10 On the possibility of a diachronic speech corpus of Japanese

MARUYAMA Takehiko
Senshu University / NINJAL

Abstract
This study investigates the possibilities of the establishment of a diachronic speech corpus of Japanese. After identifying the conditions and limitations that underlie the process of compiling a diachronic speech corpus, this study discusses its potentiality with analyses of intonation patterns and grammatical expressions.

Keywords: Diachronic speech corpus of Japanese, Okada collection, Danwago data, Intonation patterns, Grammatical expressions

1 Introduction

Since the Corpus of Spontaneous Japanese (CSJ) was released to the public in 2004, research on Japanese spontaneous speech, as opposed to reading aloud, has made drastic progress. CSJ, including 7.52 million words with 651 hours of spontaneous speech, not only provides new research data in the field of the linguistic study of speech such as phonetics, phonology and syntax, but also greatly contributes to a wide range of pursuits in linguistics and related fields, most notably in variation studies in sociolinguistics, and in the development of techniques for speech processing systems such as ASR (Automatic Speech Recognition) and NLP (Natural Language Processing). On the other hand, there is an ongoing requirement to develop a corpus of daily conversation, as most of the spoken data collected in CSJ are monologues. In response to this request, NINJAL (National Institute for Japanese Language and Linguistics) started a new project to establish a new corpus called CEJC, which contains 200 hours of daily conversation in various contexts (Koiso et. al. 2016). When development has been completed and after its release, it will be possible to establish linguistic resources focusing on contemporary spoken Japanese, of both monologues and dialogues, which may lead to further progress in studies of spoken language.

Taking the aforementioned research as a starting point, this study investigates the possibilities of the establishment of a diachronic speech corpus of Japanese. While diachronic corpora usually target written language, the compilation of chronological speech data for the diachronic study of speech holds great prospects. In what manner will it contribute to spoken language studies? After identifying the conditions and limitations that underlie the process of compiling a diachronic speech corpus, this study discusses its potentiality with some case studies.

2 Previous research

In the history of corpus linguistics, there have been very few attempts to compile diachronic speech corpora. The Diachronic Corpus of Present-day Spoken English (DCPSE) is an example of such an attempt, presenting spontaneous speech data of British English from the 1960s to the 1990s.

In 2006, the DCPSE (by the "Survey of English Usage" project at the University College London) was released to the public.[1] This diachronic speech corpus contains colloquial British English from the late 1960s to the early 1990s. The recorded resources between the 1960s and the 1970s were derived from the London-Lund Corpus, whereas the data from the 1990s is from ICE-GB. Each of these two corpora contain approximately 400 thousand words.

All these transcribed texts were annotated with morphological and syntactic information. Aarts et al. (2015) quantitatively clarified that (1) the usages of auxiliary verbs, "must," "may," and "shall" drastically declined within this 30-year period; (2) the usages of "would," "could," and "should" also declined; and (3) the usages of "will" and "can" conversely increased. Although there are some problems underlying this corpus (concerning the validity and representativeness of diachronic speech corpora as will be discussed in Section 3), this is an exemplary attempt at compiling a diachronic speech corpus for the purpose of quantitatively and diachronically analyzing the linguistic changes in spoken language.

Concerning research on existing recorded resources of old spoken Japanese, Shimizu and Kanazawa have collected and analyzed wax cylinder recordings and Standard Playing (SP) records (Shimizu 1988, 1994, 2011, 2014, Kanazawa 1991, 2000, 2015). However, these materials are mainly composed of old recordings of *rakugo* (Japanese traditional comic storytelling), and thus are different from natural speech data.

Old recordings of spoken Japanese can also be found in the "Okada Collection" archive, mainly consisting of political speeches recorded in the early first half of the 20th century. In the early 1950s NINJAL began recording daily conversations with a tape recorder, resulting in approximately 80 hours of recorded material. The contents of this material will be described in Section 4.

3 Conditions and limitations of a diachronic speech corpus

This section will consider the necessary conditions for the realization of a diachronic speech corpus. Three conditions, according to the following key terms: "diachronic," "speech," and "corpus", are identified here.

1 http://www.ucl.ac.uk/english-usage/projects/dcpse/

Concerning the first condition, "diachronic," the corpus must be a collection of speech data from various time periods. It should be well organized in order to enable an analysis of the changes in spoken Japanese. Second, in terms of the "speech" condition, the recorded resources must be preserved so that playback and listening are possible. The essential resource of any speech corpus is the recorded data itself; thus, a collection of transcriptions alone cannot be called a speech corpus in the truest sense. Furthermore, the quality of recorded data should be as clear as possible; particularly concerning conversation, it is preferable to have a multiple track recording. A condition necessary for a true "corpus" is that "a corpus should contain vast range of digitized examples with various information for linguistic retrieval" (Maekawa 2013). This means that it must include not only a variety of electronic recorded data, but various annotations such as transcriptions, morphological information including POS, syntactically parsed information, and various metadata such as information on speakers, recorded time, speaking style, etc.

