DOI: 10.17469/O2106SLI000003

TAKEHIKO MARUYAMA

Designs and Analyses of Japanese Speech Corpora

Since 2000 a series of Japanese speech corpora has been under development in NINJAL. In this paper, three of them, the CSJ (Corpus of Spontaneous Japanese), the CEJC (Corpus of Everyday Japanese Conversation) and the SSC (Showa Speech Corpus) will be introduced. These corpora are transcribed and morphologically analysed in a unified format so as to be mutually comparable. Users can access them through the web site "*Chunagon*" and search the data with part of speech information. Also, all the sound files, transcription and related data are distributed individually.

Keywords: speech corpora, CSJ, CEJC, SSC, filled pauses.

1. *Introduction*

A "corpus" can be defined as "a collection of written or spoken material stored on a computer and used to find out how language is used" (*Cambridge Dictionary*). The Survey of English Usage project, initiated in 1959 under Randolph Quirk at University College London, gathered a total of 100 million words of written and spoken British English and used it for the survey. In 1964 the first electric corpus, the "Brown Corpus", was created at Brown University; it consists of 100 million words of written American English. These two corpora opened up corpus linguistics as a new field of linguistic research.

The term "speech corpus" used here refers to a systematic and large collection of speech recorded in real-world situations. It contains digital audio files of various types of speech, such as monologues, conversations, and historical recordings. Additionally, various linguistic annotations have been added to speech corpora, including transcriptions, part of speech tags, utterance units, and meta-data. Using large speech corpora with rich annotation, linguists can empirically analyse how spoken language previously has been and presently is used in real life.

Since 2000 a series of Japanese speech corpora has been under development in NINJAL. In this paper, three of them, the CSJ (Corpus

of Spontaneous Japanese), the CEJC (Corpus of Everyday Japanese Conversation) and the SSC (Showa Speech Corpus) will be introduced. The distributions of filled pauses and first-person pronouns across these speech corpora will be compared to illustrate the utility of this kind of corpus design.

2. *Japanese speech corpora in NINJAL*

2.1 NINJAL

Established in 1948, NINJAL (National Institute for Japanese Language and Linguistics) has conducted scientific research on Japanese language for more than 70 years. In 1952 NINJAL started recording daily conversations in various situations in order to describe the intonation patterns, vocabulary, sentence length and structures, and types of words used in colloquial Japanese. The results were published in the report *Danwago no Jittai* ('Research in Colloquial Japanese') in 1955. Approximately 40 hours of speech were recorded and approximately 30 hours were analysed, making this a world-pioneering work in corpus-based research into colloquial speech.

Recently NINJAL has been designated as the Centre for Excellence of corpus creation in Japan. NINJAL has designed and constructed a series of Japanese corpora: the Corpus of Spontaneous Japanese (CSJ), Balanced Corpus of Contemporary Written Japanese (BCCWJ), Corpus of Historical Japanese (CHJ), International Corpus of Japanese as a Second Language (I-JAS), Corpus of Everyday Japanese Conversation (CEJC), Corpus of Japanese Dialects (COJADS), and Showa Speech Corpus (SSC). These corpora cover a wide range of Japanese language data, both written and spoken (monologue and dialogue), a span of 1300 years of Japanese history, a wide variety of dialects, and both learners and native speakers. Users may access the corpora through the web application "*Chunagon*" (https://chunagon.ninjal.ac.jp/).

2.2 Corpus of Spontaneous Japanese (CSJ)

The CSJ is the first of these corpora, released to the public in 2004 (NINJAL 2006). It includes 651 hours and 7.52 million words of spontaneous speech (mainly monologue) recorded from 1999 to 2001. This provided a new language resource to the field of linguis-

tics, especially for Japanese phonetics, phonology and syntax. It also contributed to the development of techniques for speech processing systems in areas such as automatic speech recognition and natural language processing (Maekawa 2004; NINJAL 2006).

The speech recorded in the CSJ can be classified into two categories: Academic Presentation Speech (APS) and Simulated Public Speaking (SPS). APS is composed of live recordings of presentations to various academic societies. SPS contains general speeches and comments by laypeople on everyday topics, given to small audiences. A relatively formal speaking style is observed in APS, while a casual speaking style is observed in SPS. Most monologues in the and SPS are 10-15 minutes long.

