# Design and Preliminary Analysis of the *Corpus of Everyday Japanese Conversation*

**Hanae Koiso[†], Yasuyuki Usuda[†], Haruka Amatani[†], Yoshiko Kawabata[†], Yasuharu Den[‡,†]**

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

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

**Abstract**

Conversations emerge in various ways in everyday life. To capture the diversity of real-life conversations, we started the compilation of a large-scale corpus of everyday Japanese conversation, the *Corpus of Everyday Japanese Conversation*, CEJC. The CEJC is designed to contain various kinds of everyday conversations in a balanced manner so as to capture the diversity of everyday conversations and to observe natural conversational behavior. The CEJC targets conversations embedded in naturally occurring activities in daily life, without the exogenous intervention by researchers imposing topics or displacing the context of action. Since the start of the project in 2016, we have compiled 94 hours of conversations in the CEJC, corresponding to about a half of the target size of the entire corpus, and have morphologically annotated 38 hours of data. In this paper, we first outline the design of the CEJC including corpus size, recording methods, and annotations to be included in the corpus. Then, we conduct a preliminary analysis on some linguistic aspects of the corpus, based on the morphologically annotated data, showing that the CEJC captures the diversity of real-life conversations.

**Keywords:** Corpus of everyday Japanese conversation, corpus design, corpus analysis, linguistic aspects, morphological annotation

## 1. Introduction

Conversations emerge in various ways in everyday life. To capture the diversity of real-life conversations, we have to collect a variety of conversations occurring in natural settings in our daily life. Yet, most of the corpora constructed so far have targeted conversations in artificially created settings in terms of topics and recording situations, such as task-oriented, experimental dialogs or chats among participants recruited for recording purpose.

To overcome these undesirable circumstances, we started the compilation of a large-scale corpus of everyday Japanese conversation, the *Corpus of Everyday Japanese Conversation*, CEJC (Koiso et al., 2018). The CEJC is designed to contain various kinds of everyday conversations in a balanced manner so as to capture the diversity of everyday conversations and to observe natural conversational behavior. The CEJC targets conversations embedded in naturally occurring activities in daily life, without the exogenous intervention by researchers imposing topics or displacing the context of action (Mondada, 2012). Since the start of the project in 2016, we have collected more than 400 hours of conversations and compiled 94 hours out of them in the CEJC, corresponding to about a half of the target size of the entire corpus. Among them, 38 hours of data have also been morphologically analyzed.

In this paper, we first outline the design of the CEJC including corpus size, recording methods, and annotations to be included in the corpus. Then, we conduct a preliminary analysis on some linguistic aspects of the corpus, based on the morphologically annotated data, showing that the CEJC captures the diversity of real-life conversations.

## 2. Design of the CEJC

**Corpus size** We plan to publish more than 200 hours of conversations. Based on the data we have recorded, transcribed, and morphologically annotated so far, the total number of words, conversations, and conversants are estimated at 2.1 million words, 400 conversations, and a total of 1200 conversants, including 600 different participants.

**Recording** In order to record convesations embedded in naturally occurring activities in daily situations, we mainly adopt a recording method called the *individual-based* method. In this method, we recruit 40 informants balanced in terms of sex and age (man/woman $\times$ 20s/30s/40s/50s/over 60 $\times$ 4 informants), provide them with portable recording devices for approximately two to three months, and have them record about 15 hours of conversations in their daily activities. The informant carries portable recording devices and record his/her everyday activities in a variety of situations such as at home, at a restaurant, and outdoors. The project members do not mediate their field recordings. About four to five hours of conversations, among 15 hours, per informant, i.e., total of about 180 hours, are selected for the CEJC by taking into account the balance of conversation variations, quality of recorded data, and legal and ethical issues.

In order to precisely understand the mechanism of our real-life social conduct, not only audio but also video data are collected and published. Two types of compact action cameras, Kodak PIXPRO SP360 4K and GoPro Hero 3+, are used when recording indoors (Koiso et al., 2016a). Figure 1 shows video images of a conversation between a customer and a barber at a barbershop.

Figure 1: Video images of a conversation between a customer and a barber at a barbershop.

**Annotation** The speech is manually transcribed, and two types of POS information, short–unit word and long–unit word, are automatically annotated. The *Core* data set, consisting of 20 hours of conversations, is designed to contain manually corrected POS information and manually annotated dependency structures, dialog acts, and intonation labels.

See Koiso et al. (2018) for the details of the design of the CEJC.

