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

LREC 2022

Hanae Koiso*, Haruka Amatani*, Yasuharu Den+, Yuriko Iseki*, Yuichi Ishimoto*, Wakako Kashino*, Yoshiko Kawabata*, Ken'ya Nishikawa*, Yayoi Tanaka*, Yasuyuki Usuda*, and Yuka Watanabe*

* National Institute for Japanese Language and Linguistics, Japan + Chiba University, Japan

Introduction

- ✓ The *Corpus of Everyday Japanese Conversation* (CEJC)
  - Conversations enbedded in naturally occurring activities in daily life
  - Balanced to capture the diversity of our everyday lives
  - Including audio and video data
- ✓ The feature of the CEJC reported here:
  - Recording methods and devices
  - The structure of the corpus
  - The evaluation on conversants and conversation attributes

Recording methods

■ Individual-based method 185 hours

- ✓ Informants are balanced in terms of age and gender
- ✓ (About 50 people consist of five generations with various occupations)
- ✓ Video and audio are recorded by the informants in their naturally-occurred activities in their everyday lives (about 15-18 hours each)
- ✓ Conversations are selected due to proportion and ethical issues
  - 4-5 hours per informant (185 hours in total)

■ Situation-specific method 15 hours

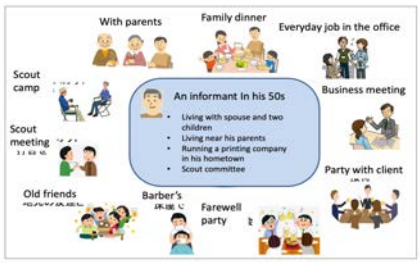
situations that are not obtained enough in the Individual-based method

- ✓ Meetings or conferences in the workplace 10 hours
- ✓ Conversation among teenagers 5 hours

Informants in Individual-based method

age	male			female		
	ID	Occupation	Duration	ID	Occupation	Duration
20s	T010	Student	4.2h	T009	Student	6.0h
	T006	Student	4.3h	K003	Student	4.4h
	T022	Teacher	3.7h	K009	Office worker	4.2h
	K007	Office worker	5.5h	K013	Office worker	4.0h
30s	T001	Freelance	5.6h	K001	Office worker	5.0h
	T005	Office worker	4.6h	T003	Homemaker	5.6h
	S002	Office worker	4.7h	K005	Freelance	5.4h
	K012	Office worker	3.1h	T008	Freelance	4.8h
40s	T016	Office worker	3.6h	C001	Office worker	4.5h
	T002	Freelance	4.8h	T011	Part-time worker	4.8h
	T019	Teacher	3.9h	K004	Part-time worker	5.0h
	T020	Office worker	5.0h	T014	Freelance	4.4h
50s	T015	Office worker	6.0h	C002	Office worker	4.2h
	S001	Office worker	4.6h	K002	Freelance	4.6h
	T024	Teacher	4.2h	K008	Freelance	4.6h
	T018	Teacher	4.2h	K011	Office worker	4.5h
Over 60	T013	Teacher	4.3h	T004	Homemaker	5.1h
	T007	Retired	5.8h	K006	Freelance	4.4h
	K010	Office worker	4.8h	T017	Office worker	4.2h
	T023	Retired	4.6h	T021	Freelance	4.3h

Conversations in the Individual-based method



GoPro Hero3+

Used for recording a normal flat image

Kodak PIXPRO SP360 4K 360-degree global view

Used for inside-out image of the conversation field

Sony ICD-SX734

All conversants wear IC recorders, recording their voices

All conversants' voices are recorded by another IC recorder on the center of the conversation space.

Sony ICD-SX1000

The size of CEJC

Total duration	200 hrs.
No. of conversation	577
No. of conversants (total)	1675
No. of conversants (diff.)	862
No. of words	2.4m

The structure of CEJC

The whole CEJC 200hrs.

Audio and Video data
Transcription
[Manual correction] POS info (Short-unit word, SUW)
[Automatic annotation] POS info (Long-unit word, LUW)

The *Core* 20hrs.

