Language Atlas of Japanese and Ryukyuan (LAJaR): A Linguistic Typology Database for Endangered Japonic Languages

Kanji KATO[*]
jiteng.ganzhi@gmail.com

So MIYAGAWA[†]
runa.uei@gmail

Natsuko NAKAGAWA[†]
nakagawanatuko@gmail.com

Abstract

LAJaR (Language Atlas of Japanese and Ryukyuan) is a linguistic typology database focusing on micro-variation of the Japonic languages. This paper aims to report the design and progress of this ongoing database project. Finally, we also show a case study utilizing its database on zero copulas among the Japonic languages.

1 Introduction

Linguistic typology databases have been created for describing and comparing languages of the world (e.g. World Atlas of Language Structure Online: Dryer and Haspelmath 2013; Grambank: Skirgård et al. 2023) and have been a useful research resource for linguistic typologists. However, due to the large scale of these studies covering the world's languages, they have not been able to capture the fine-grained variations within a single language family.

The Japonic language family is a language family comprising the indigenous languages of the Japanese archipelago (excluding Ainu), i.e. Japanese and Ryukyuan languages, many varieties of which are endangered (Moseley, 2010). Significant differences exist between the Japanese and the Ryukyuan languages and within Japanese dialects. These differences make mutual intelligibility shallow (cf. Shimoji, 2022).

Despite Japan's linguistic diversity, WALS mainly features Tokyo-Japanese, Shuri-Okinawan, and Ainu, largely ignoring other dialects. This overlook suggests a partial portrayal of the diversity of Japanese languages. Although these languages have been studied for decades, such studies remain unincorporated, highlighting a significant gap in the field due to the lack of a comprehensive dataset.

In response to this lacuna, the current research proposes the development of the web platform LAJaR: Language Atlas of Japanese and Ryukyuan. This project aims to contribute to linguistic typology by employing tools such as leaflet.js to visualize Japonic and Ryukyuan grammatical features within a geographic context. We targeted languages from a specific region, in this case, the Japonic language family.

2 Method

In the LAJaR project, we assemble a dataset from grammatical descriptions and fieldwork, adopting the WALS model to chart global languages' traits, like phonemic inventories, geographically. This facilitates empirical research into linguistic diversity's distribution and supports comparative analysis of linguistic features, revealing potential correlations. This methodology aligns with linguistic typology's goals to identify universal language attributes via detailed distribution analysis.

While LAJaR is based on WALS, this project does not intend to add data to WALS, but to create a bespoke dataset. This is to add features that are not in WALS and that are subject to question in the study of Japonic languages (e.g., the presence of dual forms of pronouns). It could also be based on Grambank, but the difference is that WALS treats grammatical features as categorical, whereas Grambank treats almost all features as binary values. Because of this difference, we decided that adapting to Grambank's method was not most appropriate for depicting grammatical features.

Building on the foundation of the WALS model, the proposed LAJaR platform intends to incorporate the methodologies and tools of the Cross-Linguistic Linked Data (CLLD) framework. CLLD is a platform designed to enable the aggregation, comparison, and visualization of linguistic data across different languages. It leverages the Cross-

[*] ROIS-DS Center for Open Data in Humanities
[†] National Institute for Japanese Language and Linguistics

Linguistic Data Formats (CLDF, Forkel et al., 2018) to standardize data representation, making it more accessible and interoperable.

Incorporating CLLD into LAJaR will offer several advantages. Firstly, it will allow LAJaR to benefit from CLLD's established infrastructure, including its robust data formats and visualization tools. Secondly, by adopting CLLD's standardized data formats, LAJaR can ensure consistency in data representation, facilitating easier comparison and integration with other linguistic databases. Moreover, CLLD supports the inclusion of a diverse range of linguistic data, from phonetics to syntax. This flexibility aligns well with the aim of LAJaR to cover a wide array of grammatical features across the Japonic and Ryukyuan languages.

3 Case study: Zero Copula

We discuss the following as a case study of how LAJaR can promote linguistic research. We annotated whether a language allows zero copula for predicate nominals (120A). We examined 25 languages from 25 sources.

The result shows that some languages of western Japan and Ryukyuans do not have copula for non-past declaratives (1a), whereas other languages do (1b).

(1) a. *wan=ja sinsii*
I=TOP teacher
'I am a teacher.'(Amami, Kato 2022, p.38)

b. *an hita sensee=zjad=do*
that person.TOP teacher=COP=SFP
'That person is a teacher.' (Kagoshima-Japanese, Hiratsuka 2018, p.115)

WALS simply states that Japanese does not allow zero copula, but LAJaR indicates that there is actually variation within the Japonic language family. The result suggests the effectiveness of LAJaR with more finely grained data extraction.

Since WALS speculate that the non-existence of copula correlates with the existence of case markers, we also wanted to examine case markers; however, we encountered some difficulties. First, different sources adopted into LAJaR refer to the region in various granularities. It is necessary to normalize the level of granularity. Second, case markers in many languages of Japan are optional. In such a case, linguists tend to describe only overt case markers and tend not to mention the possibility or impossibility of case marker drops, which makes it challenging to annotate. Third, different sources contradict each other. This might be due to language change or individual differences. It is necessary to decide which source to choose or to invent a way to represent contradicting results. The reliability of the descriptions also varies from one literature to another and must be examined to determine whether they are worthy of adoption.