## 3. Preliminary Analysis

So far, we have compiled 94 hours of conversations in the CEJC. Based on this data, Koiso et al. (2018) provided a preliminary evaluation on the issue of balancedness by reference to the survey results of everyday conversational behavior described in Koiso et al. (2016b). In this paper, on the other hand, we conduct a preliminarily analysis on some linguistic aspects of the CEJC based on the morphologically annotated data. As a result, we show that the CEJC captures the diversity of real-life conversations.
The data analyzed here amounts to about 38 hours, which consists of 81 conversations, and contains a total of 159 conversants, including 275 different participants.

### 3.1. Variance in the number of words

In this section, we focus on the number of words in a unit of time. Since the CEJC is designed to target conversations in various activities that occur naturally in daily life, participants may, on one occasion, be involved in an activity in which they do not have to speak all the time such as cooking, and, on another occasion, engage in an activity that is mainly conducted verbally such as business meetings. Therefore, the diversity of activities involved in the CEJC may result in a wider range of numbers of words compared to corpora recorded in restricted situations.
We compared the distribution of the words per minute (WPMs) in the CEJC with those in the *Chiba Three-party Conversation Corpus* (Chiba3Party) (Den and Enomoto, 2007) and in casual conversations included in the *Corpus of Spontaneous Japanese* (CSJ) (Maekawa, 2004). Conversational data in the Chiba3Party and the CSJ were recorded in rather artificial settings. The Chiba3Party consists of 12 conversations among university friends recruited for recording purpose, and amounts to 2 hours. Casual conversations in the CSJ includes 16 dyadic conversations between recruited interviewers and informants, and amounts to 1 hour. The WPM for each conversation in each corpus was calculated as the number of words in the conversation divided by the duration, in minute, of the conversation.

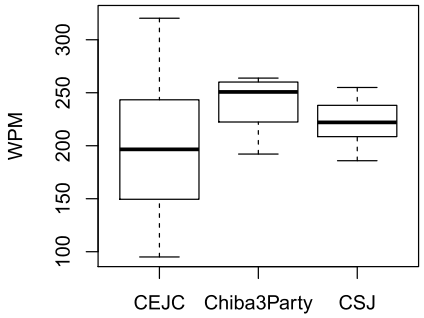


Figure 2: Distributions of the words per minute in the CEJC, the Chiba3Party, and the CSJ

Figure 2 shows the distributions of the WPMs in the three corpora. It is found that the mean of the WPM is smaller and the variance of the WPM is larger in the CEJC than in the Chiba3Party and the CSJ. In the Chiba3Party and the CSJ, the participants were recruited for recording purpose, and, thus, might feel pressure to continue speaking, resulting in greater and condensed WPMs. In the CEJC, by contrast, participants often engage in other activities while having a conversation, such as eating, cleaning, cooking, and doing homework. In these situations, they sometimes concentrate on non-conversational activities without talking. The larger variance of the WPM in the result might suggest that the CEJC captures the diversity of real-life conversation embedded in naturally occurring activities.

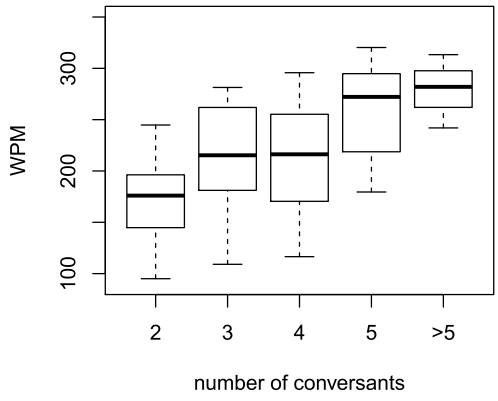


Figure 3: Distributions of the words per minute according to the number of conversants

There may be, however, another factor that could account for the larger variance in the CEJC. The numbers of conversants in the CSJ and the Chiba3Party are fixed at two and three, respectively, while the CEJC has much variety in the number of conversants. Since the number of conversants may have an influence on the variance of the WPM, we next investigate the distributions of the WPMs in the CEJC according to the number of conversants.

Figure 3 shows the result. It is found that the WPM increases when the number of conversants increases. This tendency would predict that the WPM in the CEJC, on average, is larger than those of the CSJ and the Chiba3Party, because the CEJC contains conversations with more than three conversants, which are never included in the CSJ or the Chiba3Party. What we saw in Figure2, however, is opposite to this prediction; the median of the WPM is smaller in the CEJC than in the CSJ and the Chiba3Party. This result reinforces the argument that conversants in the CEJC sometimes concentrate on non-conversational activities without talking, which is one of the key features of real-life conversations.