In reality, however, it is extremely difficult to fulfill all the aforementioned conditions. For instance, the condition of "speech data recorded various periods" is limited due to the fact that recording devices were only developed in the late 19th century, only becoming broadly available in and after the 20th century. This means that a diachronic speech corpus can only target spoken data collected in and after the 20th century.[2] While a diachronic corpus of written language enables us to collect written Japanese in and after the 8th century, a diachronic speech corpus must be extremely limited in scope as well as in quantity. Second, the condition of "good quality recording" is also limited as the quality of recorded materials from an earlier period is comparatively poor and not well-preserved. As explained in Section 4, the conversational speech data collected by NINJAL in the 1950s are sometimes insufficient in volume and marred by noises. Furthermore, the speech data that are not yet digitized may become inaudible in the near future, as the original media will inevitably deteriorate. Therefore, digitizing them is an urgent matter, but also time-consuming and costly. Finally, the condition of "a good quantity of data from various contexts" is also difficult to achieve as the amount of existing speech data is severely limited. Therefore, with respect to quantity and quality of data, a diachronic speech corpus cannot be expected to have the same quantity and quality of data as a large-scale corpus. In addition to the fundamental difficulty of achieving a balance in speech corpora generally (Maekawa 2013), problems are exacerbated when we target historical spoken data. Furthermore, even when original recorded data are extant, difficulties in using them broadly may arise due copyright.

Considering the limitations of preserved spoken data, we must accept that a diachronic speech corpus must be restricted in quantity and quality in its balance and

2 According to Shimizu (2014), the oldest recorded material of Japanese speech (discovered so far) is a reading of the Bible by Ichitaro Hitomi, which was recorded by the Societe d'Anthropologie de Paris on July, 1900, at the Expositions universelles de Paris.

diversity. In other words, it is crucial to collect the existing data in as wide an area as possible. This condition is the same as that faced by linguists working on Old Japanese texts from 8th century, who must conduct research with limited linguistic resources. As long as researchers are using old data, it is unavoidable for them to face the problem of quantitative limitation.

Taking these conditions into account, it is necessary to collect as many types of recorded material as possible and annotate the "metadata" required to categorize them (Maruyama 2012) for the purpose of compiling a single diachronic speech corpus. With regard to material, for example, public speeches and lectures in the collection of "Historical Speech Data"[3] made public on the NDL (National Diet Library) website can be one source of valid data. Annotating metadata requires an investigation into how to categorize the spoken data according to their characteristics. The categories of metadata annotated in the CSJ, such as monologue/dialogue, situation of speech, speaker (gender, age, and place of birth), speaking styles (high or low), and the degree of spontaneity, should be useful. In this manner, determining the criteria is essential to analyze and categorize the various types of speech data in multiple ways.

4 Data

In this section, we will consider the possible types of linguistic research when using a diachronic speech corpus such as the one proposed, also considering several concrete examples of available data to inform the discussion. In addition to the CSJ previously mentioned, the following two spoken resources will also be used as data for analysis:

1. Okada Collection I, *Kichō Ongen* Collection, *Sōryū-sha* Academic Resource Series
2. Recorded data in *Danwago no Jittai* (Research in Colloquial Japanese)

The first set of data will be henceforth referred to as the Okada Collection, and the second will be referred to as the *Danwago* Data. The Okada Collection is a set of spoken data recorded in SP vinyl from the late *Meiji* era (1867–1912) to the beginning of the *Shōwa* era (1926–1989). In total, 18.5 hours of speech data, comprised of 165 original speeches, were digitized and published[4] from among 35 thousand vinyl recordings collected by Mr. Norio Okada. All data are monologues, categorized into political speeches, general lectures, Buddhist sermons, recitations, and so on. Although several unclear segments exist due to low sound quality, the Okada Collection is undoubtedly important as a rare collection of spoken data from approximately 100 years ago.

3 http://rekion.dl.ndl.go.jp/

4 http://www.nichigai.co.jp/database/sp/

Professor Hiroyuki Kanazawa transcribed the Okada Collection in a NINJAL project titled "An Analysis for the Dynamic State of the Contemporary Japanese from Multifaceted Approach" (2009–2015, leader: Professor Masao Aizawa). A collection of papers with the results of this project has been published (Aizawa & Kanazawa 2016). In this study, I have used 109 lectures categorized as "political speech" or "lectures": 14.5 hours of speech data overall. The number of speakers is 76. Table 1 shows examples of speech data.

