6th, July,2018 Qualico 2018, Wrocław, Poland

Quantitative Characteristics of the BTSJ Japanese Natural Conversation Corpus (BTSJ-Corpus) ver. 2018: Focusing on the Differences of the Use of Polite Forms According to Sub-groups

Mayumi Usami, Makoto Yamazaki
National Institute for Japanese Language and Linguistics (NINJAL)

Introduction

BTSJ Japanese Natural Conversation Corpus ver. 2018 (hereafter BTSJ-Corpus) has compiled 333 Japanese Natural Conversations with recordings and transcriptions. The BTSJ-Corpus is one of the largest corpus (928,070 words) which compiles Japanese spontaneous oral conversations in various settings such as those between unacquainted people and between intimate friends of different social sub-groups. BTSJ is the abbreviation of the transcribing rules named ‘Basic Transcription System for Japanese’ which had developed by considering the characteristics of Japanese language and interaction style such as the phenomena that each sentence-final indicates the different politeness levels and there are many backchannels which overlap interlocutor’s utterances.

Quantitative characteristics of BTSJ-Corpus

Table 1: Fact sheet about BTSJ-Corpus 2018

No. of conversation	333
Total no. of tokens	922,444
Total no. of types	13,469
TTR	0.01460
Guiraud Index	14.024
Total no. of utterances	104,489
Av. no. of words per utterance	8.828
Total conversation time	284,784 sec.(79:06:24)
Av. time per utterance	2.725 sec
No. of different speakers	435

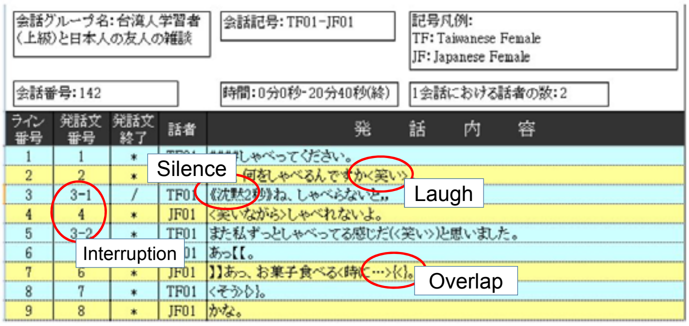
Table 2: Number of Tokens by conversation sub-group

	Native Speakers, Strangers	Native Speakers, Friends	Non-Native Speakers, Strangers	Non-Native Speakers, Friends
Tokens	220,222	400,659	62,674	32,227
Types	6,176	9,382	2,803	2,381
Guiraud Index	13.161	14.822	11.196	13.263

Methods

The percentage of the use of polite-forms of the following four groups were calculated.
Conversations between unacquainted people:
1) between native speakers. 2) between native speaker and non-native speaker.
Conversations between intimate friends:
3) between native speakers. And 4) between native speaker and non-native speaker.

Transcription of conversation



会話グループ名：台湾人学習者（上級）と日本人の友人の雑談

会話記号：TF01−JF01

記号凡例：
TF: Taiwanese Female
JF: Japanese Female

会話番号：142

時間：0分0秒−20分40秒(終)

1会話における話者の数：2

ライン番号	発話文番号	発話文終了	話者	発話内容
1	1	*	TF01	＊＊＊しゃべってください。
2	2	*	[illegible]	何をしゃべるんですか<笑い>。
3	3-1	/	TF01	《沈黙2秒》ね、しゃべらないと,,
4	4	*	JF01	<笑いながら>しゃべれないよ。
5	3-2	*	TF01	また私ずっとしゃべってる感じだ(<笑い>)と思いました。
6	[illegible]	[illegible]	[illegible]01	あっ[[。
7	6	*	JF01	]]あっ、お菓子食べる<時に…>{<}。
8	7	*	TF01	<そう>{>}。
9	8	*	JF01	かな。

Figure 1: A sample of BTSJ-Corpus transcript sheet

Japanese polite forms

desu : attached to nouns
Watashi wa gakusei ***desu.*** (I’m a student. *polite form*)

masu : attached to verbs, adjectives
Watashi wa mainich gakko ni iki ***masu***. (I go to school everyday. *polite form*)

Results

Table 3: Frequency and percentage of politeness forms by conversation sub-group

	Native Speakers, Strangers		Native Speakers, Friends		Non-Native Speakers, Strangers		Non-Native Speakers, Friends	
Frequency and Percentage of politeness forms	desu	5,272 (23.8%)	desu	981 (2.1%)	desu	1,685 (18.3%)	desu	275 (6.9%)
	masu	1,282 (5.8%)	masu	191 (0.5%)	masu	731 (7.9%)	masu	159 (4.0%)
	others	15,631 (70.5%)	others	44,368 (97.4%)	others	6,810 (73.9%)	others	3,554 (89.2%)
Total no. of utterances and percentage（total)		**22,176 (100.0%)**		**45,538 (100.0%)**		**9,221 (100.1%)**		**3,986 (100.1%)**

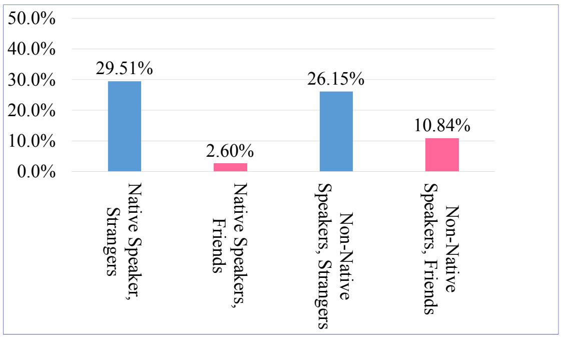


Figure2: Percentage of politeness forms by conversation sub-group

Discussion

These results show that the major difference in the use of polite forms between native and non-native speakers is the fact that non-native speakers do not switch the use of polite-forms depending on the interlocutor’s social factor such as superior and friend. And there are possibilities that this may cause some uncomfortableness especially in conversations between friends.

Future Plans

We will release the BTSJ-Corpus ver. 2018 in July 2018, which features new speaker ID.

Acknowledgement

This study was conducted as a part of the sub-project ‘The study on the Language Use of Japanese Language Learners’ (project leader: USAMI Mayumi) in Multiple Approaches to Analyzing the Communication of Japanese Language Learners" (Project Leader: ISHIGURO Kei) , NINJAL.