Speech corpora in NINJAL, Japan demonstration of corpus concordance systems : Chunagon and Kotonoha

The 3rd International Symposium on Linguistic Patterns in Spontaneous Speech (LPSS2019)
Institute of Linguistics, Academia Sinica, Taipei
November 21-22, 2019

Speech Corpora in NINJAL, Japan
Demonstration of Corpus Concordance Systems: Chunagon and Kotonoha

Hanae Koiso[1], *Masayuki Asahara*[1], *Salvatore Carlino*[1], *Ken'ya Nishikawa*[1],

Kazuki Aoyama[1], *Yuichi Ishimoto*[1], *Aya Wakasa*[1], *Michiko Watanabe*[1], *Yoshimi Yoshikawa*[1],

Nobuko Kibe[1], *Kikuo Maekawa*[1]

[1]National Institute for Japanese Language and Linguistics, Japan

kotonoha at ninjal dot ac dot jp

Abstract

The National Institute for Japanese Language and Linguistics, Japan (NINJAL, Japan) provides a demonstration site in the LPSS 2019 conference. This manuscript presents an overview of the demonstration of three corpora: Corpus of Spontaneous Japanese, Corpus of Everyday Japanese Conversation, and Corpus of Japanese Dialects.

NINJAL also demonstrates two concordance systems. The first is "Chunagon (中納言)" which is a morpheme based concordance system that was made publicly available in 2011. The second is the currently developing system "Kotonoha" released in 2018 that enables query of multiple corpora in terms of register type and period.

1. Introduction

The National Institute for Japanese Language and Linguistics (NINJAL), Japan, has developed several corpora in various registers for linguistic research. NINJAL also provides two corpus concordance systems: "Chunagon" (中納言), and "Kotonoha". The corpus concordance systems are developed in order to provide query services for the following NINJAL maintained Japanese corpora:

- Corpus of Spontaneous Japanese (CSJ) [1]
- Corpus of Everyday Japanese Conversation (CEJC) [2]
- Corpus of Japanese Dialects (COJADS) [3,4,5]
- Balanced Corpus of Contemporary Written Japanese (BCCWJ) [6]
- Corpus of Historical Japanese (CHJ) [7]
- International Corpus of Japanese as a Second Language (I-JAS) [8]
- Nagoya University Conversation Corpus (NUCC) [9]
- Gen-Nichi-Ken Corpus of Workplace Conversation (Shokuba) [10]
- NINJAL Web Japanese Corpus (NWJC) [11]

NINJAL also provides another corpus concordance system "BonTen" (梵天) [12] for the written corpora of BCCWJ and NWJC. Since we focus on speech corpora, we omit the demo of "BonTen". This paper is organized follows. Section 2 describes the concordance systems. Section 3, 4, and 5 provide detailed descriptions of CSJ, CEJC, COJADS, respectively.

2. Corpus Concordance Systems

This section provides an overview of the developed corpus concordance systems developed by NINJAL.

2.1. NINJAL corpus design

Regarding NINJAL's corpus design, first, original data, such as speech, video or document, are textized. Speech or video data are manually transcribed into text data. Documents such as newspapers, books, and magazines are digitized and have their OCR errors manually modified.

Second, morphological information is annotated on the data. The morphological information is two layered based on the lexical segmentation unit. A Short Unit Word (SUW: 短単位) is designed for the examination of individual elements, while a Long Unit Word (LUW: 長単位) is designed for the examination of linguistic properties. While SUW is suitable for word frequency counting, LUW is utilized to define the "Bunsetsu" (文節) segmentation, which is a fundamental lexical segmentation unit for syntax.

Some corpora include corpus-specific information. For example, CSJ and BCCWJ include Bunsetsu-based syntactic dependencies, COJADS includes standard Japanese translation, and I-JAS includes error and correction information of L2 learners.

2.2. Corpus Concordance System "Chunagon" (中納言)

The Chunagon corpus concordance system enables query of the CSJ, CEJC, COJADS, BCCWJ, CHJ, I-JAS, NUCC and Shokuba corpora. The system was made publicly available in 2011 as a search system for BCCWJ. Since then, the system has been improved to incorporate other corpora. The number of registered users are over 22,000 .

