Construction of the *Corpus of Everyday Japanese Conversation*: An Interim Report

Hanae Koiso[†], Yasuharu Den[‡,†], Yuriko Iseki[†], Wakako Kashino[†], Yoshiko Kawabata[†], Ken'ya Nishikawa[†], Yayoi Tanaka[†], Yasuyuki Usuda[†]

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

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

Abstract

In 2016, we launched a new corpus project in which we are building a large-scale corpus of everyday Japanese conversation in a balanced manner, aiming at exploring characteristics of conversations in contemporary Japanese through multiple approaches. The corpus targets various kinds of naturally occurring conversations in daily situations, such as conversations during dinner with the family at home, meetings with colleagues at work, and conversations while driving. In this paper, we first introduce an overview of the corpus, including corpus size, conversation variations, recording methods, structure of the corpus, and annotations to be included in the corpus. Next, we report on the current stage of the development of the corpus and legal and ethical issues discussed so far. Then we present some results of the preliminary evaluation of the data being collected. We focus on whether or not the 94 hours of conversations collected so far vary in a balanced manner by reference to the survey results of everyday conversational behavior that we conducted previously to build an empirical foundation for the corpus design. We will publish the whole corpus in 2022, consisting of more than 200 hours of recordings.

Keywords: Corpus of everyday Japanese conversation, corpus design, legal and ethical issues, corpus evaluation

1. Introduction

Everyday conversation is the most basic form of human communication. In order to understand our diverse and situated interactional behavior, it is needed to collect and analyze various kinds of conversations in our daily life. Although several corpora of Japanese conversations have been developed, most of them are biased in terms of speaker attributes and situations, mainly targeting conversations in experimental settings, such as map task dialogs, and artificial situations, such as chats among university students recruited for recording purposes. There are few corpora of Japanese conversations that covers real situations in daily life.[1]

In 2016, we launched a new corpus project, in which we are building a large-scale corpus of everyday Japanese conversation, the *Corpus of Everyday Japanese Conversation*, CEJC. The main features of the CEJC are i) that we target conversations embedded in naturally occurring activities in daily life, without the exogenous intervention of researchers imposing topics or displacing the context of action (Mondada, 2012); ii) that we collect various kinds of everyday conversations in a balanced manner so as to capture the diversity of everyday conversations and to observe natural conversational behavior in our daily life; and iii) that we collect and publish not only audio but also video data in order to precisely understand the mechanism of our real-life social behavior.

In this paper, we first introduce an overview of the corpus, including corpus size, conversation variations, recording methods, structure of the corpus, and annotations to be included in the corpus. Next, we report on the current stage of the development of the corpus development and legal and ethical issues discussed so far. Then we present some results on the preliminary evaluation of the data being collected.

[1] For corpora of other languages that cover everyday situations, see e.g., Burnard and Aston (1998) and Nelleke (2000).

2. Corpus Design

2.1. Corpus size

We plan to publish more than 200 hours of conversations. Based on the data we have recorded and transcribed so far, the total number of words, conversations, and conversants are estimated at 2.1 million words (short-unit words, see below), 400 conversations, and a total of 1200 conversants, including 600 different participants.

2.2. Conversation variation

The CEJC will contain various kinds of everyday conversations in a balanced manner. To estimate distributions of various conversational attributes in our daily life, we conducted a survey of everyday conversational behavior with about 250 Japanese adult informants (Koiso et al., 2016b). The questionnaire included when, where, how long, with whom, and during what kind of activity informants were engaged in conversations in their daily life. Based on the results, we derived rough distributions of conversation forms, conversation places, and accompanying activities as a measure of the design of a balanced corpus. The survey result will be compared with the conversation data collected so far in Section 3.

Figure 1: Video images of a conversation between husband and wife while cooking at home. The left image was recorded by a Kodak PIXPRO SP360 4K camera located on the table, while the top- and bottom-right images were recorded by two GoPro cameras placed facing each other on the bookshelf and the sideboard. As for speech, the two conversants wear IC recorders (SONY ICD–SX734), and their voices were recorded with their own recorders. All conversants' voices were also recorded by another IC recorder located on the center of the table. Due to the restriction stated in the consent form, the faces of the participants are airbrushed for the protection of personal information in a printed material, although they are left intact in the video data to be published.

2.3. Recording method

In order to record various kinds of naturally occurring conversations in daily situations, we employ two methods, *individual-based* and *situation-specific* methods (Koiso et al., 2016a).

