Plan Overview

A Data Management Plan created using DMPonline

Title: Classifiers in the landscape: a Baniwa case study

Creator: Sandra Cronhamn

Principal Investigator: Sandra Cronhamn

Affiliation: Lund University

Funder: Swedish Research Council

Template: Swedish Research Council Template

ORCID iD: 0009-0005-6727-6436

Project abstract:

Previous research has revealed a remarkable variation in the way speakers of different languages categorize landscape features, but this research has focused primarily on landscape nouns. Classifiers are grammatical systems that explicitly divide nouns into categories based on shape and other physical characteristics, providing an excellent opportunity for potential new insights into landscape categorization. Baniwa, an Arawakan language spoken in Northwestern Brazil, has a complex, shape-based classifier system, and preliminary observations indicate that Baniwa classifiers interact with nouns in interesting ways in the landscape domain.

Over the course of three years, this project will address the previously unexplored role of classifiers in landscape categorization, from both language-specific and cross-linguistic perspectives. The language-specific studies will focus on Baniwa, using a combination of field-based descriptive and experimental approaches to investigate the linguistic and cognitive landscape categorization of Baniwa speakers, and the role of classifiers therein.

The cross-linguistic perspective will consist of a typological study, where I survey published sources on classifier languages world-wide in order to establish the current state of knowledge on the topic. The joint results of these complementary studies have the potential to significantly advance our understanding of both landscape categorization and the semantic underpinnings of classifier systems.

ID: 186247

Start date: 18-08-2025

Last modified: 30-10-2025

Grant number / URL: 2025-00350

as they would like in their own plans, and customise it as necessary. You do not need to credit the creator(s) as the source of the language used, but using any of the plan's text does not imply that the creator(s) endorse, or have any relationship to, your project or proposal

Classifiers in the landscape: a Baniwa case study

General Information

Project Title

Classifiers in the landscape: a Baniwa case study

Project Leader

Sandra Cronhamn

Registration number/corresponding

2025-00350

Version

2

Date

2025-09-11

Description of data – reuse of existing data and/or production of new data

How will data be collected, created or reused?

This project will both produce new data and make use of existing data.
The new data will be on the Baniwa language (Arawakan, Northwest Amazonia). It will be collected through fieldwork in São Gabriel da Cachoeira, Brazil, and consist primarily of audio and video recordings of language, annotated recordings, field notes, and text files and/or spreadsheets with data compilations. The data will be collected via different methods, including free narrative and conversational speech, director/matcher tasks, and controlled experiments.
The existing data will be collected from published linguistic descriptions of various other languages, such as grammars, grammar sketches, dictionaries, and research articles. It will be compiled into test files and/or spreadsheets, with bibliographical references for each data point. Bibliographical references will also be given whenever the individual datapoints are referenced in publications.

What types of data will be created and/or collected, in terms of data format and amount/volume of data?

This project will primarily collect and handle data in the form of audio (wav) and video (mp4) recordings, annotated ELAN files (eaf), text files (docx, txt), spreadsheets (csv, xlsx), and scanned field notes (pdf). Supporting documents in the form of, e.g., images and photos (jpeg, png, tiff) may also be part of the data collection. Decisions on file formats have been made in collaboration with archiving experts at the Lund University Humanities Lab, where the data will be archived. The formats have been chosen for their long-term durability, and reflect standard practices for data repositories.

While the final data volumes are yet unknown, I expect the audio and video recordings (which will constitute the largest volumes in terms of storage space) to comprise somewhere between 20 and 40 hours.

Documentation and data quality

How will the material be documented and described, with associated metadata relating to structure, standards and format for descriptions of the content, collection method, etc.?