Table 1: Examples of Speech Data in Okada Collection

Years	**Speaker (Year of Birth)**	**Title of Speech**	**Recorded Time**
1915	Yukio Ozaki (1858)	A Speech by the Minister of Justice, Yukio Ozaki	0:28:10
1916	Shigenobu Ōkuma (1838)	The Power of Public Opinion in Constitutional Politics	0:17:14
1926	Shimpei Gotō (1857)	Ethics of Politics	0:12:54
1931	Tsuyoshi Inukai (1855)	Necessity of a Strong Cabinet	0:04:09
1937	Senjurō Hayashi (1876)	Announcement to Japanese Citizens	0:06:13
1941	Fumimaro Konoe (1891)	Concerning the Conclusion of the Tripartite Pact	0:10:25

The latter data, the *Danwago* Data, is a collection of resources from the 1950s to the 1960s recorded at NINJAL. Since its establishment in 1948, NINJAL has been investigating colloquial Japanese, including the Tokyo dialect and other local dialects. The earliest research using a recording device was done in October 1950 at Shirakawa city in Fukushima prefecture. Beginning in 1952, NINJAL began collecting daily conversations in various contexts. They analyzed intonation patterns, vocabulary, *bunsetsu* (phrases in Japanese), the length and structure of sentences, types of words, and so on. The results were published in three reports, *Danwago no Jittai* (Research in Colloquial Japanese) in 1955 and *Hanashi Kotoba no Bunkei 1, 2* (Research of Sentence Patterns in Colloquial Japanese 1, 2) in 1960 and 1963 (NLRI 1955; 1960; 1963). Approximately 40 hours of speech were recorded, and approximately 30 hours of speech were analyzed in these reports.

Although most of the recorded materials have been currently digitized, they are not well organized for linguistic analysis. The author of this paper has transcribed some

of them, including 33 conversations (in total approximately 19.5 hours) and 21 monologues (approximately 17 hours). Figure 1 illustrates the examples of the spoken data included. Most conversations are chats among laypeople, whereas all the monologues were lectures or talks held at NINJAL.[5]

Conversations: *Kudan High School Students, Three Young People, Kamakura Housewives, Old Men and Women, Fish Shop's Son*

Monologues: *Particles and Auxiliary Verbs, Lecture on Japanese Language, Accent etc. in Japanese, Talk for the 10th Anniversary of NINJAL*

Figure 1. Examples of Speech Data in the *Danwago* Data

Table 2 indicates the data statistics. The Okada Collection is divided into three periods: *Taisho* (from 1915–1926), *Shōwa* 1–9 (1926–1934), and *Shōwa* 10–19 (1935–1944). The *Danwago* Data is separated into two categories: conversations and monologues. The total frequency of words is calculated from the result of morphological analysis (UniDic 2.1.2 + MeCab 0.996) by removing supplementary-symbols (punctuations, brackets, and so on).

Table 2: Data Statistics of the Okada Collection and the *Danwago* Data

	Okada Collection (1915–1944)			***Danwago* Data (1950s–1960s)**	
	1915–1926 (*Taisho*)	1926–1934 (*Shōwa* 1–9)	1935–1944 (*Shōwa* 10–19)	Conversations	Monologues
Number of Files	19	52	38	33	21
Number of Speakers	16	42	30	unknown	21
Time of Recording	3 hours	6 hours	6 hours	19.5 hours	17 hours
Total Number of Words	23,022 words	46,998 words	49,070 words	218,497 words	182,619 words

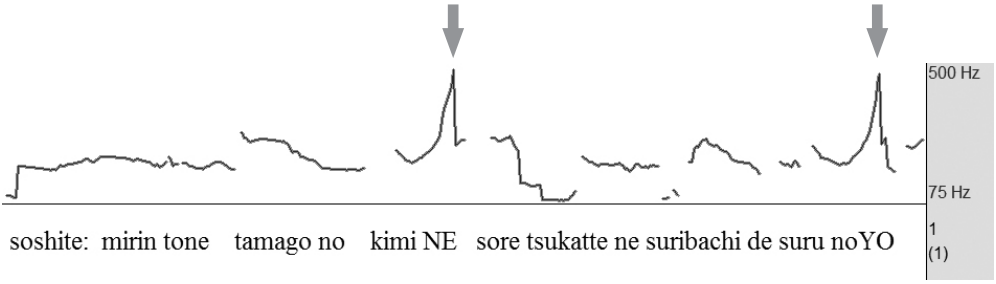
5 Concerning the conversation data, unfortunately, information such as the age of speakers, their occupation, place of birth, accurate dates and places of recording is partly not available.