Chunagon has two query systems. One is string search, used to explore "strings" in the transcribed or digitized texts. Regular expressions can be used in the query. The other is query by morphological information, in which the query can be formed based on lexical segmentation unit, surface form, part-of-speech, conjugation information with the contextual words. The query can be built using a GUI-based query builder (Figure 1).

This concordance system has two characteristics in terms of corpus provision. The speech corpus is originally developed as a sound dataset that allows the user to play the sound data of

CSJ, CEJC, and COJADS (Figure 2). The time span of the sound data is linked to the script text regions. When the corpus includes supplemental resources such as metadata and annotations, the user can download the files through the concordance system.

Figure 1: GUI-based Query Builder on Chunagon (Query by Morphological Information)

Figure 2: Result View on Chunagon (with playing audio)

2.3. Integrated Search System "Kotonoha"

Kotonoha is an integrated search system based on Chunagon. Chunagon incorporates varied corpora by period and register type. The user needs to explore each corpus individually. Kotonoha, on the other hand, enables users to explore the multiple corpora based on their morpheme and count example frequency by period or register type. Figure 3 shows the diachronic result view on Kotonoha of a query ("赤" red) by period from the Nara period to the contemporary period. Figure 4 shows a query ("とても" very) by register type (spoken vs. written) and speaker (native vs. non-native).

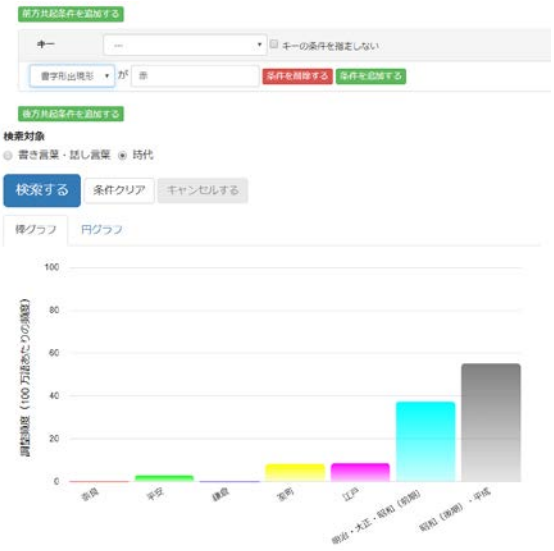

Figure 3: Result View of Kotonoha (by Period)

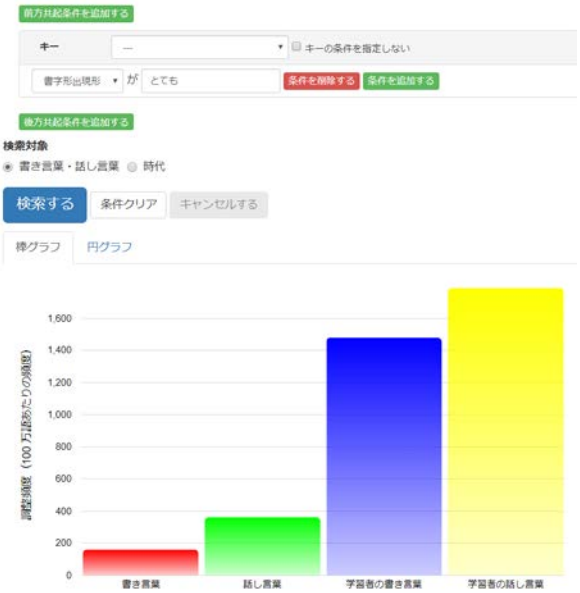

Figure 4: Result View of Kotonoha (by Register Type)

3. Corpus of Spontaneous Japanese, CSJ

The *Corpus of Spontaneous Japanese*, or CSJ, is a large-scale database of spontaneous Japanese developed through the collaboration between NINJAL, National Institute of Information and Communications Techonology, Japan (NICT), and Tokyo Institute of Technology (TITECH). It has been used for a wide variety of research purposes, such as spoken language processing, natural language processing, phonetics, psychology, sociology, Japanese education, and dictionary compilation. Since its release in 2004, approximately 1,000 copies have been sold. In 2016, CSJ was incorporated into Chunagon, which allows us to search the corpus for morphological information.