The new, primary data (audio and video recordings) will be annotated with metadata relating to the recording session, in accordance with the FAIR principles (Findable, Accessible, Interoperable, Reusable). The metadata will be encoded in a cmdi file, containing information about the language, people involved (speakers, researcher), date and place of recording, data type (method, genre), and a brief description of the content. The files in question will have unique ID:s built up of some of this information, namely:

1. The ISO 639-3 language code (BWI for Baniwa)
2. The initials of the researcher in charge of the data collection (SC for Sandra Cronhamn)
3. The date of the recording (YYMMDD)
4. A shorthand for the data type (e.g., NARR for unstructured narrative data)
5. An anonymous identifier letter representing the speaker (assigned randomly; if there are several speakers in the recording, only the letter of the main or first speaker is used in the ID)
6. A chronological session number

For example, the ID *BWI_SC_221005_ELIC_K_02* features a recording of an elicitation session of the Baniwa language, collected by Sandra Cronhamn, recorded on the 5th of October, 2022, with the speaker “K”, and it is the second session of this kind with this speaker recorded on this particular day.

Secondary data files (e.g., field notes, annotated ELAN files, photographs and other supporting images) will be connected to the primary data files they accompany. The unique file ID:s allow for cross-referencing within the archive.

When archived, the data will be accompanied by a guide document (pdf) placed on the landing page and containing information about the collection’s structure, contents, file formats, collection methods, data access, citation information, etc.

It is not yet known exactly what the final structure of the archived collection will look like, but the data will be stored in some organized folder structure deemed suitable for the accessibility and re-usability of the data (e.g., by collection method).

How will data quality be safeguarded and documented (for example repeated measurements, validation of data input, etc.)?

The data collected via fieldwork will be cross-checked across several Baniwa speakers in order to assure that it is representable for the speech community as a whole.

For the published data that I am collecting for the typological comparison, I will have to rely on the information provided by other researchers.

Storage and backup

How is storage and backup of data and metadata safeguarded during the research process?

During the project periods spent at Lund University and the University of Copenhagen, project data will be continuously and automatically backed up on Lund University’s servers.

During fieldwork in Brazil, the collected data will be stored on a laptop computer and backed up to at least two external hard drives or memory cards daily, as is standard practice. During this period, internet access is not reliable enough to rely on automatic backups on Lund University’s servers.

After fieldwork, the collected data will be uploaded onto Lund University Humanities Lab Archive for long-term storage. Data will also be stored in Lund University's data management system, at least during a limited period upon the termination of the project.

How is data security and controlled access to data safeguarded, in relation to the handling of sensitive data and personal data, for example?

The data and metadata relating to the project will be stored on the Lund University Humanities Lab Archive Server. Where no restrictions apply, the data will be openly available; this concerns for instance the cross-linguistic compilation of published data for the typological study. Data such as audio and video recordings, where speakers are featured and identifiable via their faces and voices, will be available upon request provided that the speakers agree to it. In the openly browsable metadata, individual speakers will be anonymized by being assigned an identifier letter combination at random.

See just above for data storage on external hard drives during fieldwork periods.

Legal and ethical aspects

How is data handling according to legal requirements safeguarded, e.g. in terms of handling of personal data, confidentiality and intellectual property rights?

The data and metadata relating to the project will be stored on the Lund University Humanities Lab Archive Server. Where no restrictions apply, the data will be openly available; this concerns for instance the cross-linguistic compilation of published data for the typological study. Data such as audio and video recordings, where speakers are featured and identifiable via their faces and voices, will be available upon request provided that the speakers agree to it. In the openly browsable metadata,

individual speakers will be anonymized by being assigned an identifier letter combination at random.

In accordance with the Swedish Act concerning the ethical review of research involving humans (2003:460), studies that do not satisfy any of the listed criteria are exempt from the requirement for ethical approval. In particular, if the study does not involve animals, children or other vulnerable groups, the collection of biological materials, physical manipulation, an obvious risk of harm to the participant, sensitive questions such as race, sexual practices, etc., ethical review does not need to take place (http://www.researchethics.lu.se/researchethics-information/ethical-review/when-is-ethical-permission-required). However, to be on the safe side, I will apply for an ethics review at the Swedish Ethical Review Authority before collecting any data.