Here we should note that these two collections do not necessarily cover a wide range of spoken Japanese. The Okada Collection is a collection mostly of political speech, while the *Danwago* Data includes daily conversation and lectures by researchers. It is needless to say that an optimal speech corpus should contain a wide range of speech in various situations, since distributions of intonation, vocabulary, grammatical expression and speaking style may vary in different situations. However, at least at this stage, we have to proceed with analyses using these limited data as a single corpus, since no other sizable collections of speech data have been found. The following sections present concrete case studies using the Okada Collection, the *Danwago* Data, and the CSJ to discuss the possibilities of a diachronic speech corpus.

5 Analysis of the Okada Collection and *Danwago* Data

5.2 Analysis of intonation

What is examined here is the characteristic pattern of intonation seen in the *Danwago* Data. Figure 2 shows a pitch contour of an utterance which appeared in an excerpt of recorded material called "Three Ladies" from 1957. The utterance is "*soshite: mirin tone tamago no kimi NE sore tsukatte ne suribachi de suru noYO*" (I added syrup and egg yolk, and then used a mortar to grind and mix them). We can see in the contour that the pitches on *NE* at the end of a phrase and on *YO* at the end of the utterance rise very rapidly. It is certain that this rising intonation does not signify a question addressed to the listener.

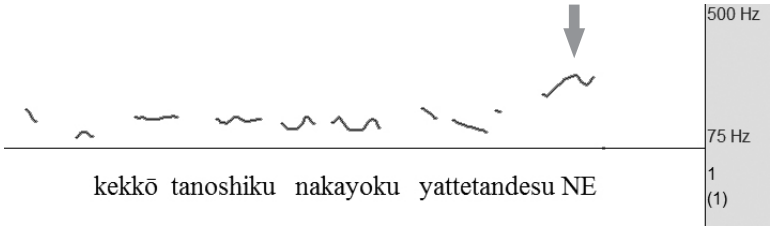


"*soshite: mirin tone tamago no kimi NE sore tsukatte ne suribachi de suru noYO*"

Figure 2. Rapid Rising Intonation (*Danwago* Data: "Three Ladies")

This rising intonation in Figure 2 reminded me of scenes spoken by actresses in old Japanese films from the 1950s. For example, in the film "Tokyo Story" (directed by Yasujirō Ozu in 1953), rising intonations like those in Figure 2 are frequently observed in the utterances by the actress Setsuko Hara. This suggests that such intonation patterns might have been natural for women in the 1950s.

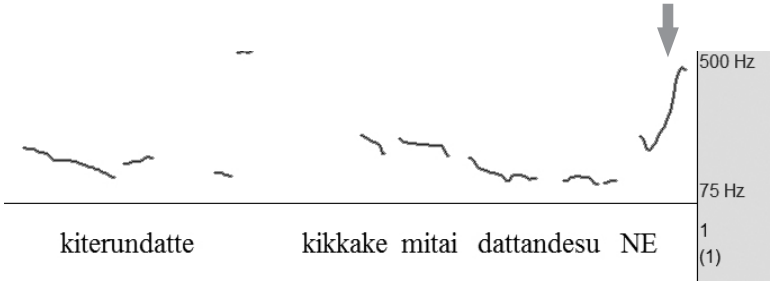
Rising intonations at the end of phrases in non-interrogative contexts can be seen even in the CSJ, which contains utterances by contemporary Japanese speakers. For instance, in Figure 3 the utterance "*kekkō tanoshiku nakayoku yattetandesu NE* (I had a fun and convivial time)" shows a rising intonation at the end. This, however, is not equivalent to the drastic rise seen in Figure 2.



"*kekkō tanoshiku nakayoku yattetandesu NE*"

Figure 3. Rising Intonation (CSJ:S05F1600)

On the other hand, Figure 4 shows a rising intonation at the end of a phrase seen in the CSJ, "*kiterundatte kikkake mitai dattandesu NE* (So it seems that this was the start of (it) coming, you know)," which seems to be very much similar to the intonation pattern in Figure 2.