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 (compact action cameras and IC recorders) for approximately two to three months, and have them record about 15 hours of conversations in their daily activities. The informant him/herself carries portable recording devices and records his/her everyday activities in a variety of situations such as at home, at a restaurant, and outdoors. In principle, the project members do not mediate their field recordings. We developed the individual-based method by referring to the approach adopted for the demographically sampled part of the British National Corpus (Crowdy, 1995; Burnard and Aston, 1998). Figure 1 shows video images of a conversation between husband and wife while cooking at home.

About four to five hours of conversations, among 15 hours, per informant, i.e., a 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.

The informant also has to i) judge, for instance, whether recording is permitted where they are conversing, and get permission if necessary, ii) explain the purpose of the recording to other conversants, iii) obtain their consents to publish the recorded conversation, including video data, iv) have them fill in informant sheets including their date of birth, residence, birthplace, sex, occupation, and relationship to the informant, and v) note the recording date and time, an overview of the conversation, and the layout of the conversants and the recording devices.

Situation-specific method In addition to the individual-based method, we also use the situation-specific method to compensate for a lack, or shortage, of recordings in institutional settings, e.g., meetings at workplaces and exchanges with store employees, for which recording based on the individual-based method is technically and/or ethically difficult. In this method, we select specific situations and the recording staff set up a recording environment. Although the project members coordinate recording settings, only conversations in these naturally occurring activities are recorded. The size and types of conversations collected based on this method will be decided by referring to data collection status based on the individual-based method.

2.4. Structure of the CEJC

Figure 2 shows the layered structure of the CEJC. About 600 to 800 hours of conversations will be recorded, and

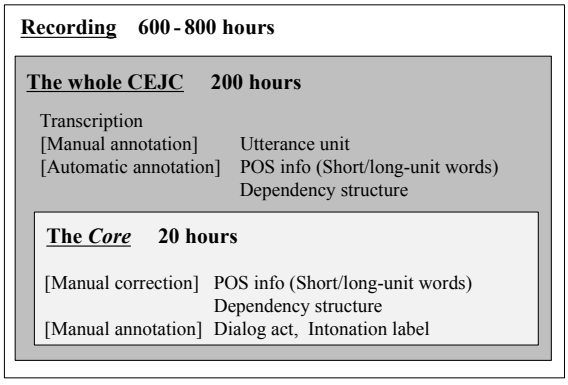

Figure 2: Layered structure of the CEJC

speakerID	startTime	endTime	text	note
IC01	2502.617	2503.920	(U Kono mae) nomikai doko de non da no.	(U xx) : transcription of questionable or inaudible talk
			Last time, where did you drink?	.: boundary of an utterance unit
IC03	2504.661	2505.651	Etto Akasaka.	
			Um, in the Akasaka area.	
IC04	2507.718	2508.495	Akasaka no	
			In Akasaka,	
IC03	2508.791	2509.744	(L)	(L) : laughter
IC04	2509.287	2510.202	ryootei.	
			at a fine dining restaurant?	
IC03	2510.912	2511.480	(L Iya iya).	(L xx) : speech while laughing
			No, no.	
IC01	2511.432	2512.185	Chigau chigau.	
			Different, different.	
IC01	2512.749	2513.451	Izakaya.	
			At a casual restaurant.	
IC03	2513.641	2514.236	(W Isakaya|Izakaya).	(W xx|yy) : 'xx' reduced or incorrect pronunciation
			At a casual restaurant.	'yy' supposed-to-be correct word
IC03	2515.464	2516.201	(U Futaherumo).	
			Futaherumo.	
IC03	2516.999	2519.648	Dooki no (D hi)(D fu)	(D xx): word fragment
			dooki to futari de non da gurai de.	
			I had a drink with the same-age, hi, fu,	
			a same-age peer.	

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

among them about 200 hours will be selected for the corpus. The whole corpus contains video and audio data, transcript, and four kinds of annotations to be described in § 2.6., three of which are automatically labeled. There is a subset of the corpus, named the *Core* data set, which consists of 20 hours of conversations, corresponding to 10% of the whole corpus, which includes six kinds of manually labeled, or corrected, annotations.

2.5. Transcription

The speech is divided into transcription units at the locations of perceptible pauses and the boundaries of utterance units (see below). Each unit is orthographically transcribed by hand with reference to video and audio data using ELAN,[2] and about 20 kinds of tags, which are 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), are inserted in the text. Figure 3 shows a sample transcript.

