## From corpus to grammar: a bottom-up approach to grammatical description

The past decades have brought to light numerous issues with the traditional method of reference grammar writing, both from a scientific and a community-based perspective (Ameka et al. 2006; Nordhoff 2008, 2012; Thieberger 2009). Amongst others, traditional paper-based reference grammars often provide limited access to the primary data they are based on; they are not well suited to address language-internal variation, with an additional risk of becoming sources of prescriptivism, potentially hindering revitalisation efforts (Florey 2004); and they may come with a prohibitive price tag detrimental both to communities and to academics in developing countries.

This paper discusses the inception and development of a methodology aimed at addressing some of these concerns, by integrating an open-access, hypertext grammatical description with a corpus. Rather than approaching the process from the envisioned end state, i.e. a comprehensive, linearly organized reference grammar, the method takes an existing language corpus as its starting point and from there, builds the description from the ground up. This description is based on an existing corpus of Idi (ISO-369: idi; Pahoturi River family), a Papuan language spoken in Western Province, Papua New Guinea. This corpus of naturalistic speech (59.3 hours/80,000 words transcribed) was collected throughout six years of involvement with the Idi speech community in Dimsisi village, in close collaboration with a local language committee to ensure both ongoing broad community support and a representative sample of interspeaker variation. Raw audio and video recordings are archived with PARADISEC (Author1 2015).

Transcribed audio recordings, annotated transcripts and relevant metadata are stored in and accessible through LaBB-CAT, an open-source, browser-based linguistic data management and research tool (Fromont & Hay 2021). This tool has wide-ranging functionality in terms of tagging and annotation, integration with commonly used linguistics software such as Transcriber, ELAN and Praat, options for exporting, and performing other customized operations using e.g. Python, R or Java. LaBB-CAT has been widely and successfully used to manage and perform research on corpora of te reo Māori and German, in addition to multiple varieties of English.

In this paper, we discuss the process of adapting the existing LaBB-CAT infrastructure in order to develop a modular grammatical description linked to the Idi corpus. In addition to addressing technical issues and challenges, we will also touch upon broader issues that arise when approaching language description from a data-driven perspective. These are, for instance, the balance between 1) striving for accountability to the complete corpus and 2) attempting to be as illustrative and reader-friendly as possible in the description, or the issue of internal consistency between different elements of the description.

Our approach to integrating corpus and grammatical description provides important lessons for establishing best practices while the field of language documentation and description is entering a new era. As Evans and Dench (2006: 30) put it, ‘Thus where the traditional corpus, as language documentation, may have been defined by the descriptive agenda, there is the possibility that hypertext grammars linked to multi-modal corpora may come to be shaped instead by the documentary enterprise.’ The present study hopes to contribute to this goal, and in this way, to improve and reorient current practice.

## References

Author 1. 2015. Recordings of the Idi language (WSDS1). *Digital collection managed by PARADISEC* [Open access]. DOI: 10.4225/72/56E97A18E14F3.

Ameka, Felix K., Alan Dench, & Nicholas Evans (Eds.). 2006. *Catching Language: The Standing Challenge of Grammar Writing*. Berlin: De Gruyter Mouton.

Evans, Nicholas & Alan Dench. 2006. Introduction: Catching language. In Felix Ameka et al. (Eds.), pp. 1-40.

Fromont, Robert & Jennifer Hay. 2012. LaBB-CAT: an annotation store. *Proceedings of Australasian Language Technology Association Workshop*, pp. 113-117. Dunedin: Australasian Language Technology Association.

Florey, Margaret. 2004. Countering purism: confronting the emergence of new varieties in a training program for community language workers. In Peter K. Austin (Ed.), *Language Documentation and Description*, vol. 2, pp. 9-27. London: SOAS.

Nordhoff, Sebastian (Ed.). 2012. *Language Documentation and Conservation Special Publication* 4: Electronic Grammaticography.

Nordhoff, Sebastian. 2008. Electronic Reference Grammars for Typology: Challenges and Solutions. *Language Documentation and Conservation* 2(2): 296-324.

Thieberger, Nicholas. 2009. Steps toward a grammar embedded in data. In Patience Epps & Alexandre Arkhipov (Eds.), *New Challenges in Typology: Transcending the Borders and Refining the Distinctions*, pp. 389-407. Berlin: Mouton de Gruyter.