audio data (see Table 1).

Given this situation, we have engaged since 2016 in the construction of the *Corpus of Everyday Japanese Conversation*, CEJC. The main features of the CEJC are i) a focus on conversations embedded in naturally occurring activities in daily life, ii) the variety of everyday conversations, collected in a balanced manner to capture the diversity of everyday conversations and observe natural conversational behavior in our daily life; and iii) the publication of audio and video data in order to precisely understand the mechanism of real-life social behavior. We published a part of the CEJC, 50 hours of conversations, in 2018 on a trial basis, and will publish the whole corpus, comprising 200 hours, in March 2022.

In this paper, we present an overview of the corpus, including the recording method and devices, structure of the corpus, formats of video and audio files, transcription, and annotations in the CEJC. We then report the results of the evaluation of the CEJC in terms of conversant and conversation attributes.

2. Corpus Design

2.1. Recording method

To capture various kinds of conversations in daily life, conversations in the CEJC were recorded on the basis of two recording methods — *individual-based* and *situation-specific* methods (Koiso et al., 2016a)— the designs of which were based on the recording methods of the spoken language part of the British National Corpus (Crowdy, 1995; Burnard and Aston, 1998).

Individual-based method Of the 200 hours comprising the CEJC, 185 hours, featuring 533 conversations, were recorded using the individual-based method. We recruited 40 informants, balanced in terms of gender and age (man/woman, 20s/30s/40s/50s/over 60), provided them with portable recording devices for approximately three months, and had them record around 15 hours of conversation in their daily activities. Table 2 shows the attributes of the informants. All informants recorded their everyday activities in a variety of situations, such as at home, at the office, at a restaurant, and outdoors. About four to five hours of conversations, around 15 hours per informant, were selected for the CEJC by taking into account the balance

Table 1: Major corpora of Japanese conversations

Corpus Name	Size	Contents
Japanese Topic-Oriented Conversational Corpus (J–TOCC) (Nakamata et al., 2020)	1800 conversations 150 hours	Chats between two close undergraduate students on campus (15 kinds of topics assigned, audio files unavailable)
BTSJ Japanese Natural Conversation Corpus (Usami, 2021)	446 conversations 112.5 hours	Chats among friends, professor-student mentoring, telephone conversations, etc. (audio files available only for some portion)
Chats in Everyday Life	96 conversations 21 hours	Natural conversations in everyday life (audio files unavailable)
Chiba Three-Party Conversation Corpus (Den and Enomoto, 2007)	12 conversations 2 hours	Chats among three undergraduate/graduate students on campus (topics assigned)
Sakura Corpus	18 conversations	Chats among four undergraduate students (topics assigned)
Meidai Conversational Corpus (Fujimura et al., 2012)	129 conversations 100 hours	Chats among friends (audio files unavailable)
Women's Language in the Workplace Men's Language in the Workplace	111 conversations 21 hours	Natural conversations in formal and informal situations in the workplace (audio files unavailable)
CALL HOME Japanese (Den and Fry, 2000)	120 conversations 20 hours	Telephone conversations between Japanese living in the U.S. and their families/friends in Japan
CallFriend Japanese	31 conversations	Telephone conversations between Japanese living in the U.S.

Table 2: Attributes of informants. Number and total duration of conversations for each informant.

	Male			Female		
Age	Occupation	No. conversations	Duration	Occupation	No. conversations	Duration
20s	student	10	4.3h	student	14	4.4h
	student	10	4.2h	student	21	6.0h
	teacher	14	5.5h	office worker	15	4.2h
	teacher	17	3.7h	office worker	8	4.0h
30s	office worker	12	3.1h	office worker	12	5.0h
	office worker	11	4.7h	freelance	22	4.8h
	freelance	11	5.6h	freelance	17	5.4h
	civil servant	14	4.6h	homemaker	12	5.6h
40s	office worker	10	3.6h	office worker	9	4.5h
	office worker	11	3.9h	part-time worker	12	5.0h
	teacher	23	5.0h	part-time worker	10	4.8h
	freelance	13	4.8	self-employed	17	4.4h
50s	office worker	9	4.6h	office worker	14	4.2h
	office worker	17	6.0h	office worker	12	4.5h
	teacher	9	4.2h	self-employed	11	4.6h
	teacher	10	4.2h	freelance	12	4.6h
Over 60	volunteer	14	5.8h	homemaker	13	5.1h
	retired	13	4.6h	office worker	12	4.2h
	freelance	18	4.8h	self-employed	14	4.4h
	teacher	17	4.3h	freelance	13	4.3h
total		263	91.5h		270	94.0h
				total	533	185.5h

of conversation variations and quality of recorded data.