2.6. Annotation

In addition to transcripts, the following annotations are created:

Utterance Unit Utterance units are manually identified based on long utterance-units (Den et al., 2010), which are regarded as a basic unit for interaction and determined considering syntactic, pragmatic, and interactional aspects. The periods in the sample transcript in Figure 3 indicate utterance unit boundaries.

[2] https://tla.mpi.nl/tools/tla-tools/elan/

Two types of POS information Two different POS systems, short-unit word (SUW) and long-unit word (LUW), are adopted. Most SUWs are mono-morphemic 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. All the data are automatically analyzed, and those in the *Core* are manually corrected.

As for SUWs, the data are analyzed using UniDic, a dictionary developed for the POS annotation of the *Balanced Corpus of Contemporary Written Japanese* (Maekawa et al., 2014). The audio data are also automatically segmented into SUWs by means of forced alignment against morphologically-segmented texts, and those in the *Core* are manually corrected.

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

Dialog act The *Core* also contains dialog acts manually annotated according to an ISO-standard-based (ISO 24617-2, 2012) scheme extended to cover various kinds of sequence organizations observed in everyday conversation.

Intonation label Part of the *Core*, which is selected based on recording conditions and degrees of dialect, is manually labeled according to a simplified version of the intonation-labeling scheme, X–JToBI (Maekawa et al., 2002), which was developed for the CSJ.

Table 1: Attributes of informants (As of Jan. 15th, 2018)

Age	Sex		Total
	male	female	
20s	student*2 student*2 self-employed	student*2 student*2 office worker*1	6
30s	civil servant*2 self-employed*2 freelance	housewife*2 office worker*2 office worker self-employed	7
40s	office worker*2 freelance*2 office worker*1 teacher*1	office worker*2 office worker*2 part-time*2 freelance	8
50s	office worker*2 manager*2 self-employed	self-employed*2 office worker*2 office worker*1	6
over 60	volunteer*2 teacher*2	volunteer*2 office worker self-employed freelance*1	6
Total	15	18	33

*1: in the process of recording
*2: finished the data selection for the CEJC

3. Current stage of the corpus development

As of January 15th, 2018, 28 informants have finished recording and five are in the process of recording. Table 1 shows the attributes of these informants. In the cases of 20 out of the 28 informants who completed recording, we have selected conversations to be compiled into the CEJC. The selected data contains about 94 hours, corresponding to 47% of the whole, 210 conversations, and a total of 783 conversants, including 424 different participants.

4. Legal and ethical issues

A notable characteristic of the CEJC is that not only audio but also video data are collected and published. There have, however, been virtually no corpora that contain video recordings of everyday conversations, and guidelines on the release of such data have not been established. Based on a variety of data collected so far, we are discussing, with a lawyer specializing in copyright and portrait-right issues, how to deal with legal and ethical problems from the aspect of portrait-right, copyright, and the protection of personal information.

The video data often contains i) the faces of third parties who have not consented to publish their faces and ii) copyrighted works, such as TV programs and books.

When the faces of third parties are inside the scope of protection of portrait rights, those parts are airbrushed by means of an image effect. The faces of people performing common activities, not sensitive activities, in public places are regarded as outside the scope of protection of portrait rights, provided that the recordings will be used for research purposes and that their faces in themselves will not be the target of the research. When a short exchange between a third partiy and conversants who have agreed to have their faces published is transcribed, the faces of the third party is concealed (See Figure 4).

Figure 4: Video image which includes a face of a waitress talking with conversants at a restaurant. Although exchanges between the waitress and the main conversants are transcribed, the face of the waitress is concealed.

Figure 5: Video image which includes a television program. TV screen is not concealed.

If the use of copyrighted works included in the video data can be interpreted as incidental, i.e., an unexpected appearance as described in the copyright provisions for 'Disclosure of Photo or Image in which Copyrighted Work Appears,' they are not concealed (See Figure 5).

Personal information including conversants' names, affiliations, and individual identification information, as well as any parts of recordings for which conversants have not given their permission for publication are replaced by anonyms or turned letters in transcripts, and the corresponding regions of the audio files are made inaudible.

5. Preliminary evaluation of CEJC

In this section, using the 94 hours of conversations that have been compiled into the CEJC, we give a preliminary evaluation of the issue of balanced by reference to the survey results of everyday conversational behavior described in Koiso et al. (2016b).