3.1. Sources: APS and SPS

As shown Table 1, most of the speech material comprises spontaneous monologs, of which the two main types are Academic Presentation Speech (APS), which includes the live recording of academic talks held in various academic societies, and Simulated Public Speaking (SPS), which includes laypeople's talks about everyday topics, such as "my most delightful memory." The remaining part of CSJ is devoted for dialogues and readings, which were recorded for comparison with the monolog part.

Table 1: Speech Types of CSJ

Speech Type	Mode	N Speakers	N Talks	Hour
APS	Monolog	819	987	274.4
SPS	Monolog	594	1715	329.9
Other presentations	Monolog	16	19	24.1
Interviews on APS	Dialog	(10)*	10	2.1
Interviews on SPS	Dialog	(16)*	16	3.4
Task-oriented dialogs	Dialog	(16)*	16	3.1
Free dialogs	Dialog	(16)*	16	3.6
Reproductions	Reading	(16)*	16	5.5
Readings	Reading	(248)*	507	15.5
Total		1417	3302	661.6

*Numbers in parenthesis are counted as speakers of APS or SPS

3.2. Layered Structure: the Core

In its entirety, CSJ comprises digitized speech (16kHz, 16bit), transcription, morphological information (SUW and LUW), clause boundary, impressionistic rating (subjective evaluation by the listener of the way a talk is being spoken), and information about the speakers and talks for 7.5 million words (SUWs) or 660 hours (Figure 5). In addition, there is a true subset of the corpus known as the Core, which concentrates the cost of annotation. The CSJ-Core consists of talks of 500K words (SUWs) or 44 hours, and includes segmental and intonation labels (section 3.3), dependency structure, and discourse segment.

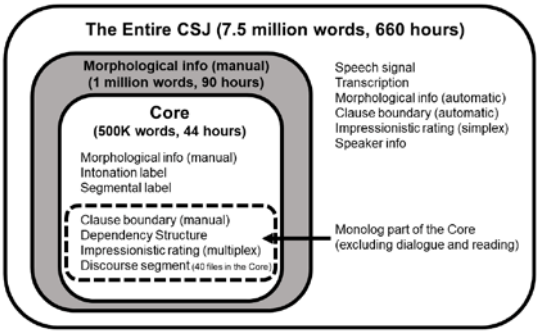

Figure 5: Layered structure of CSJ

3.3. Segmental and Intonation Labelling: X-JToBI

All speeches on the CSJ-Core are annotated in terms of segmental and intonational characteristics using the X-JToBI [13] annotation scheme (Figure 6), which is an extension for spontaneous speech of the original J_ToBI (Japanese Tones and Break Indices). Among the six tiers – Word, Segment, Tone, Break Index (or BI), Prominence, and Miscellaneous – of the X-JToBI annotation, the last four tiers are of special interest for intonation labeling. "Word" in X-JToBI is a SUW. "Segment" contains allophonic labels (e.g., "kj"), auxiliary labels (e.g., "<cl>"), and fused labels (e.g., "N,m").

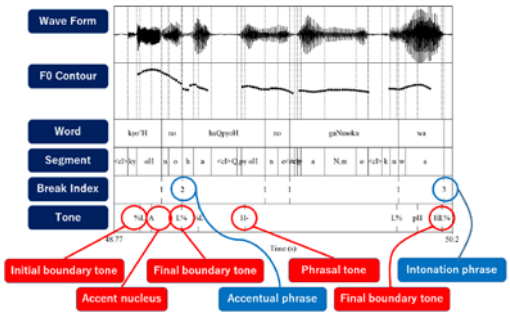

Figure 6: An example of intonation labelling (with prominence and miscellaneous tiers omitted)

3.4. Relational Database: CSJ-RDB

To enable complex searches easily and efficiently, NINJAL implemented a relational database version of CSJ (CSJ-RDB) [14] for the CSJ-Core. CSJ-RDB contains all types of information annotated to the CSJ-Core in the form of tables. As shown in Figure 7, hierarchical structures are assumed on CSJ-RDB.