"*kiterundatte kikkake mitai dattandesu NE*"

Figure 4. Rising Intonation 2 (CSJ:S01F1522)

Here, it is important to focus on the age gap between the speakers in Figures 3 and 4. The speaker in Figure 3 was in her late 20s (year of birth: the early 1970s) at that time of recording, whereas the age of the speaker in Figure 4 was approximately 50 (year of birth: the late 1940s). This means that there is an age difference of 25 years between them. At this point, we note that drastic rising intonations at the end of phrases like Figure 2 and 4 can be heard in the utterances of old women even in contemporary Japanese. It seems that these intonations occur in contexts in which older women, either of higher social

standing or with pretenses to such, speak elegantly. Furthermore, when my daughter listened to the utterance in Figure 2 she said, "It sounds like my grandmother speaking."

Supposing that the female speaker in Figure 2 was in her mid-20s at the time of recording in the 1950s, she would now be in her 80s. We can infer that the informants for the *Danwago* Data even now partially preserve and utilize the same intonation patterns from that era. For today's younger generation, however, an example like that in Figure 2 sounds like an utterance by an elderly woman or an actress in an old Japanese film.

Assuming a language change in which a new intonation pattern emerges among the younger generation thus replacing an older one, there is a point where rising intonations such as those in Figures 2 and 4 decline, and cannot be used by the younger generation. However, we must make recourse to more recorded resources and make quantitative and cross-sectional analyses in order to locate the era when such intonation patterns disappear.

5.2 Analysis of grammatical expressions: auxiliary verb *masuru*

This section focuses on grammatical expressions employing the auxiliary verb *masuru*. Hattori (2011) pointed out that the form of *masuru* began to change to *masu* at the beginning of the early modern period (from the *Azuchi-Momoyama* period (1568–1600) to the *Edo* period (1603–1867)), but it can still be frequently observed in the Okada Collection. The examples below show usages of *masuru* appearing at the end of a sentence (EOS) in (1), and appearing in verb phrases in both a *to*-clause and a *ga*-clause in (2).

(1) *Meiji 17 nen, sentē hēka no gorē o-itsutsu no koro to kioku o itashite orimasuru*
In Meiji 17th (1884), I remember that it was the time when the emperor was 5 years old.
(Akinobu Manabe, "*Taiko tenno go-yōji o shinobi tatematsurite,*" 1927)

(2) *Konnichi, shinbun nazo o mimasuru to, makoto ni nagekawashī koto ga takusan arimasuru ga, hitotsu ni ryōshin o kaeriminaide akuma no koe ni damasarete…*
Nowadays, while reading the newspaper, we find many lamentable matters, this is firstly because we do not care about conscience so that we are deceived by the devil…
(Motojiro Makino, "*Ryoshin undō no daiissei*", *Shōwa* 10s (1935–1944))

Such usages of *masuru* are also seen in the *Danwago* Data in the verb phrases of *kara*-clauses, *to*-clauses, and *keredomo*-clauses, as exemplified below.

(3) *Hijō ni yosan no kyūkutsu na, a:, jidai de arimasuru kara, e:, soredemotte….*
Now, it is the time that we must be very tight in budget, so…
(Yūzō Yamamoto, "Talk for the 10th anniversary of NINJAL," 1959)

(4) *Rajio news no kakikata toyū yōna hon o mi<u>masuru</u> to, e:, news niwa…*
Reading through a book titled "how to write a script for radio news" well, news…
(Kanji Hatano, "Talk for the opening anniversary of new office," 1962)

(5) *Atarashii jibiki ga 20-man go o shūsaisuru to kaite ari<u>masuru</u> keredomo, sononaka no 2 man go shika…*
Although it says that a new dictionary contains 200 thousand words, only 20 thousand of them…
(Ōki Hayashi, "Talk for the opening anniversary of new office," 1962)

During the lectures and talks in (1) to (5), the speakers also used an auxiliary verb, *masu*, such as "*kangeki itashiteoru shidai de ari<u>masu</u>* (I am very moved)" (Manabe), "*Rongo no uchi de attaka to omoi<u>masu</u> ga* (I think it was in the Analects of Confucius)" (Makino), "*muzukashiinde ari<u>masu</u> kara* (because it is difficult)" (Yamamoto), "*kiite ori<u>masu</u> to* (Listening to it,…)" (Hatano) and "*sa mo ari<u>masu</u> keredomo* (although there is a gap)" (Hayashi). Given that the two forms are functionally equivalent polite verbal endings, this means that *masu* and *masuru* are morphological variants. [6]

Table 3 shows the number of the auxiliary verbs *masu* and *masuru* that appeared in monologues in the Okada Collection, the *Danwago* Data, and the Core of CSJ (177 lecture talks, in total 41 hours).