Situation-specific method Of the 200 hours comprising the CEJC, 15 hours, representing 44 conversations, were recorded on the basis of the situation-specific method in order to compensate for the shortage of recordings using the individual-based method. By verifying the balance for 94 hours of conversations collected using the individual-based method in the middle of building the CEJC, (Koiso et al., 2018), it turned out that extra recordings on business situations and involving younger people were required. Then, total of 10 hours of business meetings and five hours of conversations by minors were recorded using the situation-specific method.

2.2. Recording devices

Video Two types of camera, the Kodak PIXPRO SP360 4K and GoPro Hero3+, were mainly used when recording indoors. Figure 1 shows video images of a conversation at a family gathering for an anniversary at a restaurant. The left image was recorded using an

Figure 1: Video images of a conversation at a family gathering for an anniversary at a restaurant. The faces of the conversants are airbrushed to protect their identity in the printed material, although they remain intact in the video data to be published.

SP360 camera located on the central table, while the top- and bottom-right images were recorded by two GoPro cameras placed facing each other on the other tables. A wearable camera (Panasonic HX-A500 camera) was used for outdoor recording. The HX-A500 was mounted on or around the conversant's shoulder to capture what they were looking at. In addition to these cameras, Sony HDR-CX675 cameras were sometimes used when recording meetings at work based on the situation-specific method.

Audio Each conversant hung an IC recorder, a SONY ICD–SX734, around their aneck, and their voice was recorded with their own recorder. All conversants' voices were also recorded by another IC recorder, a SONY ICD–SX1000, generally located at the center of the conversation space.

2.3. Structure of the CEJC

The whole corpus contains video and audio data, transcripts, and two types of POS annotations. There is a subset of the corpus, named the Core dataset, which consists of 20 hours of conversations, corresponding to 10% of the entire corpus. The Core includes five kinds of manually labeled or corrected annotations (see Figure 2).

2.3.1. Video and audio files

Table 3 shows the file formats of video and audio data in the CEJC.

As for the video and audio data collected for the CEJC, we discussed with a lawyer specializing in copyright and portrait-right issues the appropriate way to deal with legal and ethical problems regarding the portrait rights, copyright, and the protection of personal information (Koiso et al., 2018). The video files were handled as follows: (1) The faces of the main conversants who agreed to have their faces published were left unblurred. (2) The faces of third parties performing common activities that were not sensitive activities and in public places were regarded as outside the scope of protection of portrait rights and were left unblurred. (3) The faces of third parties who are interpreted as being inside the scope of protection of portrait rights were airbrushed. (4) The faces of third parties talking to the

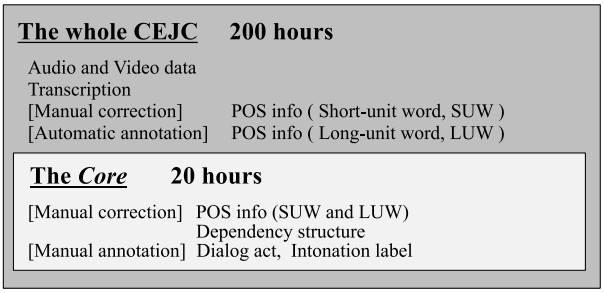

Figure 2: The Structure of the CEJC

Table 3: Audio and video file formats in the CEJC

Recording device	File format
Video	
PIXPRO SP360	mp4, H264, 1440×1440, 29.97fps
GoPro Hero3+	mp4, H264, 1280×720, 29.97fps
HX-A500	mp4, H264, 1280×720, 29.97fps
HDR-CX675	mp4, H264, 1280×720, 29.97fps
mixed file (see Fig 1)	mp4, H264, 1360×720, 29.97fps
Audio	
ICD-SX734	linear PCM, 16bit, 16kHz, mono
ICD-SX1000	linear PCM, 16bit, 16kHz, stereo

