

Survey of Conversational Behavior: Towards the Design of a Balanced Corpus of Everyday Japanese Conversation

Hanae Koiso[†] Tomoyuki Tsuchiya[‡] Ryoko Watanabe[†]
Daisuke Yokomori[‡] Masao Aizawa[†] Yasuharu Den^{§†}

[†]National Institute for Japanese Language and Linguistics, Japan

[‡]Faculty of Languages and Cultures, Kyushu University, Japan

[§]Faculty of Letters, Chiba University, Japan

Abstract

In 2016, we set about building a large-scale corpus of everyday Japanese conversation—a collection of conversations embedded in naturally occurring activities in daily life. We will collect more than 200 hours of recordings over six years, publishing the corpus in 2022. To construct such a huge corpus, we have conducted a pilot project, one of whose purposes is to establish a corpus design for collecting various kinds of everyday conversations in a balanced manner. For this purpose, we conducted a survey of everyday conversational behavior, with about 250 adults, in order to reveal how diverse our everyday conversational behavior is and to build an empirical foundation for corpus design. The questionnaire included when, where, how long, with whom, and in what kind of activity informants were engaged in conversations. We found that ordinary conversations show the following tendencies: i) they mainly consist of chats, business talks, and consultations; ii) in general, the number of participants is small and the duration of the conversation is short; iii) many conversations are conducted in private places such as homes, as well as in public places such as offices and schools; and iv) some questionnaire items are related to each other. This paper describes an overview of this survey study, and then discusses how to design a large-scale corpus of everyday Japanese conversation on this basis.

Keywords: corpus of everyday Japanese conversation, corpus design, survey of everyday conversational behavior

1. Introduction

Since everyday conversation is a foundation of our social life, it is important to describe the characteristics of conversational language and clarify the mechanisms of conversational interaction. In order to capture the diversity of everyday conversations and to observe natural conversational behavior in our daily life, we require corpora containing various kinds of everyday conversations. Although several corpora of Japanese conversations have been developed, most of them are biased towards conversations among friends or families on the telephone or in artificially created settings, and there has been no corpus that covers a diversity of everyday conversations (see Table 1).

Considering this situation, in 2016, we set about building a large-scale corpus of everyday Japanese conversation in a balanced manner. To construct such a huge corpus, we have conducted a pilot project, whose purposes are i) to establish a corpus design for collecting various kinds of everyday conversations in a balanced manner, ii) to develop a methodology for recording naturally occurring conversations, and iii) to create a transcription system suitable for precisely and efficiently transcribing natural conversations.

As for the first issue, we conducted a survey of everyday conversational behavior, with about 250 Japanese adults, last year in order to reveal how diverse our everyday conversational behavior is and to build an empirical foundation for corpus design. The questionnaire included when, where, how long, with whom, and in what kind of activity informants were engaged in conversations.

This paper describes an overview of this survey study, and then discusses how to design a large-scale corpus of everyday Japanese conversation on this basis.

2. Survey Method

In this study, conversational behavior was surveyed from three perspectives: conversation attributes, conversational situation attributes, and informant attributes (see Table 2), which were defined in reference to previous surveys conducted in Japan (National Language Research Institute, 1980; Hata, 1983; National Language Research Institute, 1987; Japan Broadcasting Corporation, 2010).

The survey was performed according to the following procedures.

1. Informants were collected from native speakers of Japanese living in the Greater Tokyo Area (20 to 25 men/women from each of the following 5 age ranges: 20–29, 30–39, 40–49, 50–59, over 60).
2. Survey materials were sent to informants via post in advance.
3. Informants could freely choose two weekdays/workdays and one weekend/day-off, carrying out the survey within two weeks after the arrival of the materials.
4. On the survey days, informants carried around, from waking up until going to sleep, a survey note, on which they wrote an overview of all conversations they had as well as the answers to the survey questions regarding the attributes of their conversations and conversational situations.
5. After completing the survey, informants returned to the survey organizers the survey notes and a questionnaire on informant attributes including his/her sex, age, and occupation.