Table 3: The number of *masu* and *masuru* that appeared in each data set

	Okada Collection						***Danwago* Data monologues**		**CSJ Core monologues**	
	1915–1926 (*Taisho*)		**1926–1934 (*Shōwa* 1–9)**		**1935–1944 (*Shōwa* 10–19)**					
masu	271	(86.6%)	752	(89.8%)	903	(92.9%)	3,918	(98.8%)	5,604	(100%)
masuru	42	(13.4%)	85	(10.2%)	69	(7.1%)	48	(1.2%)	0	(0%)

We can see that although *masuru* constituted 13.4% of all polite verbal endings in the *Taisho* era, it was eventually replaced by *masu*. The contemporary Japanese corpus, the CSJ, did not have any example of *masuru*.

I then proceeded to analyze the words that follow *masuru* in the Okada Collection and the *Danwago* Data. Table 4 presents the results for the top 10 expressions that appeared most frequently after *masuru*.

6 The years of birth for the speakers are Manabe (1878, Meiji 11), Makino (1874, Meiji 7), Yamamoto (1887, Meiji 20), Hatano (1905, Meiji 38) and Hayashi (1913, Taisho 2).

Table 4: List of Words that Appear after *masuru*

Okada Collection						*Danwago* Data monologues	
1915–1926 (*Taisho*)		1926–1934 (*Shōwa* 1–9)		1935–1944 (*Shōwa* 10–19)			
11	*to* (conjunctive particle)	28	*ba* (conjunctive particle)	16	。(EOS)	13	*keredomo* (conjunctive particle)
6	。(EOS)	11	*ga* (conjunctive particle)	13	*ga* (conjunctive particle)	8	*kara* (conjunctive particle)
5	*ga* (conjunctive particle)	10	noun phrase	8	*to* (conjunctive particle)	7	*to* (conjunctive particle)
4	*yueni*	7	*to* (conjunctive particle)	7	*ba* (conjunctive particle)	6	*shi* (conjunctive particle)
3	*naraba*	5	*ni* (case-marking particle)	5	noun phrase	6	noun phrase
3	*keredomo*	5	*kara* (conjunctive particle)	4	*no* (nominal particle)	6	*ba* (conjunctive particle)
3	*kara* (conjunctive particle)	5	。(EOS)	4	*kara* (conjunctive particle)	4	*ga* (conjunctive particle)
2	*no* (nominal particle)	4	*keredomo*	4	*ka* (sentence-final particle)	2	*ni* (auxiliary verb)
2	*ni* (case-marking particle)	3	*ya* (sentence-final particle)	3	*ya* (sentence-final particle)	1	*tameni*
1	*ba* (conjunctive particle)	3	*toiu* (quotation)	3	*ni* (case-marking particle)	1	*yueni*

While there are some instances of *masuru* appearing at EOS in the Okada Collection, no instance of *masuru* at EOS is observed in the *Danwago* Data. This result coincides with the view of Hattori (2011) that a prominent characteristic of *masuru* is that it never appears at EOS. This observation was obtained through his analysis of this auxiliary verb as it appears in the numerous amount of data of minutes of the National Diet. Further results reveal that although there is a low frequency of the conjunctive particle *keredomo* after *masuru* in the Okada Collection, *keredomo* achieves its greatest

frequency in the *Danwago* Data. It seems that in time, the form tended to be avoided for terminating a sentence, and became preferable as the style for connecting with the conjunction *keredomo*.

What we can infer from this is that the auxiliary verb *masuru* tended to be used in formal monologues (e.g., lectures, talks, and formal speeches) by a relatively small number of people. Here, the speakers' usages are strongly affected by their respective dates of birth. In time, the younger generations ceased to use *masuru* in their speech, yet it is impossible to know exactly when it began to disappear in monologues at this stage of research due to a lack of data.[7] To clarify changes in the process of alternation between *masuru* and *masu*, we need to supplement our data with more recording material from the blank period, from more registers and from a greater variety of speech situations.

5.3 Analysis of grammatical expressions: sentence final particles

This section analyzes the frequency and combination of sentence-final particles. We now take a look at examples of sentence-final particles in the conversations in the *Danwago* Data.

(6) a. *Watashi dattara Kyūshū ni ikitaiwa.* ("Sagami Female College student")
If I were you, I would go to Kyushu.
b. *Harau to shitara, taihen desu wane.* ("*Tomonokai* member")
If I must pay it, it is hard.
c. *Anta n toko no o-sakana, oishii wayo.* ("Fish shop's son")
Your fish is delicious.
d. *Sensei to o-hanashi shite kimashita noyo.* ("Kamakura housewife")
I went to talk with a teacher.