SpeakerID	StartTime	EndTime	Text	Note
IC02	13.030	14.822	(F ano) jitensha you ni *Well, for my bicycle,*	(F xx): filler (for morphological analysis)
IC02	14.952	17.524	keidensu hakaritakute kaitensuu toka. *I wanted to count my cadence, or revolution.*	.: Boundary of an utterance unit
IC01	15.359	16.261	aa. *Oh.*	
IC01	17.752	18.066	aa. *I see.*	
IC02	17.986	18.210	de *And*	
IC01	18.066	18.949	sugoi ne:. *It's amazing.*	: : prolongation
IC02	18.943	19.347	de: *then,*	
IC02	19.574	22.397	nanka yappa kyuujuu gurai de kyuujukkaiten. *you know, kinda 90, 90 revs per minute.*	
IC01	21.575	22.525	un un un. *Yep yep yep.*	
IC02	22.397	23.702	hayaku mawasu tte (W yu|yuuka) *pushing pedals down sorta faster*	(W xx|yy): "xx" / reduced or incorrect pronunciation "yy" / supposed-to-be correct word
IC01	23.347	24.671	un un un un. *Yep yep yep yep.*	
IC02	23.830	25.678	karuku(D #) karui gia de(D haya) *light, with a lighter gear fas-*	(D x): word fragments
IC01	25.872	26.638	un. *Yeah.*	
IC02	26.104	26.489	(F ano) (D itt) *well, it's g-*	
IC02	26.859	29.046	hayai kaitensuu de mawasu no ga indesu tte. *it's said to be good to rotate in a faster rate.*	
IC01	28.374	29.765	un un un un. *Yep yep yep yep*	

Figure 3: Sample transcript. In the actual transcript, texts are written in Japanese characters, and the boundary of an utterance unit is marked by an "ideographic full stop."

main conversants and whose utterances are transcribed were airbrushed. The audio and transcription files were handled as follows: Personal information, including conversants' names, affiliations, and individual identification information, as well as any parts of recordings for which conversants had not given their permission for publication, were replaced by anonyms or turned letters in transcripts, and the corresponding regions of the audio files were rendered inaudible.

2.3.2. Transcription

Figure 3 shows a sample transcript. The speech data were manually transcribed in the standard Japanese orthography with reference to the video and audio data using ELAN and/or Praat, based on the units that are divided at the locations of perceptible pauses and the boundaries of utterance units. During the transcription, about 20 different tags, such as prolongation, disfluency, and nonverbal events, were applied. These tags were defined in reference to the transcription criteria and conventions previously used in the Corpus of Spontaneous Japanese (CSJ) (Maekawa, 2004) and in the Chiba Three-party Conversation Corpus (Den and Enomoto, 2007).

2.3.3. Annotation

Two types of POS information Two different POS systems, short unit word (SUW) and long-unit word (LUW), were adopted. SUWs are basically monomorphemic words or words made up of two morphemes, while LUWs are multi-morphemic words, including compound words like compound nouns, compound verbs, and compound particles (see Figure 4).
For SUWs, the data were automatically analyzed using Mecab, a Japanese morphological analysis implementation, and UniDic, a dictionary for Mecab, which were manually corrected. UniDic was developed for POS annotation of the *Balanced Corpus of Contemporary Written Japanese* (Maekawa et al., 2014). LUWs were automatically constructed from one or more SUWs, and those in the Core were manually corrected.

