Proceedings of the Language Technologies for All (LT4All) , pages 371–374
Paris, UNESCO Headquarters, 5-6 December, 2019.
© 2019 European Language Resources Association (ELRA), licenced under CC-BY-NC

Archiving System of Endangered Languages in Japan: A Preliminary Report

Natsuko Nakagawa, Masahiro Yamada, Nobuko Kibe, Yukinori Takubo
National Institute for Japanese Language and Linguistics
10-2 Midoricho, Tachikawa, Tokyo, Japan
{nakagawanatuko, m-yamada, nkibe, ytakubo}@ninjal.ac.jp

Abstract

There are eight (UNESCO), twelve (Ethnologue), or more (degree of intelligibility) endangered languages/dialects in Japan. We present a database and digital archiving space that NINJAL (National Institute for Japanese Language and Linguistics) is developing for all of these languages where individual researchers or language communities can deposit their field data, language documentation, or audio-visual recordings. Two major features of the database/archiving space include (i) that it is a Japanese-mediated database/archive and thus virtually everyone in Japan can use it, and (ii) that it comes with an online exhibition space so that archiving is tightly connected to public use of the deposited items.

日本には 8（ユネスコ）、12（エスノローグ）、もしくはそれ以上（相互理解性）の消滅の危機に瀕した言語・方言が存在する。本発表は国立国語研究所がこれらの言語・方言のために開発中の、個別の研究者や言語コミュニティが利用可能なデータベースおよび電子的アーカイブスペースについて報告する。データベース・アーカイブスペースは以下の二つの特徴を持つ。(i) 日本語によるデータベース・アーカイブスペースであり、日本に住む誰もが利用可能である。(ii) オンライン展示スペースが付随し、アーカイブされるデータが社会一般に対する公開と密接に結びついている。

Keywords: digital archiving, exhibition, collaboration with communities, languages in Japan

1. Introduction

Atlas of the World's Languages in Danger (Moseley, 2010) reports that there are eight endangered languages in Japan: Ainu, and seven Japonic languages, namely Hachijo, and six Ryukyuan languages (Figure 1.). Although these eight languages tend to receive a special attention with respect to language conservation, there are many more mutually unintelligible sub-varieties including Japanese dialects. Indeed, there have been eight reference grammars for Ryukyuan and three for Japanese dialects, and many more grammar sketches (Shimoji, 2008; Pellard, 2009; Niinaga, 2014; Shimoji and Pellard, 2010; Heinrich et al., 2015). Under the influence of the dominant language, so-called Standard Japanese, most, if not all, of such local languages are in danger of extinction.

Since 2010, National Institute for Japanese Language and Linguistics (NINJAL) has been leading a collaborative research project termed the *Kiki-Gengo Project* (Endangered Language Project, PI: Nobuko Kibe) to describe and conserve the local languages in Japan. As of 2019, 57 researchers from 36 domestic and overseas universities/institutions are working on documentation of the local languages spoken at 40 sites across Japan and producing vocabulary lists, grammar sketches, and narratives. The project also collaborates with some of the local governments to raise the awareness of the language endangerment.

In this paper, we will first present the currently available language resources resulting from the NINJAL project. We also introduce an ongoing project of language archiving and an exhibition system that targets researchers and language communities.

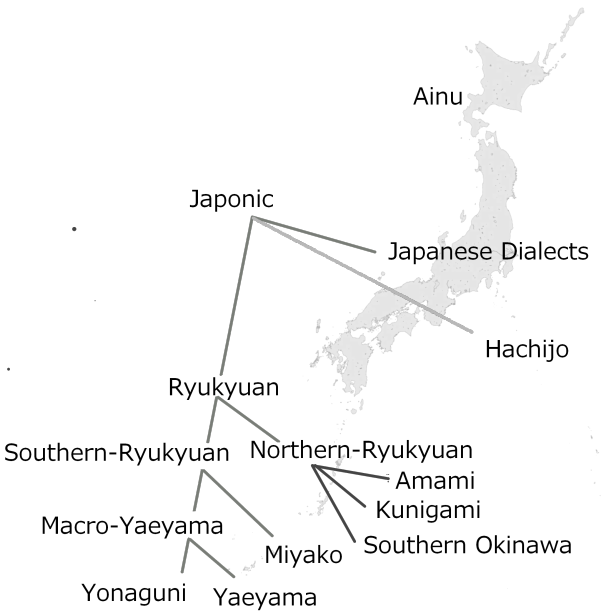


Figure 1: Language map in Japan

2. Current Resources

The documentation works done by the collaborative researchers in NINJAL's *Kiki-Gengo Project* are incrementally made available at *Kiki-Gengo Database* (Endangered Language Database). NINJAL also provides a corpus of Japanese dialects (COJADS). This section gives a brief overview of these resources. Also, some of the colleagues made a set of fonts called "Shima Shotai Font" in collaboration with a designer so that everyone can type their languages on computer. We give an overview of these re-