Table 1: Major existing corpora of Japanese conversations

Corpus Name	Size	Contents
Multilingual Corpus of Spoken Language by Basic Transcription System (BTS)	294 conversations 66 hours	chats among friends, professor-student mentoring, telephone conversations, etc. (audio files available only for some portion)
Sakura Corpus	18 conversations	chats among four undergraduate students (topics assigned)
Chiba Three-Party Conversation Corpus	12 conversations 2 hours	chats among three undergraduate/graduate students on campus (initial topics assigned)
CALL HOME Japanese	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.
Meidai Conversational Corpus	161 speakers 100 hours	chats among friends (audio files unavailable)
Women's Language at the Workplace Men's Language at the Workplace	21 speakers each	natural conversations in formal and informal situations at the workplace (audio files unavailable)

Table 2: Survey items

Attribute	Explanation	Choices
Conversation attributes		
Form	Conversation form	chat, business talk/consultation, meeting/conference, class/lesson/lecture
Length	Conversation length	below 5 minutes, 5–15 minutes, 15–30 minutes, 30 minutes–1 hour, 1–2 hours, 2–5 hours, 5–10 hours
Relationship	Relationship between the informant and his/her interlocutors	family member, relative, teacher–student, co-worker, friend/acquaintance, public/commercial service, familiar/unfamiliar person
Number of interlocutors	Number of interlocutors by relationship	(number)
Remote conversation	Voice or video conversation on the phone/Internet	(Choose if applicable)
Non-native interlocutor	Conversation including a non-native interlocutor	(Choose if applicable)
Foreign language	Conversation in/including a foreign language	(Choose if applicable)
Conversational situation attributes		
Time period	Time period in which the conversation was carried out	morning, afternoon, evening
Place	Place in which the conversation was carried out	home, workplace/school, public/commercial facility, transport facility, other indoor place, other outdoor place
Activity	Activity in which the conversation was embedded	eating, housework/chore, personal task, medical treatment, work/schoolwork, after-work/after-school activity, communal activity, leisure activity, spending time with friend, transfer, rest
Informant attributes		
Sex	Sex of the informant	male, female
Age	Age of the informant	20–29, 30–39, 40–49, 50–59, over 60
Occupation	Occupation of the informant	office worker, self-employed worker, part-time worker, student, full-time housewife, unemployed person

3. Results

3.1. Overview of the Results

Two hundreds and forty-three informants participated in the survey, who engaged in 9,272 conversations in total. Table 3 shows the distributions of answers to the survey items¹. The results are outlined below.

¹For some survey items, there were missing answers. The counts of answers in such items do not add up to 9,272, i.e., the total number of conversations.

1. Chats (62%) and business talks/consultations (32%) occur very frequently, while meetings/conferences (3%) and classes/lessons/lectures (2%) are quite rare.
2. Most conversations are short (under 30 minutes = 75%), and few are long (over one hour = 12%).
3. Most conversations involve a small number of people (two or three conversants = 75%), and few are composed of large party (more than 6 conversants = 10%).