All these examples contain sentence-final particles that are commonly understood as appearing at the end of a woman's utterance. Even though the utterances they attached to are not interrogatives, the particles *wa* and *noyo* have rising intonations. It is certainly extremely rare to observe analogous usages of *wa*, *wane*, *wayo*, and *noyo* in the conversations of younger people in contemporary Japanese. If they appear at all, these forms will most likely appear as features of a "role language" employed when its speakers take on the role of older women of higher social standing.

7 Hattori (2011) reports that the usages of *masuru* can be seen in the minutes of the National Diet even today, albeit in a smaller number.

In contrast, supposing the speakers of these utterances to be old women in the present day, the utterances all sound rather natural. To offer my own view, using *wa* and *noyo* with rising intonations is quite natural for older women when they speak elegantly.

I will now compare the conversations between the *Danwago* Data and the dialogue part of the CSJ (58 conversations, in total 12 hours). Table 5 shows which sentence-final particles appeared at the end of utterances, and indicates their frequency in both data for comparison.

Table 5: Sentence-Final Particles at the end of Utterances

	-wa	*-wane*	*-wayo*	*-noyo*	*-yo*	*-ne*
***Danwago* Data (conversations)**	153	296	116	296	1,675	5,752
CSJ (conversations)	4	2	0	0	391	4,165

Certainly, the numbers (and also the ratios) of instances of *wa, wane, wayo,* and *noyo* in the *Danwago* Data are much greater than those in CSJ. Since the intonation patterns are not considered in these totals, it is difficult to be absolutely certain, but the prediction is that instances with rising intonation are even less frequent in the CSJ particularly.

In any case, just as the instance of the rising intonation patterns shown in the Section 5, these (combinations of) sentence-final particles cannot be seen in the utterances of today's younger generations. This indicates that the usage of these grammatical expressions had gradually disappeared sometime earlier. However, at this stage it is quite difficult to identify the period when such a change in language occurred. To describe the dynamics of language change in accurate detail, it is necessary to supplement our data with more material from the 20th century, and from a greater variety of speech situations.

6 Concluding remarks

In this study, the possibility of compiling and analyzing a diachronic speech corpus of Japanese has been discussed. First, the conditions of a diachronic speech corpus were examined from the viewpoints of the key terms "diachronic," "speech," and "corpus." Also, the limitations of compiling a diachronic speech corpus of Japanese were identified; the amount of old recorded materials is limited, so only a corpus compiled from limited resources can be used for analysis. This is a general constraint which linguists must cope with when analyzing resources for the study of old language.

Following this, several case studies were presented, analyzing intonation patterns and grammatical expressions, auxiliary verbs and sentence-final particles, using three different recorded resources: the Okada Collection, the *Danwago* Data, and the CSJ. The analyses clarified some interesting linguistic findings in spoken Japanese, including

rapid rising intonation, the auxiliary verbs *masu* and *masuru*, and (combinations of) sentence-final particles.

When attempting a diachronic analysis in order to observe historical change in spoken Japanese, a serious problem arises due to the insufficiency and imbalance of existing recorded data, as seen earlier. In this study three different speech data sets were treated as a single diachronic speech corpus, however, the situations of speech and speaking styles vary among these three; the Okada Collection includes political speeches, the *Danwago* Data contains academic lectures and daily conversation, and the CSJ is mainly composed of academic and casual monologues.

In any case, there is no question about the significance of compiling a diachronic speech corpus and followed by analysis. To solve the problem of imbalance, it is crucial that we collect more varied recorded resources in order to establish a better diachronic speech corpus so that studies on spontaneously spoken Japanese in various eras can achieve full fruition.

Acknowledgement

This study is supported by the JSPS (Japan Society for the Promotion of Science), the Grant-in-Aid for Scientific Research (no. 24520523), and Grant-in-Aid for Collaborative Research Project of NINJAL "A multifaceted study of spoken language using a large-scale corpus of everyday Japanese conversation."

Literature

Aarts, B., Bowie, J., and Wallis, S. (2015) Profiling the English verb phrase over time: modal patterns. In: Taavitsainen, I., Kytö, M., Claridge, C., and Smith, J. (eds.) *Developments in English: expanding electronic evidence*, 48–76. Cambridge: Cambridge University Press.

Aizawa, M. and Kanazawa, H. (eds.) (2016) *Senzenki SP record ga hiraku nihongo kenkyū* (Japanese Language Studies by Analyzing SP Records Before World War II). Tokyo: Kasamashoin.