sources.

2.1. Endangered Language Database

The *Kiki-Gengo* (endangered language) Database is a website to make NINJAL's collaborative researchers' documentation works accessible online (Celik and Kibe, 2019).[1] It provides vocabularies and narratives site-by-site and currently covers 14 sites (mainly Hachijo and the Ryukyus). Approximately 9,000 words and expressions are currently registered and searchable online. In addition to the recordings collected by the collaborative researchers, some legacy recordings of dialogues and monologues collected as old as 1969 are also available.

Grammar sketches will be made available at the website by the end of the current project period, 2021. For now we post pdf versions of the grammar sketches along with a basic vocabularies and sentences at nine sites.

Currently the data depositing is done by a few project RAs by hand and is not quite functioning as an ideal archiving system, but the process is fairly simple for the collaborative researchers for they just need to submit their field data filled in a spreadsheet template along with the audio files.

2.2. Corpus

Corpus of Japanese Dialects, or COJADS is the largest corpus of Japanese dialects currently available (Aoyama et al., 2019; Kibe et al., to appear). Currently, the beta version is available with the data from 48 sites (two from Okinawa prefecture and one for each of the 46 prefectures of Japan); 30 minutes of narratives for each site, for a total of 24 hours (published as NINJAL (2001-2008)).

The corpus is searchable both in transcriptions and in Standard Japanese translations at NINJAL's Chunagon corpus concordance system.[2] The audio data corresponding to the search result is also presented on the same page.

The original data used on COJADS consists of more than 4,000 hours of narratives collected from 224 locations through the "Urgent Survey of Japan's Regional Dialects", a nationwide investigation project conducted between 1977 and 1985 by the Agency for Cultural Affairs. COJADS aims to cover 75 hours of narratives by 2021.

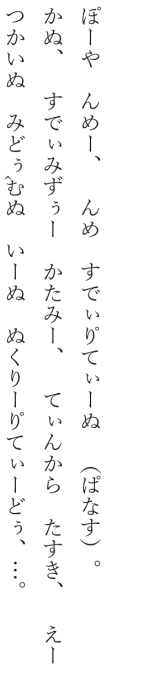


Figure 2: Shima Shotai Font

2.3. Shima Shotai Font

Many languages/dialects in Japan have sounds that cannot be written in Japanese writing, which is based on the Standard Japanese sound system. Ogawa (2015) proposed a unified writing system to express all sounds in Ryukyuan varieties. The proposed system, however, includes unconventionalized usage of diacritics (such as ˚ , ˇ , and ^) and needs a new type of font. Ogawa et al. (2019) proposed new font termed "Shima Shotai Font" (*shima* means "(our) island" and "(our) village" across Ryukyuan islands), which includes unconventionalized combinations of diacritics and kana (Japanese phonogramic characters). See Figure 2 for an example. It is *kana*-based phonogramic system instead of *kanji*-based ideogramic system because the pronunciation of *kanji* is often ambiguous. The set of fonts makes it easier for communities to use their languages on PCs (and smartphones in the future), which is expected to promote their language usage. Vertical writing is also possible with this set of fonts. In the future, Ogawa and his colleagues want the font to be included in Unicode fonts.

3. Archiving and Exhibition Space

This section discusses our ongoing project and future plan.

3.1. Language Archiving System

We are now building a language archiving system both for researchers and for community members. The system should be Japanese-mediated so that almost all citizens in Japan can access it. It should be easily depositable and immediately visible/usable; i.e., it should be easy for everyone to deposit their data and to see and search the data immediately after they deposit them.

