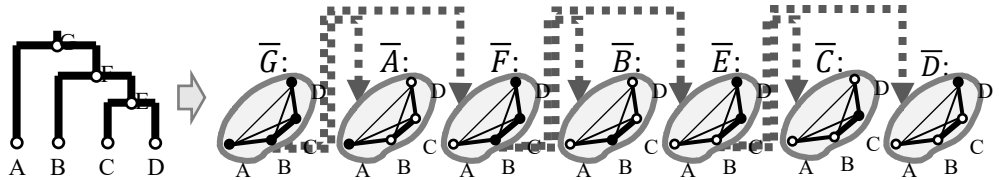
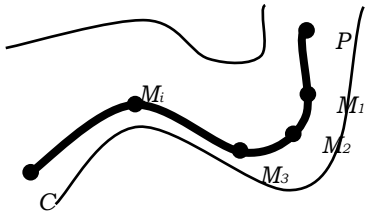
Toward Open Science in Linguistics: Adapting Practices in Japan

MURAWAKI Yūgo
Kyoto University

The Kickoff Meeting of the JSPS International Leading Research Project @ Minpaku 29 April 2025

About Me





- Background: Computational linguistics/Natural language processing (NLP)
  - Mostly working on projects involving large language models (LLMs)
- Current side/personal projects
  - Statistical modeling of spatio-temporal dynamics beyond the limits of phylogenetic approaches
  - Modeling latent representations of typological data, with a focus on diachronic change
- My goals and expectations for this project
  - Arrangement of machine-readable language resources
  - Qualitative data analysis and hypothesis generation

Collaboration with Linguists

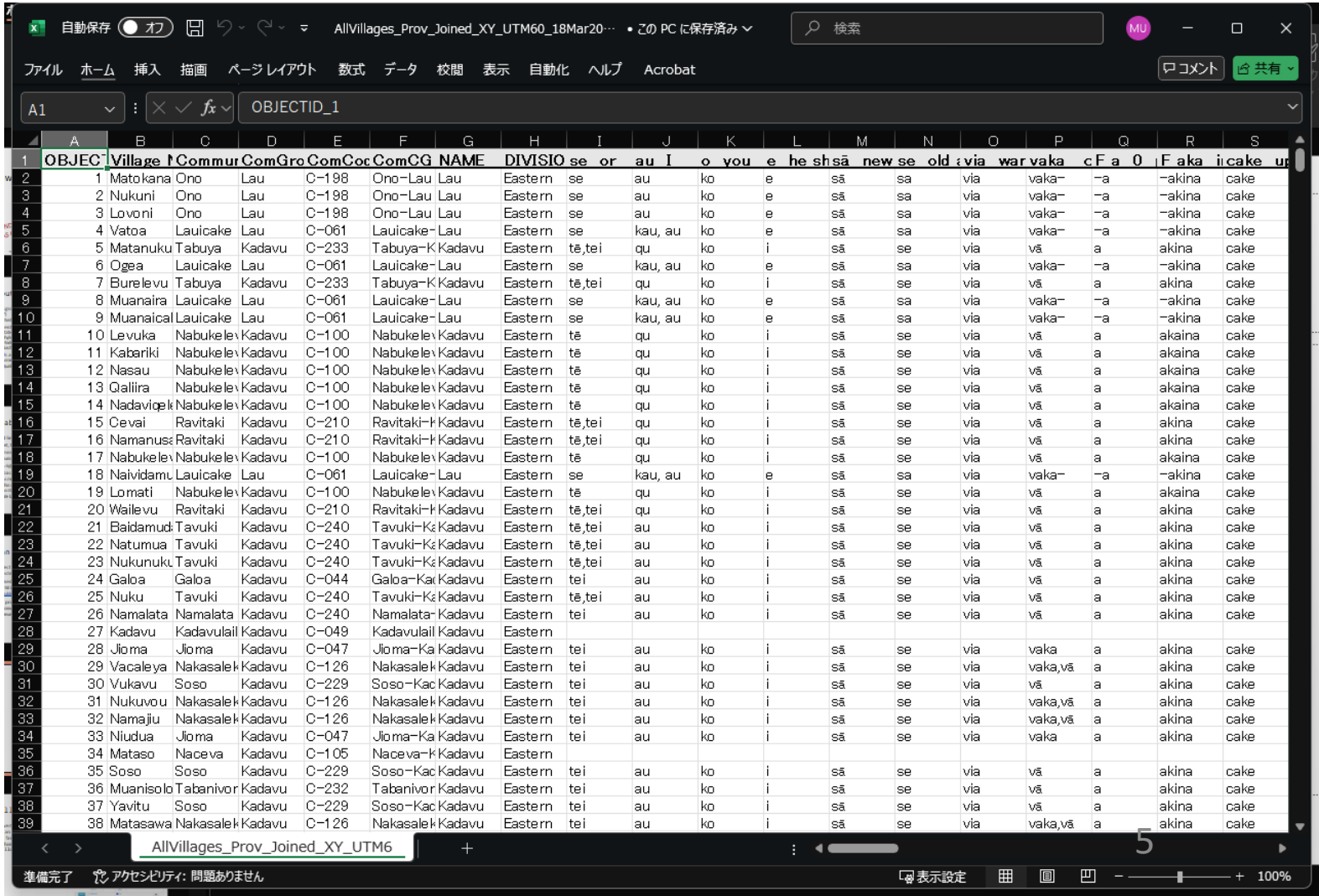
- Field linguists carry out essential, hands-on data collection
  - Yet, the resulting data is often not published in reproducible formats
- I admire the data publication initiatives in Western Europe
  - Curious how similar initiatives could be developed and sustained in Japan
  - A significant educational gap between linguistic and computational disciplines
- My background is in computer science, but...
  - As my career shits, I have less time to code
  - This remains a solo side project, with no lab members currently involved and no students trained for this work
  - We lack a dedicated engineer like Robert Forkel ☹