The distributions of forms, places, activities, and numbers of conversants in the current data set, as well as the survey results, are shown in Figure 6.

As for the conversation form, slight differences are seen in that the ratio of chats in the current data set is about 11% higher than that in the survey result, while the ratio of business talks/consultation is 11% lower. Overall, however,

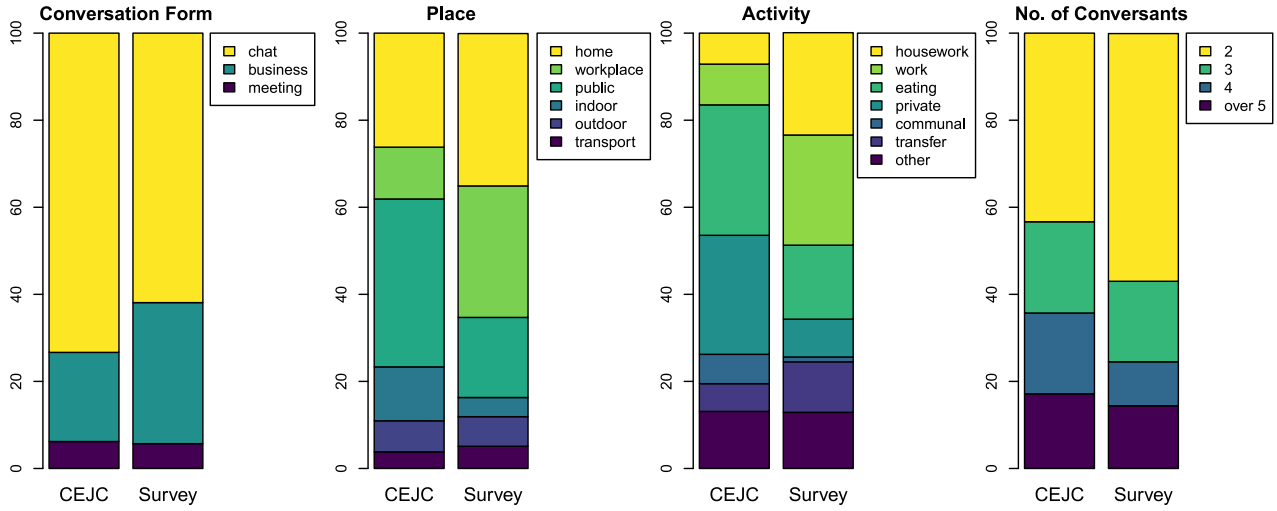

Figure 6: Distributions of conversation forms, places, activities, and numbers of conversants in the current data set and the survey results of conversational behavior

the current data set well varies in a balanced manner with reference to the survey results.

The same can be said for the number of conversants. In Figure 6, the ratio of dyadic (two-party) conversations in the current data set is about 14% lower than that in the survey result, while the ratio of conversations with more than three people is about 11% higher. Such a slight bias toward conversations with larger groups was intentionally introduced upon data selection in order to cover various kinds of conversations.

The distributions of places and activities show different tendencies between the current data set and the survey results. For example, the current data set contains more conversations at public/commercial facilities, such as restaurants and city halls, but fewer conversations at home, at school, or in the workplace than the survey results. In regard to activity, the current data set includes more conversations during leisure/communal activities and when spending time with friends but fewer conversations during housework, work, and schoolwork than the survey results.

The main reason why the current data set contains few conversations during work/schoolwork at workplaces/schools is that it is difficult to record such conversations based on the individual-based method. In the future, it will be necessary to reinforce such types of conversations based on the situation-specific method.

The current data set has considerably fewer conversations at home than the survey results, even though informants may have many opportunities for recording conversations at home. This is due to a bias in our sampling criteria. If we choose as many conversations at home as in the survey results, only similar types of conversations, such as conversations during dinner with the family at home, will be included in the corpus. We would rather select conversations with the family that were conducted outside, such as those in public/commercial facilities and at relatives' houses. This bias results in a decrease in the relative frequency of conversations at home.