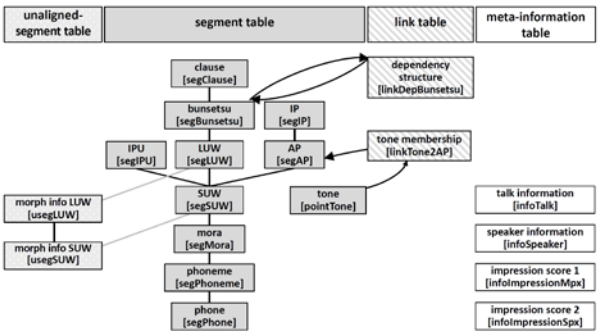

Figure 7: CSJ-RDB data structure

For example, the search for the boundary tones ("falling," "rising," etc.) of the final accentual phrases in all clauses ending with the final particle "ne" would be rather difficult without CSJ-RDB, because they belong to different annotations – Intonation and morphological (SUW) annotations. CSJ-RDB enables this search by simply joining several tables. The simple data structure of CSJ-RDB significantly facilitates corpus management and searching.

4. Corpus of Everyday Japanese Conversation, CEJC

In 2016, NINJAL started the compilation of a large-scale corpus of everyday Japanese conversation, the *Corpus of Everyday Japanese Conversation* (CEJC). Prior to the publication of the entire corpus scheduled for 2022, NINJAL published the trial version of CEJC, CEJC-T, in December 2018. This section briefly introduces the CEJC-T design.

4.1. Recording Method

CEJC's main features are: i) the focus on conversations embedded in daily, naturally occurring activities; ii) the collection of various types of everyday conversations in a balanced manner; and iii) the collection and publication of both audio and video data.

To record naturally occurring conversations in daily situations, a recording method called individual-based method was adopted. Following this method, NINJAL recruited 40 informants balanced in terms of gender and age (man/woman × 20s/30s/40s/50s/over 60 × 4 informants), provided them with portable recording devices for approximately two to three months, and had them record about 15 hours of conversations of their daily activities. Informants were required to carry the portable recording devices and record their daily activities in a variety of situations, such as at home, at a restaurant, and outdoors. About four to five hours out of 15 hours of conversation per informant were selected for the entire CEJC by considering the balance of conversation variations.

Figure 8 shows video footage of a conversation between a customer and a barber at a barbershop. The left image was recorded using a Kodak PIXPRO SP360 4K camera, while the top- and bottom-right images were recorded using two GoPro cameras. As for speech, the two conversants wore IC recorders and recorded their voices using their own recorders.

Conversants' voices were also recorded by another IC recorder located in the center of the room.

Figure 8: Video footage of a conversation between a customer and a barber at a barbershop

4.2. Size and Structure

CEJC-T consists of conversations collected by 20 out of 40 informants, as shown in Table 2.

Table 2. Attributes of informants in CEJC-T

age	gender		total
	male	female	
20s	student student	student student	4
30s	self-employed office worker	office worker housewife	4
40s	office worker freelance	office worker part-time part-time	5
50s	manager office worker	office worker self-employed	4
over 60	retired person teacher	housewife	3
total	10	10	20

The final version of CEJC comprises about 200 hours of conversations, while CEJC-T includes 50 out of these 200 hours. Total numbers are 126 conversations, 392 conversants (including 237 different participants), and 609,327 words (SUW, see Section 2.1). Of the 126 conversations, 91 are chats, 26 are business talks/consultations, and 9 are meetings.

Figure 9 shows the distribution of conversants by age and gender in the CEJC-T. From the figure, it is found that there are few children under 20 years old. This is because children under the age of 20 were not recruited as informants since the individual-based recording method places a high responsibility on informants.

The speech material was manually transcribed and a POS (part-of-speech) analysis based upon SUWs was automatically conducted and manually corrected. CEJC-T is published (1) on a hard disk containing video and audio data, transcription files (tsv, Praat, and ELAN formats), and SUWs, and (2) on the Chunagon corpus concordance system.