SUW			LUW		
Entry	POS	Gloss	Entry	POS	Gloss
その	Pronoun	DEM	その	Pronoun	DEM
講義	Noun	lecture	講義	Noun	lecture
テーマ	Noun	theme	テーマ		theme
に	Particle	DAT	に	Particle	about
つい	Verb	attach	つい		
て	Particle	*–te*	て		

Figure 4: Example of two-way POS annotation

Dependency structure Dependency structures between *bunsetsu* phrases, which are composed of content words possibly followed by one or more function words, were automatically analyzed within utterance units and manually corrected in the Core

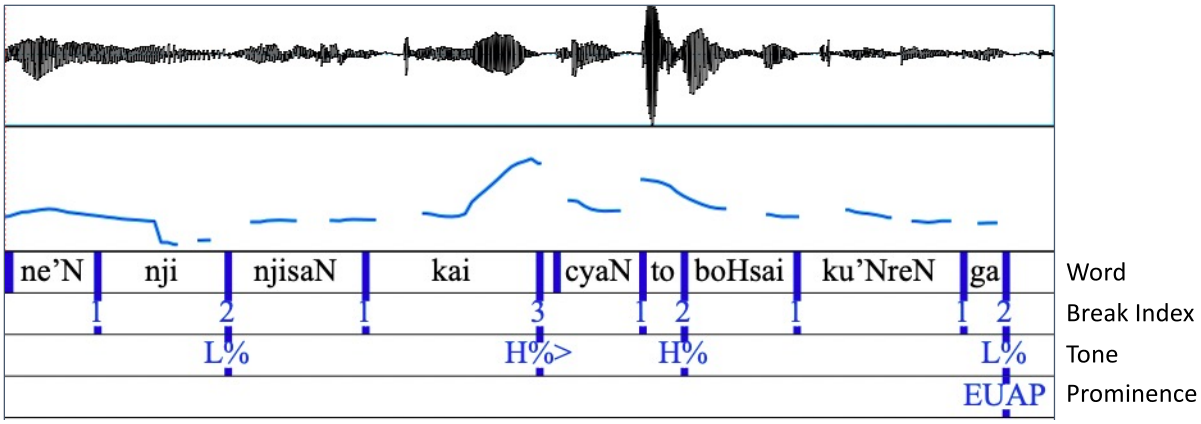

Figure 5: Sample labeling of a simplified version of X-JToBI
Break Index (BI): prosodic phrasing levels, such as a word (BI=1), accent phrase (2), and intonation phrase (3);
Tone: boundary tones, such as a falling tone (L%), rising tone (H%), and rising-falling tone (HL%);
Prominence: non–lexical prominence, such as emphasis of an unaccented accentual phrase (EUAP).

(Asahara and Matsumoto, 2016).

Dialog act Dialog acts were manually annotated according to the ISO 24617-2 scheme (ISO 24617-2, 2012) extended to everyday conversations (Iseki et al., 2019) in the Core. ISO 24617-2 includes communicative functions and functional dependence relations. Information about communicative function is divided into two types: semantic/pragmatic-level tags (e.g., question, answer, attention getting, and thanking) and interactional-level tags (e.g., repair, opening, and closing) in our framework.

Intonation label The utterances of 152 of the 157 main conversants included in the Core, which were selected on the basis of recording conditions and degrees of dialect and recording conditions, were manually labeled according to a simplified version of the intonation-labeling scheme, X–JToBI (Maekawa et al., 2002), which was developed for spontaneous Japanese speech included in the CSJ. Figure 5 shows a sample labeling of a simplified version of X–JToBI.

3. Evaluation of CEJC

The CEJC contains 200 hours of speech, 461 sessions, 577 conversations, about 2.4 million words (short-unit words, see above), and a total of 1675 conversants, including 862 different participants.
In this section, we will evaluate whether the CEJC is balanced in terms of conversant and conversation attributes.

3.1. Conversant attributes

Figure 6 shows the distributions of numbers of conversants and words by gender and age included in the CEJC. In the individual-based method that is the main recording method of the CEJC, since the informant over 20 years old shown in Table 2 recorded conversations mainly with their friends, colleagues, and family members, many conversants of the same generation as the informant are naturally included in the corpus. Figure 6 shows that although there are some differences, such as a small number of males in their 40s and 50s, and many females in their 30s, 40s, and 50s, the CEJC contains about 100 or more conversants and more than 150,000 words for both males and females of all generations over 20 years of age. Since a preliminary survey of about 100 hours of conversations recorded using the individual-based method showed that there were few children under 20 years of age in the data, we also recorded, using the situation-specific method, about five hours of conversations by junior high and high school students. However, the figure shows that the conversations of minors are far fewer than those of adults of all ages.