Table 3: Distributions of answers to the survey items

Form			Non-native interlocutor		
chat	5,719	(61.9%)	no	9,188	(99.1%)
business talk/consultation	2,997	(32.4%)	yes	84	(0.9%)
meeting/conference	317	(3.4%)	Foreign language		
class/lesson/lecture	209	(2.3%)	no	9,200	(99.2%)
Length			yes	72	(0.8%)
below 5 minutes	2,716	(29.3%)	Time period		
5–15 minutes	2,640	(28.5%)	morning	3,019	(32.7%)
15–30 minutes	1,555	(16.8%)	afternoon	3,618	(39.2%)
30 minutes–1 hour	1,264	(13.7%)	evening	2,599	(28.1%)
1–2 hours	714	(7.7%)	Place		
2–5 hours	340	(3.7%)	home	3,237	(35.0%)
5–10 hours	26	(0.3%)	workplace/school	2,802	(30.3%)
Number of interlocutors			public/commercial facility	1,700	(18.4%)
1	5,257	(56.9%)	transport facility	472	(5.1%)
2	1,713	(18.5%)	other indoor place	411	(4.4%)
3	932	(10.1%)	other outdoor place	631	(6.8%)
4	392	(4.2%)	Activity		
5	254	(2.7%)	eating	1,566	(17.0%)
6	129	(1.4%)	housework/chore	1,588	(17.2%)
7	81	(0.9%)	personal task	501	(5.4%)
8	58	(0.6%)	medical treatment	85	(0.9%)
9	40	(0.4%)	work/schoolwork	2,331	(25.3%)
over 10	388	(4.2%)	after-work/after-school activity	118	(1.3%)
Remote conversation			communal activity	97	(1.1%)
no	8,368	(90.3%)	leisure activity	429	(4.7%)
yes	904	(9.7%)	spending time with friend	251	(2.7%)
			transfer	1,069	(11.6%)
			rest	1,187	(12.9%)

- Most conversations are conducted at home (35%) or at workplaces/schools (30%).
- Most conversations are carried out during work/schoolwork (25%), housework/chore (17%), eating (17%), or resting (13%).

3.2. Relationships among Survey Attributes

In this section, we focus on the relationships among the survey attributes. Before the analysis, we combined some survey answers using the following procedures:

- The associations among survey items were analyzed by using multiple correspondence analysis.
- Hierarchical cluster analysis of the (three-dimensional) scores assigned to survey answers was conducted.
- Some survey answers were combined based on the results of the cluster analysis².

²The combined answers are as follows: **Form:** meeting (meeting/conference + class/lesson/lecture); **Length:** 1–5 hours (1–2 hours + 2–5 hours); **Number of interlocutors:** 4–5 persons (4 persons + 5 persons), over 6 persons (6 persons + ... + over 10 persons); **Place:** outdoor/transport facility (other outdoor place + transport facility); **Activity:** housework/chore (housework/chore + personal task + medical treatment), work/schoolwork (work/schoolwork + after-work/after-school activity), leisure activity (leisure activity + spending time with friend).

Figure 1 shows the occurrence rates of conversation and conversational situation attributes with respect to the informant's occupation. As for the conversation attributes, regardless of occupation, most conversations i) consist of chats and business talks/consultations, ii) involve a small number of participants, and iii) last for only a short time³. There are, however, also differences according to occupation. For example, working people's conversations often take the form of a business meeting/conference or a business talk/consultation, while full-time housewives more often engage in chats than people with other occupations do.

With regard to the conversational situation attributes, there are many occupation-based differences. For example, conversations by full-time housewives are often carried out at home or in public/commercial facilities during engaging in housework/chore; working people often have conversations at their workplaces during engaging in work; and students often have conversations not only during studying at schools but also during resting or transferring.

Figure 2 shows associations between conversation form and place with respect to the activity in which the conversation is embedded. The width and the height of a bar indicate the relative frequencies of a form and a place categories, respectively, and the area shows the relative frequency of a "form × place" combination.

³The same tendency can be seen when considered relative to sex and age.