Hattori, T. (2011) *Washa no shussē nendai to hatsuwa jiki ni motozuku gengo henka no kenkyū: kokkai kaigiroku o riyō shite* (A Study on Language Change Based on the Speakers' Date of Birth and Years of Utterances: An Analysis of Minutes of the National Diet). *Kēryō Kokugogaku*, 28 (2): 47–62.

Kanazawa, H. (1991) *Mēji-ki Ōsaka-go shiryō toshite no rakugo sokkibon to SP record*. (Storybooks of Rakugo and SP Records as the Resources of Osaka Dialect in Meiji Era). *Kokugogaku*, 167: 15–28.

Kanazawa, H. (2000) *Rokuon shiryō no rekishi to sono kanōsē*. (The History of Recorded Resources and Their Possibilities). *Nihongogaku*, 19 (11): 197–208.

Kanazawa, H. (2015) *Rokuon shiryō niyoru kindai-go kenkyū no ima to korekara* (The Current Studies on Modern Language by Recorded Resources and Its Future). *Nihongo no Kenkyū*, 11 (2): 133–140.

Koiso, H., Tsuchiya, T., Watanabe, R., Yokomori, D., Aizawa, M., and Den, Y. (2016) Survey of Conversational Behavior: Towards the Design of a Balanced Corpus of Everyday Japanese Conversation, *Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016)*, pp.4434-4439.

Koiso, H., Tanaka, Y., Watanabe, R., and Den, Y. (2016) A Large-Scale Corpus of Everyday Japanese Conversation: On Methodology for Recording Naturally Occurring Conversations, *Proceedings of LREC 2016 workshop on casual talk among humans and machines*, pp. 9-12.

NLRI (The National Language Research Institute) (1955) *Danwago no jittai* (Research in Colloquial Japanese). NINJAL Report 8. NINJAL.

NLRI (The National Language Research Institute) (1960) *Hanashi kotoba no bunkē (1): taiwa shiryō niyoru kenkyū* (A Research for Making Sentence Patterns in Colloquial Japanese: On Materials in Conversation). NINJAL Report 18. Tokyo: Shūēshuppan.

NLRI (The National Language Research Institute) (1963) *Hanashi kotoba no bunkei (2): dokuwa shiryō niyoru kenkyū* (Research of Sentence Patterns in Colloquial Japanese: On Materials in Speech). NINJAL Report 23. Tokyo: Shūēshuppan.

Maekawa, K. (2013) *Corpus no sonzai igi*. (Raison d'être of Corpus). In: Maekawa, K. (ed) *Kōza Nihongo Corpus 1 Corpus Nyūmon* (Lecture Series: Japanese Corpus 1: Introduction): 1–31. Tokyo: Asakurashoten.

Maruyama, T. (2012) *Daikibo Corpus no riyō to meta data no yakuwari*. (Usage of Large-scale Corpus and the Role of Meta Data). The First Corpus Nihongaku Workshop Proceedings, 203–210.

Shimizu, Y. (1988) *Tokyo-go no rokuon shiryō*. (Recorded Resources of Tokyo Dialect). *Kokugo to Kokubungaku*, 65 (11): 129–143.

Shimizu, Y. (1994) *Rokuon shiryō ni kiku 20-seiki hajime no Tokyo-go*. (Tokyo Dialect in early 20th Century: An Analysis of Recorded Resources). *Annual Report of the Institute for Japanese Culture and Classics*, Kokugakuin University, 73: 191–230.

Shimizu, Y. (2011) *Ōbē no rokuon archives: shoki nihongo rokuon shiryō shozō kikan o chūshin ni*. (Recorded Archives in Europe and the USA: Institutes to Collect Recorded Resources in Early Japanese). *Kokubun Mejiro*, 50: 29–19.

Shimizu, Y. (2014) *Hyakunen mae no nihongo o kiku*. (Listening to Japanese in 100 years ago). Japan Women's University.

要旨（Abstract in Japanese）

「通時音声コーパスの可能性」

丸山岳彦（専修大学／国立国語研究所）

「話し言葉の通時コーパスは実現可能か」という問題について論じる。通時コーパスと言えば、通常、書き言葉を対象としたものが想定される。それでは、話し言葉を対象とした通時コーパスは実現可能であろうか。本稿では、「通時」「音声」「コーパス」という3つの条件について検討し、「通時音声コーパス」の実現によってどのようなことが明らかになるかを示す。大正から昭和前期にかけて録音されたSPレコードの音源資料、国立国語研究所において1950年代に録音された資料などを分析対象として、そこに見られるイントネーションの型、文法形式について分析の事例を示し、話し言葉の経年変化を追うための「通時音声コーパス」が持つ可能性について論じる。