3.2. Language Museum

The language archiving system accompanies with exhibition space, called "digital museum of language", which exhibits part of the archived data mainly for non-researchers (Takubo, 2017). This kind of exhibition space not only encourages the community members to be proud of their community, but also motivates researchers to deposit more data since researchers receive feedback from the community. The goal is to receive positive feedback, but should the feedback be negative, researchers would strive to improve the content.

We plan to show our materials with coherent stories. For example, users interested in the history of Miyako Island will be able to access related videos in chronological order, while users interested in sightseeing can access videos introducing popular locales and other nearby attractions. Multiple stories will be attached to the same video, accessible based on the users' interests. In the future, we would like to make it possible for everyone to make their own stories, arrange materials in the museum, and publish them.

Materials for Miyako language (Ikema-Nishihara variety) learning are already partially available.[3] They cover pronunciation, the writing system, grammar, and videos. We will improve this website to be more user-friendly.

4. For and with Communities

This section introduces some projects we are working on with community members.

4.1. Picture books with audio, video, and text

We and other colleagues (Akiko Tokunaga-Yokoyama, Kayoko Shimoji, and Fumi Yamamoto) are making four

[1] http://kikigengo.ninjal.ac.jp
[2] https://chunagon.ninjal.ac.jp
[3] http://www.kikigengo.jp/nishihara/doku.php



Figure 3: Drawings from a picture book

picture books for children in four different Ryukyuan languages. They are based on the local stories or songs inherited from their ancestors. The pictures were drawn by Fumi Yamamoto, a professional illustrator and designer, who happened to be attracted to one of the stories in the Ryukyus (see Figure 3).

Each story was narrated by a local native speaker, was transcribed by a linguist who specializes in the language, and was accompanied with interlinear gloss and translations (in Japanese).[4] The linguists also wrote grammar sketches to help readers to read aloud the text and understand the story in the given language (Nakagawa and Yamada, 2018). Narration by the native speaker was recorded and will be available online so that people can learn how to pronounce each word and how to read a sentence with natural intonation. The books have a kid-friendly appearance and content, but it is also intended that the parent generation can practice the language while they read them to their children. The parent generation are considered to be passive speakers; they can understand the language but do not speak it (Yamada et al., to appear). The picture books are expected to encourage them to speak their language.

[4]English translations will be also available.

We are now raising a fund via crowdfunding to publish the books.[5] 213 people donated 3,032,000 Japanese yen (approximately 28,000 US dollars) as of 13th January, 2020. Comments left by the supporters tell that the supporters are not necessarily related to Ryukyu.

4.2. Dictionary

Traditionally, a single native speaker devotes their life to compiling a dictionary, and it usually takes decades to publish it. Now, younger community members collect words and expressions from elderly speakers with a help of linguists and input the collected items in Google Docs. Another linguist in Tokyo receives the data immediately, change the spread sheet into a database and makde it available online. Another linguist prints the data in a dictionary-like format (using LaTeX), and send it back to the community. In the future, we want to make an online dictionary maker which automatically generates a professional dictionary from Google Docs (or tab-separated file).

4.3. Interactive interlinear glossed transcriptions (Karaoke)

We are making a system of interactive interlinear glossed transcript called *karaoke*. For now, only karaoke system with sounds (without video) is available.[6] In this system, users can not only see the subtitles while watching a video or listening to narratives, but also can click the subtitles and watch and listen to the corresponding part. Users can also chose what they see; they can watch a video or listen to narratives without any subtitles or they can chose to see one of the transcriptions (IPA, *kana*-based style, and/or only translations). In this way, the karaoke system functions as an efficient, customized learning tool. It can also be applied to songs.

4.4. Physical “escape the room” game

We (Niinaga Yuto, Masahiro Yamada, and Natsuko Nakagawa) are also making a game called physical “escape the room” game with Igengo Lab. “Escape the room” is a kind of game played with a book or, more recently, a computer. The player is confined to a room and is required to escape from the room by collecting hints hidden in the room. The hints are often encrypted, and the player needs to infer the solution. In the physical escape the room game, the players actually play the game physically. They (usually a team) are confined in a room, seek hints often encrypted, and look for the solution.

Igengo Lab is one of such game-developping groups. They specialize in Japanese Sign Language and have made various problems (e.g., encrypted hints) that can be solved only when signers and non-signers collaborate with each other.