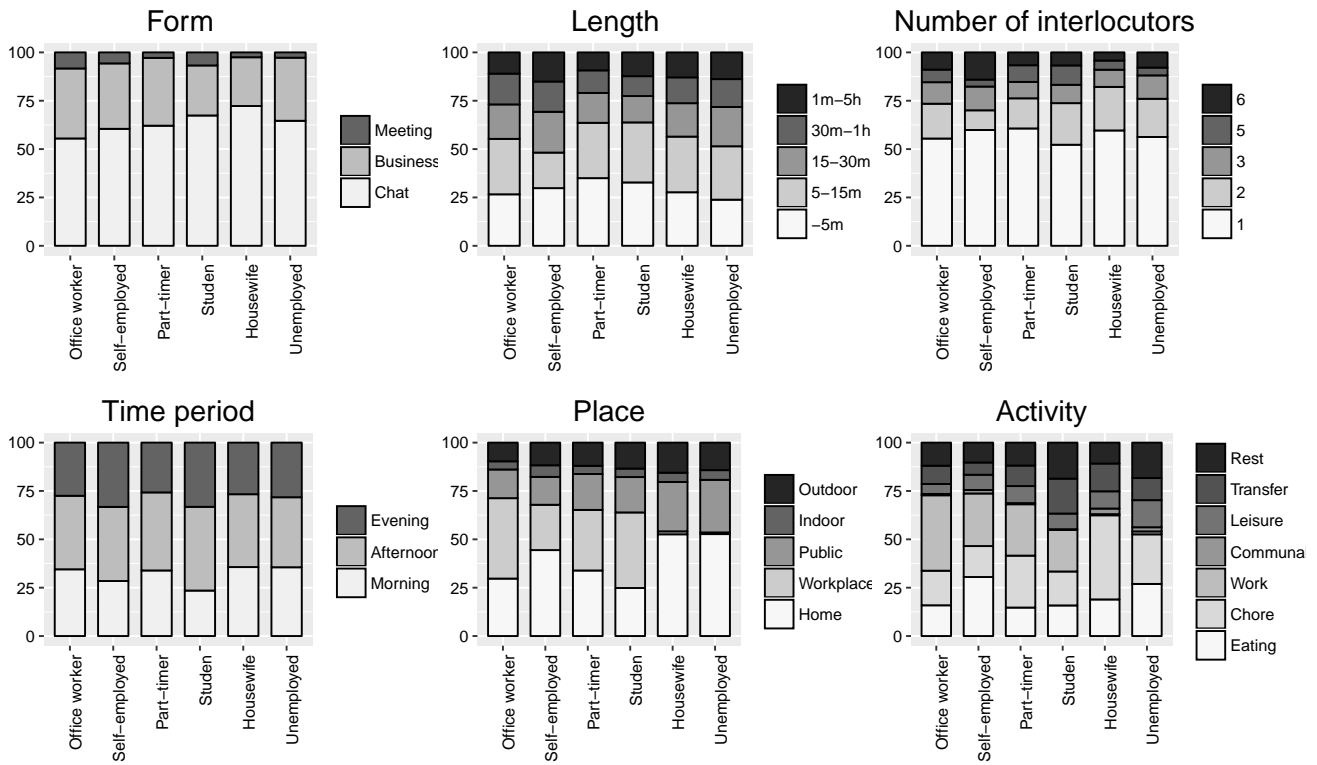


Figure 1: Occurrence rates of conversation and conversational situation attributes with respect to occupation