The CSJ is characterised by its rich annotations. It contains very precise transcription, segment labels and intonation labels for phonetics and phonology, part of speech tags for morphology and lexicology, clause-boundary labels, dependency structure, and discourse structure for syntax and discourse analysis. Also, speaker's information is provided for use in sociolinguistic analysis. All annotation is stored in a relational database which users can manipulate to retrieve data efficiently.

Table 1 shows the distribution of filled pauses appearing in the SPS core data set, which consists of 226,902 words. Let's compare two groups of speakers: male and female. The ratio of filled pauses used by male speakers is 7.7% of the total number of words, whereas the ratio of filled pauses used by female speakers is 5.0%. This indicates that male speakers produce filled pauses more frequently than female speakers.

Looking at the distribution of each form, clear differences can be observed between male and female speakers. Of their filled pauses, male speakers' use of *ma* and its elongated form *ma:* comprises 29%, while female speakers' use of those forms comprises only 17%. On the other hand, female speakers' use of *ano:* and *an*o comprises 37%, while male speakers' use of these forms makes up only 15%. This distribution shows that male and female speakers choose different types of filled pauses: male speakers using *ma:* and female speakers using *ano:*. The result shows a sociolinguistic pattern of behaviour in Japanese monologue speech.

Table 1 - *Filled pauses observed in the CSJ (SPS core)*

Male (117,411 words)			*Female (109,491 words)*		
2,080	22.9%	*e:*	1,076	19.7%	*ano:*
1,361	15.0%	*ma:*	955	17.4%	*ano*
1,276	14.1%	*ma*	732	13.4%	*e:*
767	8.5%	*ano:*	549	10.0%	*ma*
569	6.3%	*ano*	419	7.7%	*ma:*
433	4.8%	*e*	256	4.7%	*e*
315	3.5%	*e:to*	181	3.3%	*sono*
263	2.9%	*sono*	149	2.7%	*n*
203	2.2%	*n*	147	2.7%	*e:to*
9,072	100%	total	5,473	100%	total

2.3 Corpus of Everyday Japanese Conversation (CEJC)

The CEJC consists various conversations in daily situations, such as conversations during dinner with the family at home, meetings with colleagues at work, and conversations while driving (Koiso *et al.* 2016, 2022). It includes 200 hours and 2.4 million words of daily conversations in a balanced selection. To estimate distributions of various daily conversations, a survey of everyday conversational behaviour had been conducted previously with about 250 adult Japanese informants. The survey asked when, where, how long, with whom, and during what kind of activity informants was engaged in conversations in their daily life. Based on the results, conversation forms (chat, business, meeting), places (home, workplace, public, indoor, outdoor, transport), and activities (housework, work, eating, private, communal, transfer, other) were defined as a measure of the design of a balanced corpus (Koiso *et al.*, 2016, 2022).

The most distinctive feature of the CEJC is that it provides multi-directional video files, as shown in Figure 1. The entire speech situation were captured by the three cameras, thus speakers' gaze movement, nodding, and gestures can be observed. The CEJC also serves audio files, transcription, TextGrid files for Praat, eaf files for ELAN, part-of-speech tags, and speaker's information.

Figure 1 - *An example of video file in the CEJC*

Table 2 - *Filled pauses observed in the CEJC (ver. 2018)*

Male (260,183 words)			*Female (349,144 words)*		
2,160	65%	*ano*	2,042	63%	*ano*
412	12%	*sono*	320	10%	*sono*
142	4%	*e:*	186	6%	*etto*
113	3%	*e:to*	147	5%	*e:to*
75	2%	*n*	115	4%	*a:no*
73	2%	*n:*	92	3%	*n*
71	2%	*etto*	66	2%	*a*
52	2%	*a*	61	2%	*e:tto*
3,305	100%	total	3,263	100%	total

Table 2 shows the distribution of filled pauses appearing in the CEJC (ver. 2018), which consists of 609,327 words. The ratio of filled pauses used by male speakers is 1.2% of the total number of words, whereas the ratio of use by female speakers is 0.9%. Compared to the result from the CSJ, the proportion of filled pauses appearing in conversation is considerably lower than in monologue.

Comparing male and female speakers, the distribution is almost identical. For both groups, use of *ano* is just over 60% and use of *sono* is around 10% of all filled pauses. This result implies that differences in style – monologue vs. conversation – produce different linguistic

behaviours according to gender. Male and female speakers behave differently in monologue situations and behave similarly in conversation.

