Language Corpora and Documentation: A Guide to the Ende Corpus

The corpus of Ende and other Pahoturi River languages ([1]) is the largest collection of auditory/visual data with accompanying transcriptions, translations, and interlinearizations available for any language in southern Papua New Guinea, a region known for its incredible diversity ([2]) and shockingly low rates of documentation. In this presentation, we will discuss the Ende corpus based on Thieberger and colleagues' ([3]) criteria for evaluating archival collections and present the types of academic and community-level impact that have resulted from the data.

Ende is one of six significantly undocumented Pahoturi River languages and is spoken primarily in Limol, Malam, and Kinkin villages of Western Province Papua New Guinea. The Ende language project was initiated by Ende pastor Warama Kurupel and has resulted in a collection of 116 hours of audio and 54 hours of video recordings, 5,000 words for the Ende-English lexicon, several Ende-medium books, a hymnal, a documentary, and a dissertation. Since 2018, the open-access archival collection has resulted in several research projects, which will be summarized.

The data were collected in three locations in southern New Guinea: the town of Daru and the villages of Limol and Malam. The collection includes spontaneous, elicited and prompted data types with genres including elicitation, narrative, song, formulaic discourse, and interactive discourse.

We asssess the corpus in terms of Thieberger and colleagues' criteria for evaluating an archive of a previously undocumented language, specifically referring to accessibility, quality, and quantity. This corpus has followed best practice methods for archiving endangered language data, which include archiving data in a repository that provides long-term maintenance, providing each item with a unique identifier and a citation system, organizing the data in a hierarchical structure with easy access to metadata including background and contextual information, and linking the linguistic annotations to the raw data. This corpus also features an innovative method of bypassing the "researcher bottleneck" that stymies data collection and analysis: native speaker training and collaboration. This has resulted in very high quantity for this collection, including a diverse range of speakers, ages, genders, and dialects, and extensive digital annotations.

The Ende language corpus is archived in PARADISEC and includes 876 items; 306 with video (54 hours) and 780 with audio (116 hours) collected from over one hundred speakers. The collection is nearly 50% transcribed, 25% translated, and 5% interlinearized. The corpus follows a hierarchical structure, where each item is a folder containing all raw data files and associated annotation files. Each item and file has its own page with metadata, a unique ID and citation information. As a large open-access collection, the corpus gives researchers with various research and documentary efforts a place to find desired data and acts as a template of an ideal corpus collection.

The benefits of language documentation and description are twofold in terms of their linguistic and community focus. Within linguistics and academia, a thorough and holistic corpus documenting understudied languages provides a specific landing point for users to easily access data and gives linguists a collection that can prompt or address specific research questions. An extensive corpus also allows research outputs to easily cite the original sources of the data rather than derivative works. A community-based focus emphasizes speakers' desire to document and share their language and culture outside of their community and provides them with resources to promote literacy efforts within these communities.

Current and future work on the Ende language project is presently being undertaken across various universities and focuses on the historical, sociolinguistic, language variation, language description, and structural analysis of the language. The corpus is also being updated with additional files and a guide is being prepared for submission to *Language*'s new *Language Revitalization and Documentation* subsection ([4]).

The goal of this presentation is to highlight the impact that documenting understudied languages has within academia and communities while emphasizing the innovative methods behind them. The Ende corpus is an example of a collection that has served several research projects and has resulted in a high quantity collection of archived data. We hope that this presentation serves as a guide for other researchers interested in developing corpora for undocumented and understudied lanaguges.

References: [1] Author 2. 2015. *Language corpus of Ende and other Pahoturi River Languages.* Canberra: PARADISEC. [2] Evans, N. 2012. "Even more diverse than we had thought: The multiplicity of Trans-Fly languages." *Melanesian languages on the edge of Asia: Challenges for the 21st century.* University of Hawaii Press. [3] Thieberger, N. et al. 2015. "Assessing annotated corpora as research output." *Australian Journal of Linguistics.* 36.1. 1-21. [4] Fitzgerald, C. "A framework for Language Revitalization and Documentation." *Language*. 91.1. e1-e11.