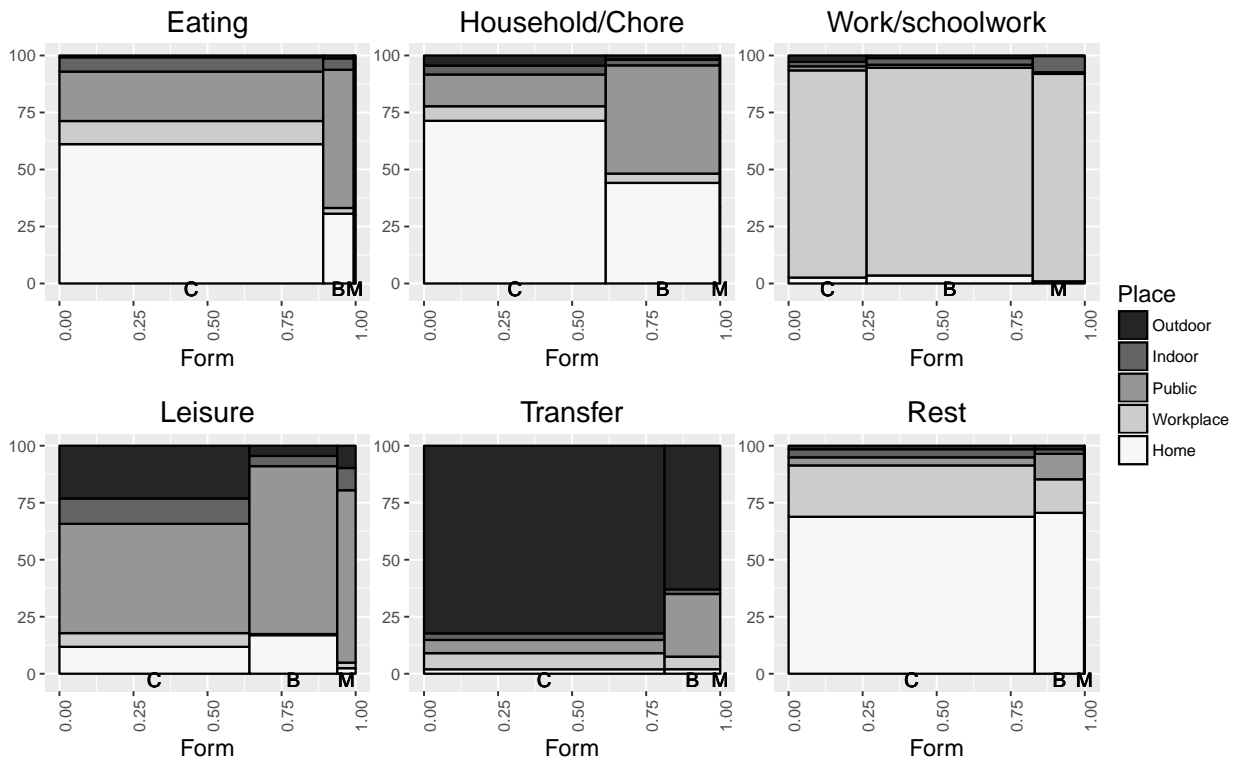


Figure 2: Associations between conversation form and place with respect to activity (X-axis labels: [C] Chat, [B] Business talk/consultation, [M] meetings/conferences)

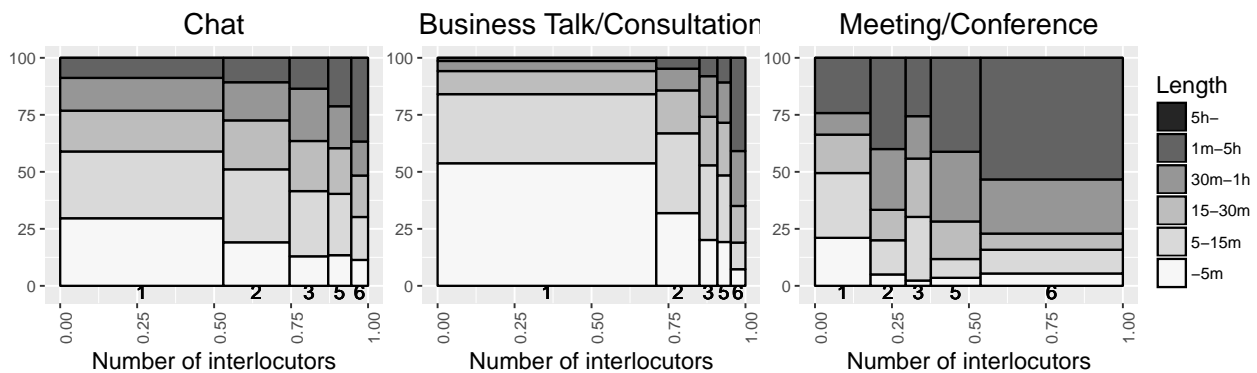


Figure 3: Associations between the number of interlocutors and conversation length with respect to conversation form

The figure reveals the dependence of the conversation form on the activity. For example, when eating, transferring, or resting, chats occupy the great majority of cases; during activities like housework/chore and leisure, chats are the most frequent, followed by business talks/consultations; and while engaged in work or schoolwork, business talks/consultations are the most frequent, followed by business meetings/conferences.