Figure 7 shows the distributions of ages, sexes, and occupations of a total of 783 conversants, including 424 different participants involved in the current data set. It is found that conversants are balanced in terms of sex. By contrast, the figure shows that children under 20 years old, from elementary school students to high-school students, account only for 5 to 7% of the data and there are no high-school students at all. Since the individual-based recording method places a heavy responsibility on principal informants, such as dealing with various types of personal information, children under 20 are not recruited as informants. Children may participate in conversations when the principal informant invites them, but the possibility of recording conversations involving children depends highly on the composition of the informant's family. To solve this problem, we will select more informants who have children in their families.

6. Conclusions

In this paper, we first introduced an overview of the CEJC, including corpus size, conversation variations, recording methods, structure of the corpus, and annotations to be included in the corpus. Then we reported on the current stage of the corpus development and legal and ethical issues encountered so far. We also presented a preliminary evaluation of the data collected so far.

We focused on whether or not the 94 hours of conversations collected so far varies in a balanced manner by reference to the survey results of everyday conversational behavior. As for the conversation form and the number of conversants, the current data set varies in a balanced manner by reference to the survey results. By contrast, the current data set contains i) fewer conversations during work/schoolwork at workplaces/schools than the survey results, due to difficulty in recording such conversations using the individual-based method, ii) fewer conversations at home than the survey results, due to a bias in our sampling criteria, and iii) few conversations involving children under 20, due to the age restriction on informants. We will adopt the situation-specific method so as to compensate for these biases in the collected data.

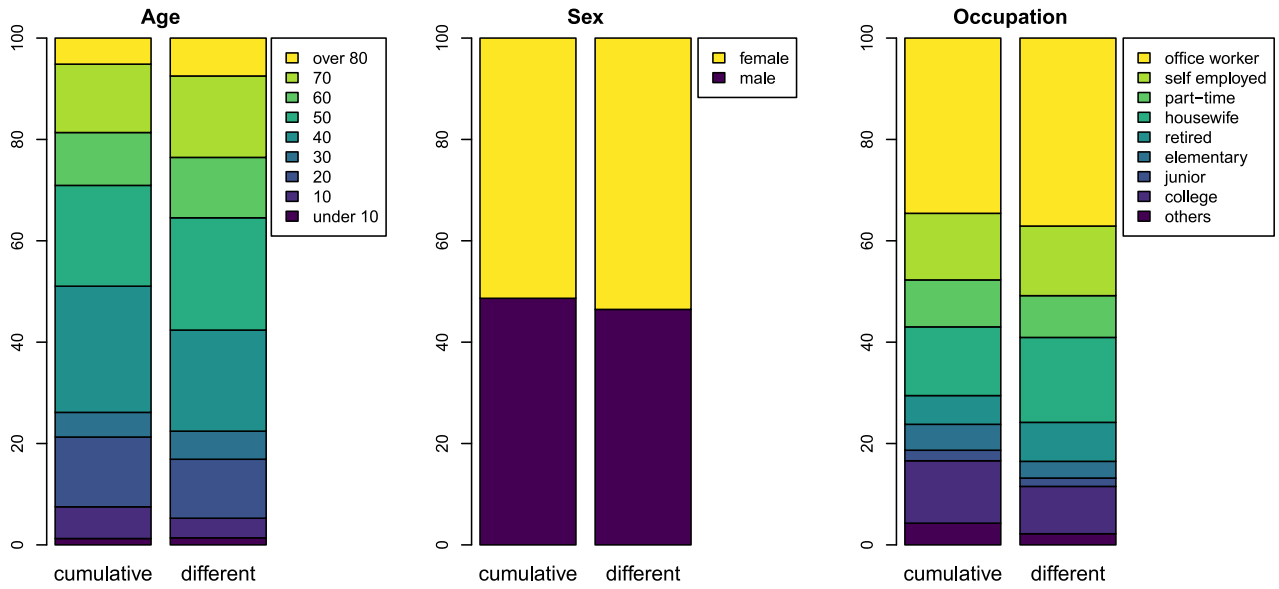

Figure 7: Distributions of ages, sexes, and occupations of a cumulative total of 783 conversants, including 424 different participants involved in the current data set

We plan to publish a part of the CEJC, about 50 hours, on a trial basis in 2018, and the entirety in 2022.

7. 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".

8. Bibliographical References

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, G. Myers, and J. Thomas, 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., 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.

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.

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 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.

Nelleke, O. (2000). Building a corpus of spoken Dutch. In P. Monachesi, editor, *Computational linguistics in the Netherlands 1999: Selected papers from the 10th CLIN meeting*, pages 147–158. Utrecht Institute of Linguistics OTS.