2.4 Showa Speech Corpus (SSC)

The SSC is a collection of conversations and monologues recorded from the early 1950s to the early 1970s (a span included in the Showa Era) by NINJAL (Maruyama 2020, 2021). The earliest of these recordings were collected in 1952 in the survey of colloquial speech mentioned above. Various types of colloquial speech were sampled according to the sampling frame, including region (uptown, downtown, outskirts), place (home, neighbourhood, school, place of work), gender, age, educational background, and number of speakers. Also, some monologues such as academic presentations and congratulatory addresses were recorded.

Figure 2 is a picture of recording a conversation in the 1950s.

Figure 2 - *Recording conversation in the 1950s*

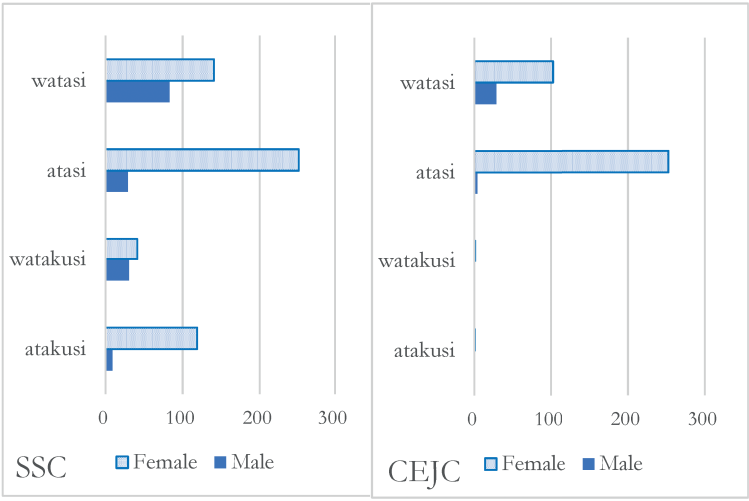
The original reel-to-reel materials, which had been stocked in the NINJAL archive for decades, were digitized by the 1990s. The author of this paper collected the sound files in the archive, newly transcribed them, morphologically analysed them for part of speech, and annotated them with meta-data. A series of this data was compiled into the SSC, which was released in 2022. The SSC includes 73 conversations

(in total approximately 27 hours) and 50 monologues (approximately 17 hours). Most of the conversations are casual chats involving men and women of all ages, whereas all the monologues were lectures or talks held at NINJAL.

One of the characteristic features of the SSC is that it provides a basis for diachronic analysis of spoken Japanese. As expected, some "old-fashioned" linguistic expressions can be observed in the SSC, including aspects of pronunciation, intonation, vocabulary, and grammar. Connecting the SSC to the contemporary speech corpora like the CSJ and CEJC creates a diachronic speech corpus, covering the late 20th century and early 21st century (Maruyama, 2020, 2021).

In Japanese, the form of the first-person pronoun *watasi* has some morphological variations, *watasi*, *atasi, watakusi,* and *atakusi*. The forms *watakusi* and *atakusi* are more polite than *watasi* and *atasi*. Comparing the SSC conversation part to the CEJC, the distributions of first pronoun variants (per 0.1M words) can be illustrated as in Figure 3.

Figure 3 - *Distribution of first-person pronoun in the SSC and CEJC*



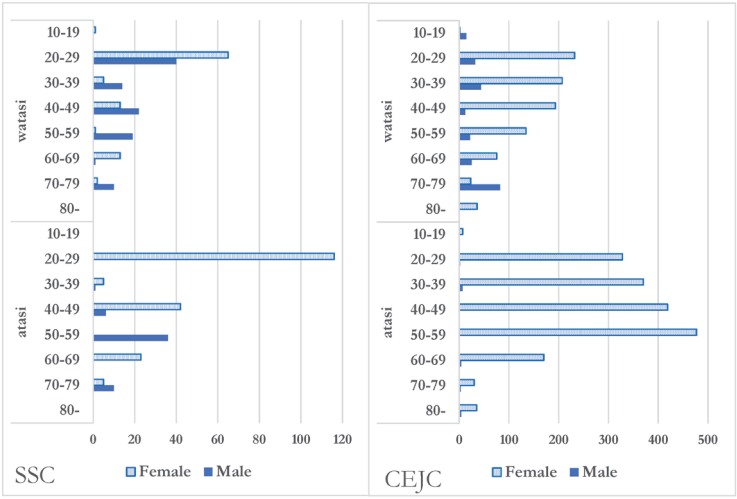
The numbers of uses of *watasi* and *atasi* by female speakers are almost the same between SSC and CEJC, so there has been no change during the 60 years. On the other hand, uses of *watasi* and *atasi* by male

speakers has declined significantly. The more polite forms *watakusi* and *atakusi* have completely disapeared in male and female speakers. These are diachronic changes in the use of first-person pronouns.