Such dependence can also be found between the conversation place and the activity: conversations during eating or resting frequently take place at home; conversations during working or studying tend to occur at workplaces and schools; conversations while engaged in leisure activities or spending time with friends often take place at public/commercial facilities; and conversations during transferring is strongly related to place outside.

Considering the perspectives of the conversation form and place together with respect to the activity, it can further be seen i) that as for eating, chats are overwhelmingly frequent at home, while not only chats but also business talks/consultations occur frequently at public/commercial facilities; and ii) that in regard to housework/chore, chats and business talks/consultations with family members are frequent at home, while conversations with people like store employees at public/commercial facilities tend to take the form of business talk/consultation.

Figure 3 shows associations between the number of interlocutors and conversation length with respect to conversation form. It is found i) that business talks/consultations involve few conversants and are short in length—70% are dyadic (two-party) conversations and 80% are shorter than fifteen minutes, ii) that meetings/conferences and classes/lessons/lectures are composed of a relatively large number of people and last for long periods of time, and iii) that chats fall in between these two. Furthermore, for all conversation forms, conversation length increases as the number of conversants increases.

4. Toward a Corpus Design

4.1. Recording Method

In order to capture the diversity of everyday conversations and to observe natural conversational behavior in our daily life, one must record conversations that naturally occur in daily situations (Mondada, 2012). The British National

Corpus (BNC), constructed in the former half of the 1990s in the U.K., provides a methodology for such purpose. The BNC is comprised of one-hundred million British English words (Crowdy, 1995; Burnard and Aston, 1998). While the majority of it contains written language, approximately 10% of the words (ten million) are from spoken language. This spoken language part of the BNC is composed of the following two data groups:

Spoken demographic: Recorded with a portable tape recorder over the course of seven days by 124 informants who were chosen so as to avoid bias in terms of age, sex, social class, and region.

Spoken context-governed: Spoken language that many people listen to (e.g., broadcasts and lectures). Divided into four categories: educational, business, public/organizational, and leisure.

The first of the above BNC methods is suited for recording conversations embedded in naturally occurring activities. With their approach in mind, we decided to record everyday conversations using the following two methods⁴.

Individual-based method: We choose a set of informants balanced in terms of sex, age, etc., provide them portable recording devices (compact action cameras and IC recorders) for approximately one to two months, and have them record conversations in their daily activities. In principle, the project members do not mediate their field recordings.

Situation-specific method: We select specific situations in which recording based on the individual-based method is technically and/or ethically difficult, e.g., exchanges with store employees, meetings at workplaces, regional activities, public events, etc., and record conversations occurring in these situations. Although the project members coordinate recording settings, only conversations in these naturally occurring activities are recorded.

⁴Koiso et al. (submitted) report how to record conversations based on the individual-based method.

Table 4: Constituent ratios with respect to conversation form, conversation place, and activity

Activity\Place	Form								
	chat			business talk/consultation			meeting/class		
	private	public	other	private	public	other	private	public	other
eating/resting	30%	5%	10%	5%	5%		5%		
work/study	15%	10%	5%	15%	40%	20%	0%	70%	10%
other		5%	20%	0%	0%	15%	0%	0%	15%

4.2. Constituent Ratio of Conversation and Conversational Situation Attributes

By reference to the constituent ratios derived from the survey results, we now draft a corpus design for collecting various kinds of everyday conversations in a balanced manner.