We and Igengo Lab made a game using three varieties of Ryukyuan: Yonaguni Dunan, Shiraho (Ishigaki, Yaeyama), and Amami (See also Figure 1.). The players are required to compare these three languages, find phonological correspondence and some grammatical rules, and communicate with local people in the local language in order to solve the problems. We will run the first trial in February 2020 in

[5]https://readyfor.jp/projects/minato

[6]http://kikigengo.ninjal.ac.jp/danwashiryo.html

Tokyo and hope to continue in the Ryukyu Islands in the future.

5. Conclusion

This paper gave a brief overview of NINJAL projects related to endangered languages in Japan. The resources already available are a vocabulary database (9000 expressions from 14 places, contemporary), spoken corpora (75 hours from 25 places, approximately 40 years ago), and the Shima Shotai Font, expected to be more ubiquitous and be registered as Unicode fonts. We are now building an archiving system with exhibition spaces that are mediated in Japanese. Finally, we introduced four ongoing projects with communities: picture books, dictionaries, a karaoke system, and a physical "escape the room" game.

6. Acknowledgements

The work reported in this article was supported by JSPS KAKENHI projects 16H01933, B16K16824, and 18K12360.

7. Bibliographical References

Aoyama, K., Asahara, M., Carlino, S., Ishimoto, Y., Kibe, N., Koiso, H., Maekawa, K., Nishikawa, K., Wakasa, A., Watanabe, M., and Yoshikawa, Y. (2019). Speech corpora in NINJAL, Japan: Demonstration of corpus concordance systems: Chunagon and Kotonoha. In *The 3rd International Symposium on Linguistic Patterns in Spontaneous Speech*, Taiwan. Institute of Linguistics, Academia Sinica.

Celik, K. and Kibe, N. (2019). Raising language diversity awareness in Japan through web-based open access application. In *6th International Conference on Language Documentation and Conservation (ICLDC)*.

Patrick Heinrich, et al., editors. (2015). *Handbook of the Ryukyuan Languages*. Mouton de Gruyter, Berlin.

Kibe, N., Sato, K., Nakanishi, T., and Nakazawa, K. (to appear). Corpus-based study of japanese dialects: Regional differences in accusative case marking system. In *Proceedings of Methods in Dialectology XVI.*

Christopher Moseley, editor. (2010). *Atlas of the World's Languages in Danger*. UNESCO Publishing, Paris.

Nakagawa, N. and Yamada, M. (2018). Taketomi-jima hoshi-suna-no hanashi-no ehon seisaku-to ippan-dokusha-muke bumpô-gaiyô-no shippitsu [making the picture book a tale of star sand in taketomi and writing a grammar sketch for a general audience]. *NINJAL Research Papers*, 14:145–167.

Niinaga, Y. (2014). *A grammar of Yuwan, a Northern Ryukyuan language*. Ph.D. thesis, University of Tokyo.

Ogawa, S., Yamada, M., Hayashi, Y., and Ueda, H. (2019). A typeface for endangered languages. In *ATypI2019*.

Shinji Ogawa, editor. (2015). *Ryukyu-no Kotoba-no Kakikata [How to write Ryukyuan languages]*. Kurosio, Tokyo.

Pellard, T. (2009). *Ōgami – Éléments de Description d'un Parler du Sud des Ryūkyū*. Ph.D. thesis, École des hautes études en sciences sociales. (Ōgami – Elements of description of a Southern Ryukyuan language).

Michinori Shimoji et al., editors. (2010). *An introduction to Ryukyuan languages*. ILCAA, Tokyo.

Shimoji, M. (2008). *A Grammar of IRabu, a SOuthern RYukyuan Language*. Ph.D. thesis, Australian National University.

Takubo, Y. (2017). The digital museum project for the documentation of endangered languages: The case of Ikema Ryukyuan. In A. Vovin et al., editors, *Studies in Japanese and Korean historical and theoretical linguistics and beyond*, pages 3–12. Brill, Leiden.

Yamada, M., Takubo, Y., Iwasaki, S., Celik, K., Harada, S., Kibe, N., Lau, T., Nakagawa, N., Niinaga, Y., Otsuki, T., Sato, M., Shirata, R., Van der Lubbe, G., and Yokoyama, A. (to appear). Experimental study of inter-language and inter-generational intelligibility: Methodology and case studies of Ryukyuan languages. *Japanese/Korean Linguistics*, 26.