Increase in the WPM with increasing number of conversants in Figure 3 might be caused by response tokens, such as "hai" and "un," produced by listeners during a speaker's utterance, since there would be more response tokens in a conversation with more conversants. The tendency, however, is retained even after response tokens, as well as other interjections, have been removed from the data. One reason for this might be that when there are more than three conversants, they can be divided into two or more groups to have a separate talk (Sacks et al., 1974). This result suggests that the number of conversants could affect conversational structures.

Conversations with four or more conversants account for about 35% of the data collected so far, and the whole corpus maintains the balancedness of the numbers of conversants by reference to the survey results of everyday conversational behavior (Koiso et al., 2018), i.e., at least 24% for more-than-three-party conversations.

## 3.2. Proportion of polite forms

In Japanese, polite and non-polite forms are expected to be distinguished in a proper way, according to the conversants' relationships and speaking situations. If the CEJC captures a broad range of conversational situations in real life, various patterns of the use of polite and non-polite forms between conversants would be observed. In this section, we investigate how conversants use polite and non-polite forms differently depending on their relationships.

In this analysis, we used 16867 utterance units (Den et al., 2010) containing verbs or adjectives as their main predicates. Utterance units with auxiliary verbs "desu," "masu," and "gozai-masu" in their predicates were defined as polite forms. We also distinguished utterance units with and without final particles "ne," "yo," "na," "sa," "wa," and "no," because some studies pointed out that polite forms with final particles show lower degree of formality (Ijuin, 2004; Ogi, 2014; Satake, 2016). Utterance units for which the addressee's relation to the speaker cannot be uniquely determined were excluded from the analysis; for instance, in a conversation with a teacher and two or more students, the teacher's utterances were included in the analysis, but the students' utterances were excluded because they may be addressed to either the teacher or other students, i.e., friends.

Figure 4 shows the proportions of polite and non-polite forms according to the conversants' relationship, i.e., the addressee's relation to the speaker. First, we take a look at the general tendency of the use of polite and non-polite forms. It is found that the proportion of the use of polite forms is the highest when talking with customers, and the second highest when talking with business connections. This suggests that people tend to speak more formally when they are talking in business situations. In other cases, familiarity and power relation among participants may influence on how politely they speak. For instance, people use more polite forms to their acquaintances than to their family members and friends. Similarly, senior and similar-age colleagues at an office are more likely to be spoken in a polite way than junior colleagues, and teachers are more likely to be addressed politely than students, presumably because of asymmetric power relations.[1]

Let us, next, glance at characteristics of utterance units with and without final particles. Overall, final particles are employed more often with non-polite forms, with a notable exception of conversation between teachers and students, which shows relatively high likelihood of the use of polite forms with final particles. In the data observed in this analysis, there were only four teacher–student conversations, and they were all informal chats outside the classroom. This may lead more use of polite forms with final particles, which is less formal.

Now, we look at our data from another perspective. Focusing on a particular conversant, Informant A, we explored the relation between how often polite forms and final particles are used and with whom the informant is talking (Figure 5). In the case of Informant A, a male professional officer in his 30s, his "family" category was

[1]In the cultural context of Japan, there is a rather clearly asymmetrical power relation between seniors and juniors. This causes juniors usually use polite forms to their seniors, which may be unusual in Western cultures. Note that demands on the use of polite forms also depend on the company's/organization's custom.