Because it is expected that conversations recorded by using the individual-based method consist mostly of chats and business talks/consultations, we first estimate the ratio with respect to the conversation form. From the survey results in Section 3.1., an approximate proportion would be chat: 60%, business talk/consultation: 30%, and meeting/conference or class/lesson/lecture: 10%.

The next issue is a breakdown within each conversation form. The survey results show that there is a strong association among the number of conversants, conversation length, and conversation form (Figure 3). This suggests that if conversation forms are well-balanced, the number of conversants and conversation length would naturally be distributed in a balanced manner as well. Therefore, we do not care the number of conversants and conversation length in the corpus design.

Next, we focus on place and activity. As pointed out in Section 3.1., most conversations are conducted at homes (35%) or workplaces/schools (30%). Thus, we classify places into three categories: i) private places, ii) public places, and iii) the others. As for activity, most conversations are carried out during work, schoolwork, or housework/chore (42%) and eating or resting (30%). We, thus, classify activities into three categories: i) eating/resting, ii) work/study, and iii) the others.

Based on these observations, we derived the constituent ratio of the conversation form, the conversation place, and the activity as in Table 4.

5. Concluding Remarks

In this paper, we reported our survey of everyday conversational behavior, which was conducted in order to reveal how diverse our everyday conversational behavior is and to build an empirical foundation for corpus design. Based on the survey results, we discussed how to design a balanced corpus of everyday Japanese conversation.

The constituent ratios of various conversational attributes presented in Table 4 were calculated only considering the results of the survey. We need to finalize the corpus design by taking into account technical and ethical issues related to recording in daily situations (Koiso et al., submitted).

In our project, more than 200 hours of conversations will be recorded. The recorded speech is precisely transcribed and

is annotated with morphological information, dependency structure, utterance boundary, dialog act, and so on. The corpus is planned to be published in 2022. It is expected that our corpus will greatly contribute to various research fields including linguistics, conversation analysis, psychology, cognitive science, Japanese language teaching, speech processing, and social robotics.

Language and behavior change with the times. In the future, our conversation corpus will also be a precious record to know our everyday language and conversational behavior in the early part of the 21st century. It is a significant role for researchers to record and preserve a diversity of everyday conversation that mirrors our culture.

6. Acknowledgements

This work is supported by Grant-in-Aid for Collaborative Research Project of NINJAL “A multifaceted study of spoken language using a large-scale corpus of everyday Japanese conversation” led by Hanae Koiso.

7. References

- Burnard, L. and Aston, G. (1998). *The BNC Handbook*. Edinburgh University Press, Edinburgh, U.K.
- Crowdy, S. (1995). The BNC spoken corpus. In Leech, G., Myers, G., and Thomas, J., editors, *Spoken English on computer: Transcription, mark-up and application*, pages 224–235. Longman, Harlow, U.K.
- Hata, H. (1983). Situation and language (in Japanese). *Kokugogaku*, 133:55–68.
- Japan Broadcasting Corporation. (2010). *Survey report of national living hours in 2010* (in Japanese). NHK Broadcasting Culture Research Institute.
- Koiso, H., Tanaka, Y., Watanabe, R., and Den, Y. (submitted). A large-scale corpus of everyday Japanese conversation: On methodology for recording naturally occurring conversations. In *Proceedings of the LREC 2016 workshop on casual talk among humans and machines*.
- Mondada, L. (2012). The conversation analytic approach to data collection. In Sidnell, J. and Stivers, T., editors, *The Handbook of Conversation Analysis*, pages 32–56. Wiley-Blackwell, Hoboken, NJ.
- National Language Research Institute. (1980). *The reality of the spoken language of the educated class in Japan* (in Japanese). National Language Research Institute.
- National Language Research Institute. (1987). *Aspects of conversational behaviour: An analysis based on round-table talk database* (in Japanese). Sanseido.