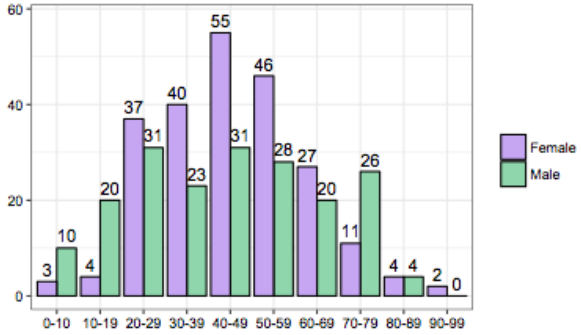

Figure 9: Distribution of informants by age and gender

5. Corpus of Japanese Dialects, COJADS

5.1. Overview

The Japanese archipelago is very rich in dialectal variations. Although there are many detailed language atlases and dialect dictionaries, there was no a multi-dialect natural conversation corpus that could meet the demands of research and the general public. A large corpus is needed not only for linguistic research, but also to preserve the dialects of Japan, many of which are endangered. To meet these demands, NINJAL started the constructing a large-scale dialect corpus that would cover the entire archipelago, from Hokkaido to the Ryukyu Islands was started. This corpus would later become COJADS (COrpus of JApanese DialectS) [3,4,5].

5.2. The Original Recordings

Figure 10 A Volume of the *Nihon no furusato kotoba shuusei* Series.

The original data used on COJADS was collected through the "Urgent survey of Japan's regional dialects" *(Kakuchi hougen shuushuu kinkyuu chousa)*, a nationwide investigation project that aimed primarily at collecting natural conversation data between native speakers of dialects. The investigation was conducted between 1977 and 1985 by the Agency for Cultural Affairs and covered all 47 prefectures of Japan, including data collected from 200 areas, and more than 4000 hours of conversation. The informants were over 50 years old at the time of the recording and would be in their 90s today. Only part of these documents has been published in the series of volumes *Zenkoku hougen danwa deetabeesu - Nihon no furusato kotoba shuusei* (Figure 10) [15]. The COJADS project is intended to provide these valuable data, including the huge unpublished portion, in the form suitable for research.

5.3. Structure

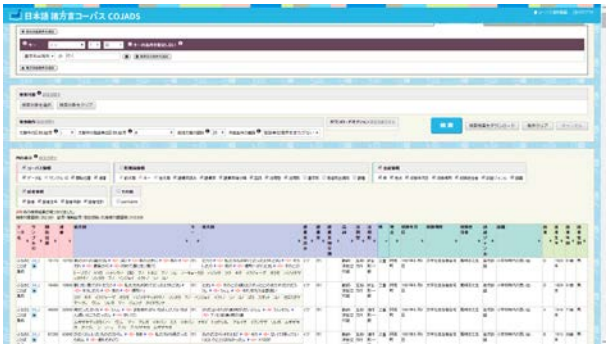

Figure 11: COJADS Search Result View.

COJADS is a parallel corpus consisting of standard Japanese text and dialect texts and audio. Standard Japanese texts are written in a mixed kanji and hiragana script, while the katakana script is used for dialect texts. In some cases, other characters, such as Latin characters, are used to represent sounds that are unrepresentable with katakana. Search is possible in both Standard Japanese and regional dialects, but dialect-form search function is limited to character string search. In the original Standard Japanese text data, proper names, fillers, personal pronouns, untranslatable words, onomatopoeia, and cases of particle omission are tagged as such. In case there is no Standard Japanese equivalent, sentence final particles, adverbial particles, and diminutives are also tagged in the dialect text. Search by tag is not available in the beta version, with the exception of particle omission, which is marked by square brackets in Standard Japanese texts. The corpus can be searched using the Chunagon corpus concordance system (Figure 11). It is also possible to filter search results by area. Search results can be downloaded in an excel file. Metadata including the speaker's gender, age, and year of birth, the recording location, and the conversation theme are available with the research results. Currently, the beta version available has conversation data from 48 areas (two from the Okinawa prefecture and one for each of the other 46 prefectures of Japan). There are 30 minutes of conversations for each area, for a total of 24 hours.

5.4. Plan

A beta version was launched in March 2019. The final version is expected to be made public in 2021, and will include over 75 hours of recorded conversations, search by tag.

6. Acknowledgements

This work was supported by JSPS KAKENHI Grants Number 16H01933, 17H00917, and 18H05521; NINJAL Collaborative research projects 'Endangered Languages and Dialects in Japan' (Project Leader: Nobuko Kibe), 'A Multifaceted Study of Spoken Language Using a Large-scale Corpus of Everyday Japanese Conversation' (Project Leader: Hanae Koiso) and projects of the Language Variation Division (COJADS team), Spoken Language Division, and Center for Corpus Development, NINJAL.