Figure 4 shows the number of uses of first-person pronoun variants broken down by speakers' age. In the SSC on the left side, *atasi* is shown to be used more by younger female speakers. On the other hand, in the CEJC on the right, *atasi* is used more by older female speakers. This can be considered to indicate that female speakers who were young in the 1950s are now in the older age group, and the tendency of using *atasi* is still present even today.

Looking at the distribution of *watasi* in the CEJC, the younger age groups use it more than the older age groups. From the past to the present day, it can be interpreted that use of the first-person pronoun by female speakers is in the process of transitioning from *atasi* to *watasi*.

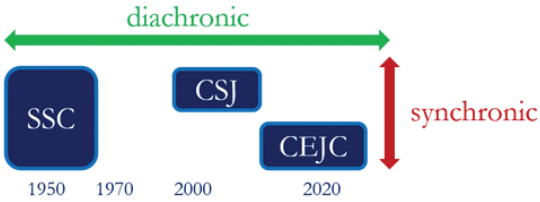
Figure 4 - *Distribution of first-person pronoun by speakers' age*



3. *Concluding remarks*

This paper outlines the CSJ, CEJC and SSC, three of the speech corpora designed and constructed in NINJAL. The positioning of the three corpora may be illustrated as follows.

Figure 5 - *Relationship of the CSJ, CEJC and SSC*



By comparing the CSJ, a corpus of monologue, with the CEJC, a corpus of daily conversation, it is possible to analyse the diversity of the synchronic spoken language. On the other hand, by linking the historical SSC corpus with the contemporary CSJ and CEJC corpora, it is possible to analyse diachronic changes in spoken language.

In the future, the development of more diverse speech corpora will enable comprehensive analysis and description of spoken Japanese in modern and contemporary times.

Acknowledgments

This study 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", and JSPS KAKENHI Grant Number 20H05630 and 16H03426. I would like to thank Dr Stephen Wright Horn for useful comments and help.

References

Koiso, Hanae & Tsuchiya, Tomoyuki & Watanabe, Ryoko & Yokomori, Daisuke & Aizawa, Masao & Den, Yasuharu. 2016. Survey of Conversational Behavior: Towards the Design of a Balanced Corpus of Everyday Japanese Conversation. In *Proceedings of LREC 2016*, 4434-4439. https://aclanthology.org/L16-1702/

Koiso, Hanae & Amatani, Haruka & Den, Yasuharu & Iseki, Yuriko & Ishimoto, Yuichi & Kashino, Wakako & Kawabata, Yoshiko & Nishikawa, Ken'ya & Tanaka, Yayoi & Watanabe, Yuka & Usuda, Yasuyuki. 2022.

Design and Evaluation of the Corpus of Everyday Japanese Conversation, In *Proceedings of LREC 2022*, 5587-5594.
http://www.lrec-conf.org/proceedings/lrec2022/pdf/2022.lrec-1.599.pdf

NINJAL 1955. *Danwago no jittai (Research in Colloquial Japanese)*. NINJAL Report 8. Tokyo: NINJAL.

NINJAL 2006. *Nihongo Hanasi Kotoba Kopasu no Kotikuho (Construction of the Corpus of Spontaneous Japanese)*. NINJAL Report 124. Tokyo: NINJAL.

Maekawa, Kikuo. 2004. Design, Compilation, and Some Preliminary Analyses of the Corpus of Spontaneous Japanese, In Yoneyama, Kiyoko & Maekawa, Kikuo (Ed.), *Spontaneous Speech: Data and Analysis*, 87-108. Tokyo: NINJAL.

Maruyama, Takehiko. 2020. On the Possibility of a Diachronic Speech Corpus of Japanese. In Bekeš, Andrej & Srdanović, Irena (Ed.), *Japanese Language from Empirical Perspective: Corpus-based studies and studies on discourse*, 219-234. Ljubljana: Znanstvena založba FF.
https://ebooks.uni-lj.si/zalozbaul//catalog/book/187

Maruyama, Takehiko. 2021. Diachronic Change of Spoken Japanese in the 20th Century: A Corpus Study. In Suzuki, Seiko & Cordereix, Pascal & Bergounioux, Gabriel (Ed.), *Mémoire sonore du Japon : le disque, la musique et la langue*, 17-32. Paris: Bibliothèque nationale de France.