[Manual correction] POS info (SUW and LUW)
Dependency structure
[Manual annotation] Dialog act, Intonation label

Two types of POS info (SUW/LUW)

- ✓ SUW (short-unit word): a single or two mono-morphemic word
- ✓ LUW (long-unit word): multi-morphemic word

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		
つい	Verb	attach	つい	Particle	about
て	Particle	-te	て		

Dependency structure

- Dependency between *bunsetsu*, which consist of a content word and more than one functional word
- Extended version of BCCWJ-DepPara for speech

label	description
D	normal dependency
Z	sentence boundary
B	connected to the next part due to dependency
F	undetermined

Dialog act

manually annotated according to the ISO 24617-2 scheme extended to everyday conversations

■ Level 1: basic functions

Task Group	Inform/Self-speech/Question/Answer/Request…
Social Obligations Management Group	Greeting/Apology/Thanking…
Feedback Group	Positive/Negative
Others	Indeterminable/Not Applicable

■ Level 2: if applicable

Preliminary Group	Pre-Pre/Preliminary/Projected Action…
Settlement Group	Settlement
Repair Group	Repair Initiation/Repair
Discourse Structuring Group	Pre-Opening/Opening…

■ Dependency

Prospective	An action requires a certain type of following action
Retrospective	A following action voluntarily reacts to the previous action
External-prospective	The previous action does not appear on the transcript
External-retrospective	The previous action does not appear on the transcript

Intonation labeling

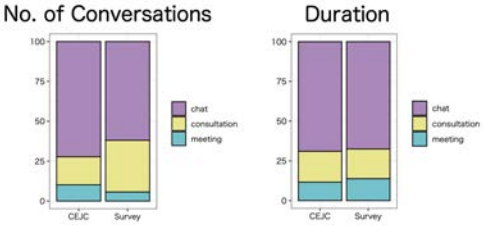
- Based on simplified version of X-JToBI for CSJ

Tier	Description	Example
Word	word boundary/word form/accent	ha'rukani
Break Index	type of intonation boundary	1 (word boundary) 2 (accent-phrase boundary) 3 (intonation-phrase boundary)
Tone	Phrase-final tones/Phrase-final boundary tones	L% (L%)H% (L%)HL% (L%)HLH% (L%)LHL%
Prominence	Tonal changes	PNLP FR HR EUAP
Others		HBP QQ

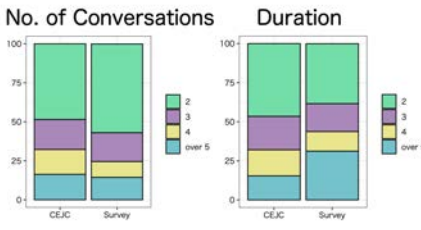
Survey of conversational behavior

- To demonstrate the variation of everyday conversation and design a corpus that contains various conversation in a balanced manner based on it
- Nov. 2014 - Feb. 2015
- 243 informants (balanced in age and gender)
- Two working days and a holiday per informant
- 9272 conversation in total

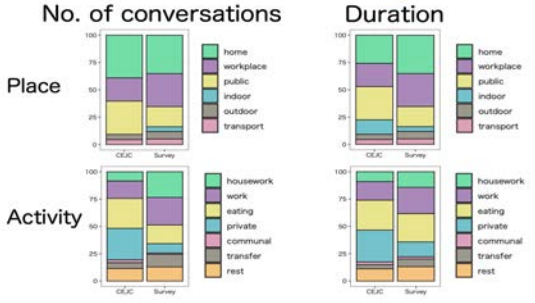
Conversation forms



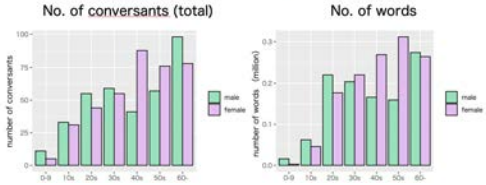
Number of conversants



Place and activity



No. of conversants and words per age and gender



The *Corpus of Everyday Japanese Conversation*

Released in March 2022
More information at:

https://www2.ninjal.ac.jp/conversation/cejc.html