The fieldwork for this project will be conducted in São Gabriel da Cachoeira, a small town in the remote Brazilian Amazon, as well as in its immediate surroundings. As the town is not situated in Indigenous territory, no federal permits are required. Instead, as I have done during previous fieldwork in this location, I will obtain permission through the local Indigenous council, FOIRN (Federation of the Indigenous Organizations of Rio Negro), in addition to individual informed consent by all participants.

For all participants, I will record their age, sex, first language, and any additional languages. Some types of data will involve personal identification in the form of audiovisual recordings. The project will not collect, handle or store sensitive personal data of any kind, and will strive to ensure that such data are not inadvertently included in recorded materials. In the unlikely case that sensitive data does turn up in the recorded material (for example, if a participant brings something up on their own accord), I will refrain from transcribing it. One caveat is that a participant's first language may, under some circumstances, hint at their ethnicity, but there is no 1-1 relationship between the two. Ethnicity as such is not a parameter that will be registered or studied in the project.

All data will be stored and managed in accordance with GDPR regulations, as well as with the CARE principles (Collective benefit, Authority to control, Responsibility, Ethics) for Indigenous Data Governance.

How is correct data handling according to ethical aspects safeguarded?

See reply just above.

Accessibility and long-term storage

How, when and where will research data or information about data (metadata) be made accessible? Are there any conditions, embargoes and limitations on the access to and reuse of data to be considered?

The data and metadata relating to the project will be stored on the Lund University Humanities Lab Archive Server. Where no restrictions apply, the data will be openly available; this concerns for instance the cross-linguistic compilation of published data for the typological study. Data such as audio and video recordings, where speakers are featured and identifiable via their faces and voices, will be available upon request provided that the speakers agree to it.

For transparency and re-usability reasons, data used for specific research articles produced within the project (e.g., spreadsheets containing cross-linguistic comparative data) may additionally be stored and made accessible in supplementary materials files, OSF/github repositories or the like, in accordance with the scientific journals' requirements.

In what way is long-term storage safeguarded, and by whom? How will the selection of data for long-term storage be made?

Long-term storage is safeguarded by Lund University Humanities Lab Archive Server. All data collected for the project will be stored there. The data will also be stored in Lund University's archive.

Will specific systems, software, source code or other types of services be necessary in order to understand, partake of or use/analyse data in the long term?

The data used for this project will primarily be stored in the folloring file formats, which are all widely used and accessible: wav, mp4, docx, txt, csv, xlsx, pdf, jpeg, png, and tiff. The only file format that is not widely used is eaf, used for storing ELAN files containing time-aligned annotations connected to audio and video recordings. Eaf files can be opened in the open access ELAN software, which can be downloaded from https://archive.mpi.nl/tla/elan.

How will the use of unique and persistent identifiers, such as a Digital Object Identifier (DOI), be safeguarded?

The data collection will receive a persistent identifier in the form of a link provided by the Lund University Humanities Lab.

Responsibility and resources

Who is responsible for data management and (possibly) supports the work with this while the research project is in progress? Who is responsible for data management, ongoing management and long-term storage after the research project has ended?

As the project manager and the sole employee in the project, I am responsible for all aspects of the data management during the course of the project, including DMP implementation and review, data collection, metadata production, data quality, and ethical and legal compliance. By agreement, the Lund University Humanities Lab Archive will be responsible for the long-term storage and backup of the data after the research project has ended.

What resources (costs, labour input or other) will be required for data management (including storage, back-up, provision of access and processing for long-term storage)? What resources will be needed to ensure that data fulfil the FAIR principles?

Data management will require continuous time on the part of the project leader. This is built into the workflow and routines of the data collection and management. When the data is ready for archiving, it will also require some time on the part of the Lund University Humanities Lab’s staff, but only during limited periods of time.

Expenses relating to data management include project costs for a computer and external hard drives/memory cards, and possibly a one-time fee for archiving at the Lund University Humanities Lab Archive.