3.2. Conversation attributes

One of the main features of the CEJC is that it contains various kinds of everyday conversation in a balanced manner. To establish a corpus design for such a balanced corpus, we conducted a survey of everyday conversational behavior with about 250 adults in order to show the diversity of our everyday conversational behavior (Koiso et al., 2016b). The questionnaire included when, where, how long, with whom, and in what kind of activity informants were engaged in conversation. On the basis of the results, we derived the approximate distributions of conversation forms, conversation places, accompanying activities, and the number of conversants included in each conversation as a measure of the design of a balanced corpus. In this section, we compare the conversations included in the CEJC with the survey results.
Figures 7 and 8 show comparisons between the CEJC and the survey results in terms of conversation forms, conversation places, accompanying activities, and the number of conversants in each conversation. Figure 7 and 8 correspond to the number of conversations and

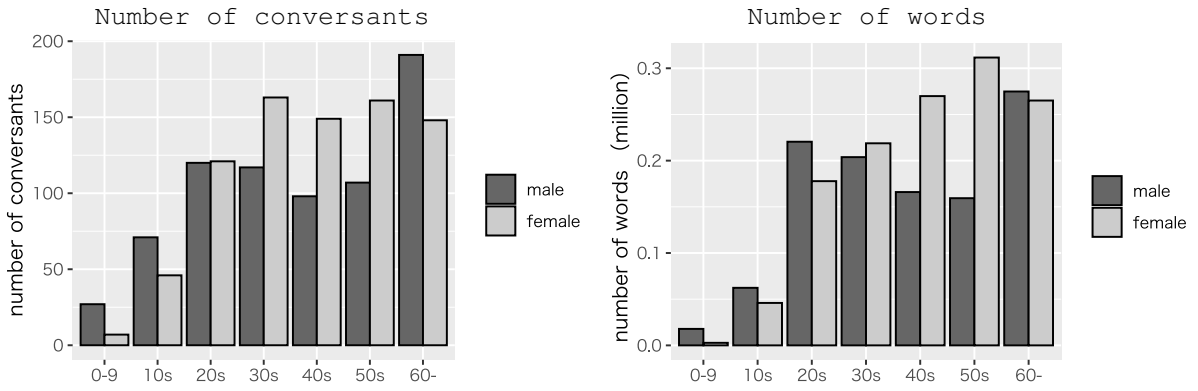

Figure 6: Distributions of numbers of conversants and words by gender and age

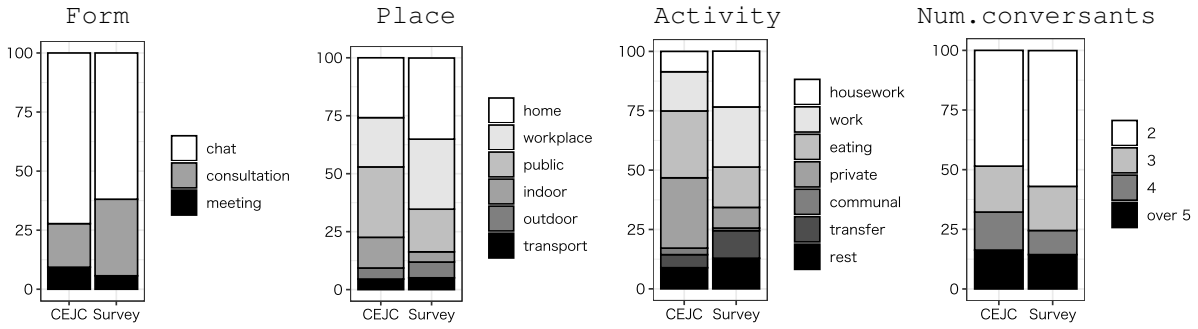

Figure 7: Distributions of the conversation forms, places, activities, and number of conversants in the CEJC and the survey results of conversational behavior