7. References

[1] Maekawa, K. "Corpus of Spontaneous Japanese: Its Design and Evaluation", Proceedings of ISCA and IEEE Workshop on Spontaneous Speech Processing and Recognition (SSPR2003), Tokyo, pp.7-12, 2003:4.

[2] Koiso, H., Y. Den, Y. Iseki, W. Kashino, Y. Kawabata, K. Nishikawa, Y. Tanaka, and Y. Usuda, "Construction of the Corpus of Everyday Japanese Conversation: An Interim Report," Proceedings of the 11th edition of the Language Resources and Evaluation Conference, 4259-4264, 2018.

[3] Kibe, N., Otsuki, T., and Sato K. "Intonational Variations at the End of Interrogative Sentences in Japanese Dialects: From the "Corpus of Japanese Dialects"", Proceedings of LREC 2018 Special Speech Sesssions, 21-28, 2018.

[4] Nakazawa K., Otsuki T., Kamimura K.,Carlino S., Sato K., Kibe N. "Regional differences in patterns of contracted forms with a focus-marking particle or case-marking particles in Japanese dialects: A study based on the "Corpus of Japanese Dialects (COJADS)." *Conference papers of the Dialectological Circle of Japan*, No. 108, 53-60, 2019

[5] Otsuki T., Kamimura K., Carlino S., Sato K., Nakazawa, K. Kibe, N. "Koopasu wo tsukatta hougen kenkyuu no kaitaku. Nihongo shohougen koopasu (COJADS) monitaaban wo tsukatte." *Conference papers of the 2019 Spring Meeting of the Society for Japanese Linguistics.* 181-186, 2019.

[6] Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., and Den, Y., "Balanced Corpus of Contemporary Written Japanese". *Language resources and evaluation*, 48:2, 345-371, 2014.

[7] Ogiso, T. ""Nihongo Rekishi Corpus" no genjou to tenbou", *Kokugo to kokubungaku*, 93:5, 72–85, 2016.

[8] Sakoda, K., Konishi, M., Sasaki, A., Suga, W., Hosoi, Y., "International Corpus of Japanese as a Second Language", *NINJAL Project Review*, 6:3, 93-110, 2016.

[9] Fujimura, I., Chiba, S., Ohso, M., "Lexical and Grammatical Features of Spoken and Written Japanese in Contrast: Exploring a lexical profiling approach to comparing spoken and Written corpora", *Proceedings of the VIIth GSCP International Conference. Speech and Corpora*, 393-398, 2012.

[10] Kashino, W., Omura, M., Nishikawa, K., and Koiso, H., "Supplemental Arrangement for Public Data Available in the Chunagon Versions of "Gen-Nichi-Ken Corpus of Workplace Conversation"", *Proceedings of Language Resources Workshop 2018*, 494-509, 2018.

[11] Asahara, M., Maekawa, K., Imada, M., Kato, S., and Konishi, H., "Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan", *Alexandria*, 26:1-2, 129-148, 2014.

[12] Asahara, M., Kawahara, K, Takei, Y., Masuoka, H., Ohba, Y., Torii, Y,. Morii, T., Tanaka, Y., Maekawa, K., Kato, S., and Konishi, H., "'BonTen' - Corpus Concordance System for 'NINJAL Web Japanese Corpus'" *Proc. of COLING-2016 Demo Session*, 2016.

[13] Maekawa, K., Kikuchi, H. , Igarashi, Y., and Venditti, J. "X-JToBI: An extended J_ToBI for spontaneous speech." *Proc. 7th International Congress on Spoken Language Processing (ICSLP2002)*, 1545-1548, 2002.

[14] Koiso, H., Den, Y., Nishikawa, K., and Maekawa, K., "Design and Development of an RDB Version of the Corpus of Spontaneous Japanese," *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, 1471-1476, 2014.

[15] Kokuritsu kokugo kenkyuusho, *Zenkoku hougen danwa deetabeesu - Nihon no furusato kotoba shuusei.* Tokyo: Kokusho Kankoukai 2001-2008.