Fijian Language GIS Project (-- March 2025)

- Project funding primarily supports the digitization of field data collected by Paul Geraghty
- I gained access to the Fijian version of 100-word basic vocabulary list (100WL)
  - Published a preliminary statistical analysis of this data in 2020
- As a proof of concept for open science practices: I …
  - Converted the data into the CLDF (Cross-Linguistic Data Formats)
  - Developed a web app using `clld`

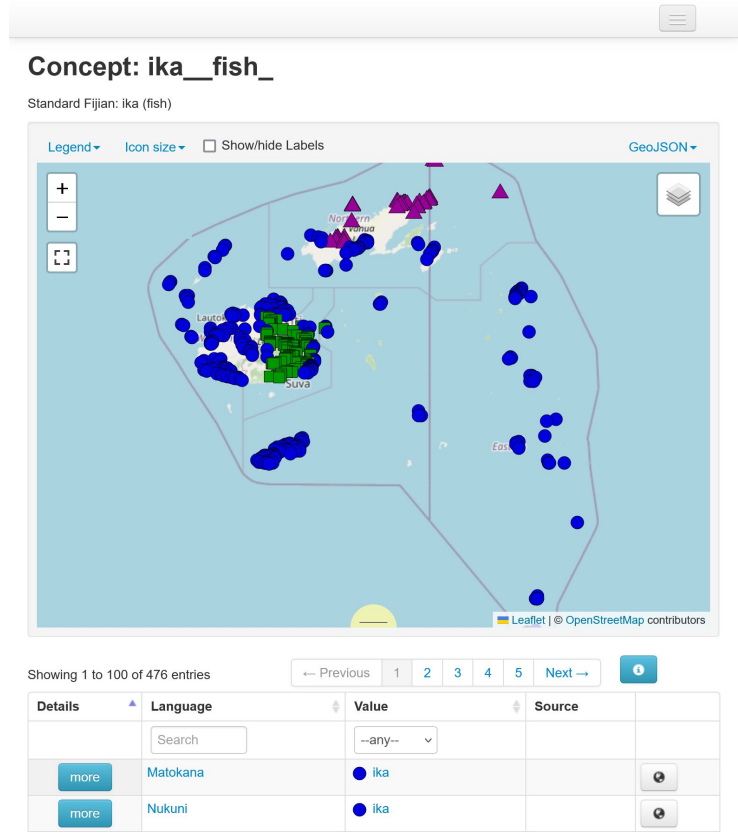
Fijian 100WL (As of 18 March 2024)

- All data is compiled in a single Excel sheet
  - This is acceptable as long as it includes all essential information
- Each village is linked to a geolocation based using the Fiji Map Grid Coordinate system



CLDF and CLLD

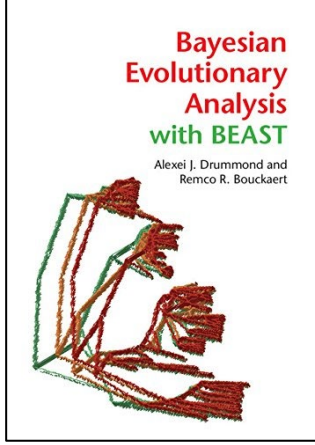
- >100 lines of Python code for CLDF conversion, available at github
  - `pycldf`, `pandas`
  - `pyproj` for coordinate system conversion
- Adapted `clld` (+861 -86 lines changed)
  - Custom `scripts/initializedb.py`
  - Support villages, communalects and communalect groups, instead of languages
  - Custom map markers



Lessons Learned

- CLDF/CLLD are well-designed resources
  - CLDF is a solid data format for long-term publication
  - CLLD is reasonably extensible, though the longevity of its underlying web frameworks is uncertain
- But too hard-coded for linguistic use
  - Potential barriers to interdisciplinary collaboration, such as with archaeologists
- Linguists cannot be expected to build infrastructure on their own
- Vibe coding has the future (?)

Promoting Bayesian Phylogenetics



- Open science practices could be facilitated if linguists themselves can benefit from open data
- Bayesian phylogenetics, especially using BEAST, is increasingly common in linguistics research abroad, but remains virtually absent in Japan
- There are a textbook and a tutorial for linguists but they assume an advanced level of statistical knowledge that most linguists lack
- Explaining these methods in a linguist-friendly way would likely require something on the scale of a full university lecture course
- Challenges:
  - Is there real demand for this?
  - Do I have the capacity to take it on?

Applying Large Language Models (LLMs) to Data Construction (and Analysis)

- Idea: Can LLMs help streamline linguistic research?
  - Many routine tasks linguists perform, especially data construction, might be partially delegated to LLMs
  - Insights from this could also feed back into LLM development (my main work)
- Proposal:
  - Create a benchmark dataset showcasing tasks that, if done well by LLMs, would benefit linguists
  - Encourage LLM developers to compete for performance on this benchmark
- Next step: Identify tasks that are promising candidates for LLM delegation