While we are "digging deeper" into the Japanese data, we are essentially taking the same approach as WALS and are facing the same problems encountered by the WALS authors. On this matter, we are exploring ways to more accurately reflect the variation within languages while referencing the approach of WALS. For instance, we are considering approaches such as allowing multiple values for a specific feature.

4 Conclusion

This paper presented an overview of our ongoing project on the linguistic typology database LAJaR. We have developed a database from existing literature on Japonic languages, which until now has largely been overlooked. Additionally, through our case study, we demonstrated how this database can facilitate new research in linguistic typology.

The LAJaR project marks a significant advance in depicting the Japonic language family's diversity more fully and accurately. Using sources written in Japanese includes a broader array of linguistic data, often overlooked in current databases. Furthermore, the dataset helps members of the endangered language communities know the grammatical diversity of Japonic languages, aiding in understanding their own languages' grammatical features.

Future work will focus on further expanding and refining the database, particularly on endangered Japanese dialects and Ryukyuan languages that have received minimal attention. This expansion will enhance our understanding of the structural complexity and variation within the Japonic languages. Additionally, we aim to utilize the LAJaR database as a platform for international collaboration and exchange among researchers, further advancing the field of linguistic typology. Through this endeavor, we hope to significantly contribute to the field of language typology and raise awareness about the importance of preserving and studying Japonic languages.

References

Matthew S. Dryer and Martin Haspelmath. 2013. Wals online (v2020.3). [last accessed on 2023/12/18].

Robert Forkel, Johann-Mattis List, Simon J. Greenhill, Christoph Rzymski, Sebastian Bank, Michael Cysouw, Harald Hammarström, Martin Haspelmath, Gereon A. Kaiping, and Russell D. Gray. 2018. Cross-linguistic data formats, advancing data sharing and re-use in comparative linguistics. *Scientific Data*, 5(1):180205.

Yusuke Hiratsuka. 2018. Kagoshimaken Kagoshimashi hôgen. In *Zenkoku Hôgen Bumpô Jiten Shiryôshû 4: Katsuyô Taikê 3*, pages 107–116. Hôgen Bumpô Kenkyûkai. [The dialect of Kagoshima-city, Kagoshima Prefecture].

Kanji Kato. 2022. Tokunoshima (Kagoshima, Northern Ryukyuan). In Michinori Shimoji, editor, *An introduction to the Japonic languages : grammatical sketches of Japanese dialects and Ryukyuan languages*, chapter 2, pages 25–57. Brill, Leiden ; Boston.

Christopher Moseley, editor. 2010. *Atlas of the world's languages in danger*. UNESCO.

Michinori Shimoji, editor. 2022. *An introduction to the Japonic languages : grammatical sketches of Japanese dialects and Ryukyuan languages*. Brill, Leiden ; Boston.

Hedvig Skirgård, Hannah J. Haynie, Damián E. Blasi, Harald Hammarström, Jeremy Collins, Jay J. Latarche, Jakob Lesage, Tobias Weber, Alena Witzlack-Makarevich, Sam Passmore, Angela Chira, Luke Maurits, Russell Dinnage, Michael Dunn, Ger Reesink, Ruth Singer, Claire Bowern, Patience Epps, Jane Hill, Outi Vesakoski, Martine Robbeets, Noor Karolin Abbas, Daniel Auer, Nancy A. Bakker, Giulia Barbos, Robert D. Borges, Swintha Danielsen, Luise Dorenbusch, Ella Dorn, John Elliott, Giada Falcone, Jana Fischer, Yustinus Ghanggo Ate, Hannah Gibson, Hans-Philipp Göbel, Jemima A. Goodall, Victoria Gruner, Andrew Harvey, Rebekah Hayes, Leonard Heer, Roberto E. Herrera Miranda, Nataliia Hübler, Biu Huntington-Rainey, Jessica K. Ivani, Marilen Johns, Erika Just, Eri Kashima, Carolina Kipf, Janina V. Klingenberg, Nikita König, Aikaterina Koti, Richard G. A. Kowalik, Olga Krasnoukhova, Nora L.M. Lindvall, Mandy Lorenzen, Hannah Lutzenberger, Tônia R.A. Martins, Celia Mata German, Suzanne van der Meer, Jaime Montoya Samamé, Michael Müller, Saliha Muradoğlu, Kelsey Neely, Johanna Nickel, Miina Norvik, Cheryl Akinyi Oluoch, Jesse Peacock, India O.C. Pearey, Naomi Peck, Stephanie Petit, Sören Pieper, Mariana Poblete, Daniel Prestipino, Linda Raabe, Amna Raja, Janis Reimringer, Sydney C. Rey, Julia Rizaew, Eloisa Ruppert, Kim K. Salmon, Jill Sammet, Rhiannon Schembri, Lars Schlabbach, Frederick W.P. Schmidt, Amalia Skilton, Wikaliler Daniel Smith, Hilário de Sousa, Kristin Sverredal, Daniel Valle, Javier Vera, Judith Voß, Tim Witte, Henry Wu, Stephanie Yam, Jingting Ye, Maisie Yong, Tessa Yuditha, Roberto Zariquiey, Robert Forkel, Nicholas Evans, Stephen C. Levinson, Martin Haspelmath, Simon J. Greenhill, Quentin D. Atkinson, and Russell D. Gray. 2023. Grambank reveals global patterns in the structural diversity of the world's languages. *Science Advances*, 9.