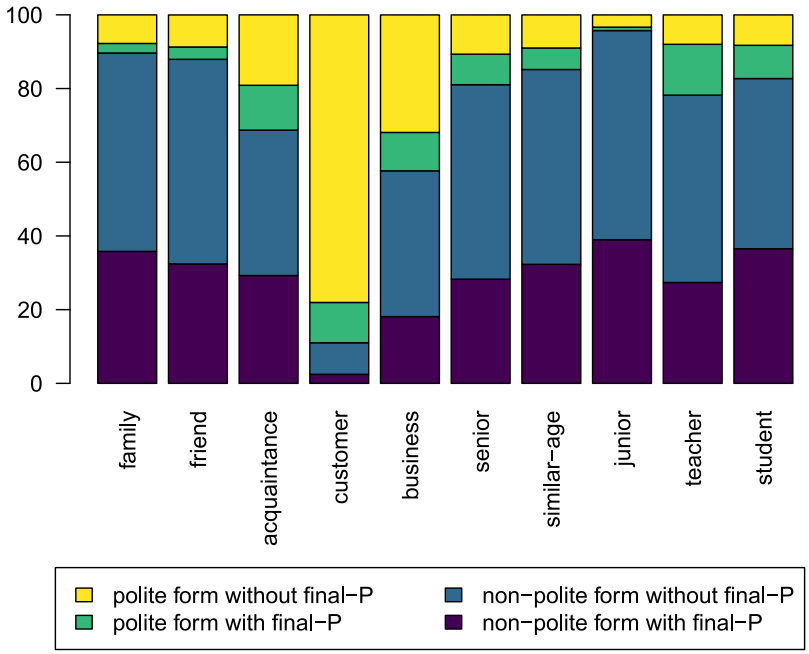


Figure 4: Usage of polite and non-polite forms according to the conversants' relationship

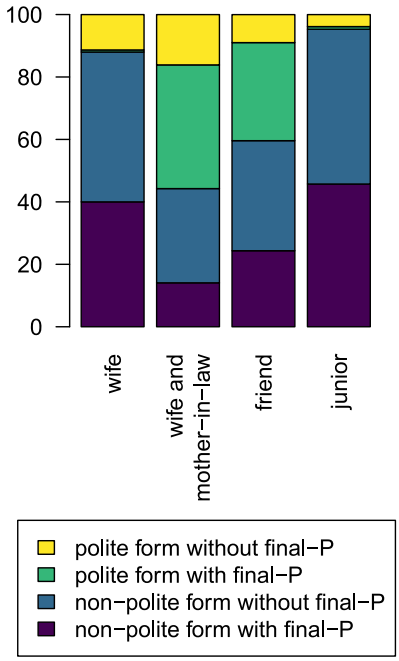


Figure 5: Informant A's usage of polite and non-polite forms according to the relation to the addressees

divided into "wife only" and "wife and mother-in-law (his wife's mother)." It is found that Informant A tends to use far more non-polite forms when talking with his wife and junior colleagues than other people. Polite forms are less likely to be employed in very close relationship or when he is in power. On the other hand, he speaks using polite forms when talking with his mother-in-law. Interestingly, in quite a few cases when talking with his mother-in-law, he uses more polite forms with final particles than those without final particles. This suggests that he downgrades formality so that he avoids a sense of unfamiliarity caused by being too polite.

In this way, the characteristics of the use of polite and non-polite forms reveal the diversity of the CEJC in terms of relationships among participants, which is another key feature of real-life conversations.

## 4. Conclusions

In this paper, we outlined the design of the *Corpus of Everyday Japanese Conversation*, which we have been constructing since 2016. We also presented some results on a preliminary analysis based on the morphologically annotated data from linguistic aspects, such as the variance in the number of words according to the number of conversants and the proportion of polite forms according to conversants' relationships. We pointed out that the CEJC covers conversations in various situations in real life. We plan to publish a part of the CEJC, about 50 hours, on a trial basis in 2018, and the entirety in 2022.

## 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

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., Koiso, H., Maruyama, T., Maekawa, K., Takanashi, K., Enomoto, M., and Yoshida, N. (2010). Two-level annotation of utterance-units in Japanese dialogs: An empirically emerged scheme. In *Proceedings of LREC 2010*, pages 2103–2110, Valletta, Malta.

Ijuin, I. (2004). A comparison of speech style adaptation between native situations and contact situations (in Japanese). *The Japanese Journal of Language in Society*, 6(2):12–26.

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

Mondada, L. (2012). The conversation analytic approach to data collection. In J. Sidnell and T. Stivers, editors, *The handbook of conversation analysis*, pages 32–56. Wiley-Blackwell, Hoboken, NJ.

Ogi, N. (2014). Language and an expression of identities: Japanese sentence-final particles *ne* and *na*. *Journal of Pragmatics*, 64:72–84.

Sacks, H., Schegloff, E. A., and Jefferson, G. (1974). A simplest systematics for the organization of turn-taking for conversation. *Language*, 50(4):696–735.

Satake, K. (2016). Nichijo-danwa-ni mirareru keigo shiyo-no jittai (An empirical study on honorifics in everyday discourse) (in Japanese). In Gendai Nihongo Kenkyukai, editor, *Danwa-shiryo: Nichijo-seikatsu-no kotoba (Discourse material: Language in everyday life)*, pages 191–212. Hitsuji Shobo, Tokyo.