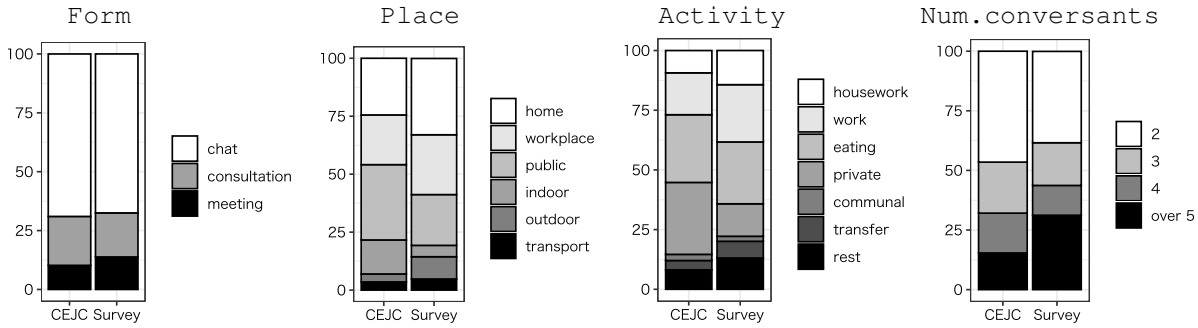

Figure 8: Distributions of total duration of conversation forms, places, activities, and number of conversants in the CEJC and the survey results of conversational behavior

the total duration of conversations for each category, respectively.

Conversation form Figure 7 shows that there are slight differences in that the ratio of the number of chats in the CEJC is higher than that in the survey, while the ratio of the number of consultation/business talks is lower, although Figure 8 indicates that the CEJC shows similar tendencies to the survey in terms of total duration. This difference comes from the fact that the survey included more short consultations/business talks of less than five minutes than the CEJC did. Although there is only a slight difference, it can be said that, with regard to conversation forms, the CEJC contains conversations in a balanced manner with reference to the survey results.

Place and Activity The distributions of places and activities differ between the CEJC and the survey as follows: (1) the CEJC includes more conversations in public and commercial facilities and indoors, such as restaurants, city halls, friends' houses, and parents' houses, and fewer conversations at home, at school, and in the workplace than the survey; (2) the CEJC contains more conversations during private activities, such as spending time with family and friends, while

there were fewer conversations during housework and work/schoolwork than the survey. As mentioned in Section 2.1, by verifying the balance of the 94 hours of conversations collected using the individual-based method in the middle of building the CEJC, we also collected 15 hours of business meetings and conversations by minors using the situation-specific method. The augmentation of data significantly reduced the gap between the survey and CEJC, but not to the same extent.

The reasons why there are fewer conversations at home in the CEJC, even though informants recorded many conversations at home, are as follows. As reported by Koiso et al. (2018), we decided to include fewer conversations at home in the CEJC than in the survey to reduce the similar types of conversation, such as conversations while eating with a family member at home, and to ensure a variety of conversations.

Number of conversants in each conversation In Figure 7, the proportions of the number of conversants in each conversation show similar tendencies for the CEJC and the survey. However, differences can be seen in Figure 8. The total duration of conversations among five or more people is shorter in the CEJC than in the survey. This difference arises from the criteria of data collection for the CEJC. Although Koiso et al. (2016b) showed that conversations of five or more conversants tend to be long conversations over one hour, we asked informants to record at most an hour of conversation in order to collect many kinds of conversation for the CEJC.

4. Conclusions

In this paper, we first provided an overview of the corpus of the CEJC, including the recording method and devices, structure of the corpus, formats of video and audio files, transcription, and annotations. Next, we reported some results of the evaluation of the CEJC in terms of conversant and conversation attributes. We showed that the CEJC contains a good balance between adult conversants in terms of gender and age. On the other hand, it was also shown that the CEJC includes fewer children than adults. To correct for this bias, we plan to start a new project in April 2022, in which we will build a corpus whose main targets are children. As for conversational attributes, we evaluated the balance of the CEJC by comparing it with a survey of everyday conversational behavior. Although we found fewer conversations during work/schoolwork at workplaces/schools in the CEJC than in the survey, we showed that the CEJC contains a variety of conversations in terms of conversation forms, conversation places, and accompanying activities, as well as number of conversants.

The CEJC will be published in March 2022 in two ways: 1) a web-based search system accessible to two-way POS data, provided free of charge, and 2) the entire corpus, including audio/video data and all annotation, provided for a fee.

5. Acknowledgments

The work reported in this article is supported by the NINJAL collaborative research project "A Multifaceted Study of Spoken Language Using a Large-scale Corpus of Everyday Japanese Conversation".

6. Bibliographical References

Asahara, M. and Matsumoto, Y. (2016). BCCWJ-DepPara: A syntactic annotation treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pages 49–58, December.

Burnard, L. and Aston, G. (1998). *The BNC handbook*. Edinburgh University Press, Edinburgh, U.K.

Crowdy, S. (1995). The BNC spoken corpus. In G. Leech, et al., editors, *Spoken English on computer: Transcription, mark-up and application*, pages 224–235. Longman, Harlow, U.K.

Den, Y. and Enomoto, M. (2007). A scientific approach to conversational informatics: Description, analysis, and modeling of human conversation. In T. Nishida, editor, *Conversational informatics: An engineering approach*, pages 307–330. John Wiley & Sons, Hoboken, NJ.

Den, Y. and Fry, J. (2000). Callhome japanese corpus (in Japanese). *Joumal of the Phonetic Society of Japan*, 4(2):24–30.

Fujimura, I., Chiba, S., and Ohso, M. (2012). Lexical and grammatical features of spoken and written japanese in contrast: Exploring a lexical profiling approach to comparing spoken and written corpora. In *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*, pages 393–398.

Iseki, Y., Kadota, K., and Den, Y. (2019). Characteristics of everyday conversation derived from the analysis of dialog act annotation. In *Proceedings of the 22nd Conference of the Oriental COCOSDA*, pages 1–6.

ISO 24617-2. (2012). Language resource management — semantic annotation framework (SemAF) — Part 2: Dialogue acts.

Koiso, H., Tanaka, Y., Watanabe, R., and Den, Y. (2016a). A large-scale corpus of everyday Japanese conversation: On methodology for recording naturally occurring conversations. In *Proceedings of LREC 2016 Workshop: Just talking — Casual talk among humans and machines*, pages 9–12, Portoroz, Slovenia.

Koiso, H., Tsuchiya, T., Watanabe, R., Yokomori, D., Aizawa, M., and Den, Y. (2016b). Survey of conversational behavior: Towards the design of a balanced corpus of everyday Japanese conversation. In *Proceedings of LREC 2016*, pages 4434–4439, Portoroz, Slovenia.

Koiso, H., Den, Y., Iseki, Y., Kashino, W., Kawabata, Y., Nishikawa, K., Tanaka, Y., and Usuda, Y. (2018). Construction of the Corpus of Everyday Japanese Conversation: An interim report. In *Proceedings of LREC 2018*, Miyazaki, Japan.

Maekawa, K., Kikuchi, H., Igarashi, Y., and Venditti, J. J. (2002). X–JToBI: An extended J_ToBI for spontaneous speech. In *Proceedings of INTERSPEECH 2002*, pages 1545–1548, Denver, CO.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y. (2014). Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, 48(2):345–371.

Maekawa, K. (2004). Design, compilation, and some preliminary analyses of the *Corpus of Spontaneous Japanese*. In K. Yoneyama et al., editors, *Spontaneous speech: Data and analysis*, pages 87–108. The National Institute for Japanese Language and Linguistics, Tokyo.

Nakamata, N., Ota, Y., Kato, E., Sawada, H., Shimizu, Y., and Mori, A. (2020). J-tocc: Japanese topic-priented conversational corpus. In *11th International Conference of Practical Linguistics of Japanese*.

Usami, M. (2021). Btsj-japanese natural conversation corpus with transcripts and recordings. Technical report, NINJAL Institute-based projects: Multiple Approaches to Analyzing the Communication of Japanese Language Learners.